# Business Analytics



**ISBN:** 978-93-91540-81-4

# Table of Contents

# Course Outcomes

After studying this book, you would be able to delve into a comprehensive and systematic exploration of the field of business analytics. The book provides a holistic understanding, empowering individuals to apply analytics in real-world scenarios for informed decision-making. The book comprises the following twelve chapters:

**Chapter 01: Getting Started with Business Intelligence-** This chapter provides an overview of the book, explaining the concepts of data, information, knowledge and wisdom. It also discusses how to manage data. Further, it explains the concept of business view of information technology applications and defines business intelligence.

**Chapter 02: Introduction to Business Analytics-** This chapter gives insights into the concepts of types of business analytics, relation between business intelligence and business analytics, and the role of business models in analytics. Towards the end, the chapter explains importance of business analytics and emerging trends in business intelligence and business analytics.

**Chapter 03: Resource Considerations to Support Business Analytics**- This chapter delves into the analytics personnel and their roles. It then introduces required competencies for personnel in analytics. Further, it explains business analytics data. It concludes by explaining about the technology for business analytics.

**Chapter 04: Introduction to Statistics-**This chapter commences with the concepts of measures of central tendency, probability theory, and statistical inference. Further, the chapter discusses hypothesis testing, correlation and regression analysis.

**Chapter 05: Decision Making and Support-** This chapter provides an overview of the concepts of decision making, decision support system, and techniques of decision making. It concludes by throwing light on data mining with decision trees and application of data science for decision making in key area.

**Chapter 06: Introduction to OLTP & OLAP-** This chapter lays the groundwork by elucidating the basic concepts of online transaction processing (OLTP), online analytical processing (OLAP), and different OLAP architectures. Further, it provides a comparison between OLTP & OLAP and discusses OLAP operations.

**Chapter 07: Data Warehousing-** This chapter commences by explaining the concept of data warehousing: an informational environment and its key components. It also explains data warehouse design techniques and ETL process.

**Chapter 08: Descriptive, Predictive, Prescriptive and Diagnostic Analytics-** This chapter introduces the concepts of descriptive analytics, descriptive statistics, and predictive analytics. Further, it discusses the concept of predictive modelling. It also delves into model comparison and improvement. Towards the end, it discusses prescriptive analytics and diagnostic analytics.

**Chapter 09: Data Representation and Visualisation-** This chapter begins by throwing light upon the ways of representing visual data, various techniques used for visual data representation, and types of data visualisation. Further, it discusses the applications of data visualisation, visualising big data, tools used in data visualisation, data visualisation for managers, and visualising and exploring data in Excel.

**Chapter 10: Quantitative Techniques-** This chapter commences by explaining the concept of linear programming. It then delves into the various problems. Towards the end, the chapter explains additional problems in quantitative techniques.

**Chapter 11: System Management and KPI-** This chapter begins by introducing the need for a system management and data quality. Further, it explains the concept of business metrics and Key Performance Indicators (KPIs).

**Chapter 12: Business Analytics in Practice-** This chapter begins by highlighting the basic concepts of financial and fraud analytics, HR analytics, marketing analytics, and healthcare analytics. Further, it explains supply chain analytics, web analytics and stock market analytics, and analytics for government and NGOs.

CHAPTER 1

# Getting Started with Business Intelligence

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the different types of data
- Explain how to manage data
- Discuss the business view of information technology applications
- Elucidate the need for Business Intelligence (BI)

## 1.1 INTRODUCTION

Information is the most valuable asset in any modern organisation. However, merely possessing information cannot guarantee success in business. Business organisations need to analyse information to generate actionable insights for strategic decision-making. This gives a competitive advantage to organisations and a much sought-after value addition that sets a business apart from the competition.

Business Intelligence (BI) is one such tool that helps organisations in analysing information and making an effective business decisions. It also helps in identifying new opportunities for generating revenue through the following ways:

- Optimise the existing technology and business framework
- Speed up decision-making and problem resolution
- Eliminate operational inefficiencies
- Streamline the regulatory compliance and adhere to industry-best practices
- Identify new business opportunities
- Test new opportunities or approaches through segmentation and data testing

The capability to earn lucrative returns from business data brings a competitive edge to modern organisations. An organisation can achieve an edge over competitors by using effective BI solutions or software according to its needs and environment.

This chapter first explains the relation between data, information and knowledge. Further, this chapter discusses different types of data and its management. It also describes the impact of information technology in business. Towards the end, this chapter explains the importance of business intelligence.

## 1.2 DATA, INFORMATION, KNOWLEDGE AND WISDOM

Data, simply put, is the raw material that does not make any definite sense unless you process it to any meaningful end. It needs to be processed with a context, before being logically viable.

The examples of data are as follows:

2, 4, 6, 8

Mercury, Jupiter, Pluto

Notes

The above data alone does not represent the true picture. Maybe the sequence above is simply the table of two or a sequence denoting the difference of two between numbers. The names may just be the names of conference rooms in an organisation rather than being planet names, unless you give it logic and define the reasoning for its existence. The data does not have a standalone existence by itself.

Information is the result that we achieve after the raw data is processed. This is where the data takes the shape as per the need and starts making sense. Standalone data has no meaning. It only assumes meaning and transitions into information upon being interpreted. In IT terms, characters, symbols, numbers, or images are the data. These are joint inputs which a system running a technical environment needs to process in order to produce a meaningful interpretation.

Information can offer answers to questions such as which, who, why, when, what and how. Information put into an equation should look like:

Information = Data + Meaning

The examples of Information are as follows:

2, 4, 6 and 8 are the results of the first four multiples of 2. Mercury, Jupiter and Pluto are the names of the planets.

When we allocate a situation or meaning, only then the data becomes information.

Generally, the data is available in raw form. Information is processed from data, and knowledge is gained from information. Knowledge is the proper assembly of meaningful information whose intent is to be valuable. It is a deterministic process.

Knowledge can be of two types:

- Obtaining and memorising the facts
- Using the information to crack problems

The first type is called explicit knowledge, meaning the knowledge that can be simply transferred to others. Explicit knowledge and its offspring can be kept in a certain format, e.g., encyclopedias and textbooks.

The second type is termed as tacit knowledge, referring to the type of the knowledge that is complex and intricate. It is gained simply by passing on to others and requires elevated and advanced skills in order to be comprehended. For example, it will be tough for a foreign tourist to understand the local customs or rituals of a specific community located in a country whose language is different from the tourist's language. In such a case, the tourist needs to be conversant with the language or requires additional resources in order to understand the rituals. Similarly, the ability to speak a language, use a computer or similar other things requires knowledge that cannot be gained explicitly but rather learned through experience.

### 1.2.1 HOW DATA, INFORMATION AND KNOWLEDGE ARE LINKED?

Data signifies an element or statement of procedures without being related to other things. Information symbolises the relationship of some types that act as a bridge between the data and information.

The topics are hierarchical in the following order, as shown in Figure 1:


Figure 1: Transforming Data into Knowledge

For example, the temperature fell 15 degrees followed by rains. Here, temperature falling by 15 degrees is the information and the inference is that rains lead to fall in temperature.

Knowledge signifies a design that links and usually provides a high-level view and likelihood of what will happen next or what is described.

For example, if the humidity levels are high and the temperature drips considerably, the atmosphere is pretty much unlikely to hold the moisture and the humidity. Hence, it rains. The pattern is reached on the basis of comparing the valid points emanating from data and information resulting into the knowledge or sometimes also referred to as wisdom.

Wisdom is systematic and allows you to comprehend the interaction taking place among temperature gradients, raining, evaporation and air currents.

### Self Assessment Questions

1. _____ is something that is inferred from the data and the information.
2. Which types of knowledge and its offspring can be kept in a certain media format, e.g., encyclopedias and textbooks?
   a. Explicit
   b. Implicit
   c. Tacit
   d. None of these
3. Wisdom exemplifies the understanding of essential values, personified within the knowledge, that act as a foundation for the knowledge in its current form. (True/False)

## 1.3 TYPES OF DATA

Data that comes from multiple sources – such as databases, Enterprise Resource Planning (ERP) systems, weblogs, chat history and GPS maps – varies in its format. However, different formats of data need to be made consistent and clear to be used for analysis. Data acquired from various sources can be categorised primarily into the following types of sources:

- Internal sources, such as organisational or enterprise data
- External sources, such as social data

NOTES

Table 1 compares the internal and external sources of data:

TABLE 1: Comparison between the Internal and External Sources of Data

| Data Source | Definition | Examples of Sources | Application |
|---|---|---|---|
| Internal | Mostly provides structured or organised data that originates from within the enterprise and helps run the business | • Customer Relationship Management (CRM)<br>• Enterprise Resource Planning (ERP) systems<br>• Customers' details<br>• Products and sales data<br>• Generally online transaction processing (OLTP) and operational data | This data (current data in the operational system) is used to support daily business operations of an organisation |
| External | Mostly provides unstructured or unorganised data that originates from the external environment of an organisation | • Business partners<br>• Syndicate data suppliers<br>• The Internet<br>• Government<br>• Market research organisations | This data is often analysed to understand the entities mostly external to the organisation such as customers, competitors, market and environment |

### 1.3.1 STRUCTURED DATA

Structured data can be defined as the data that has a defined repeating pattern. This pattern makes it easier for any program to sort, read and process the data. Structured data is:

- Organised data in a predefined format
- The data that resides in fixed fields within a record or file
- Formatted data that has entities and their attributes mapped
- Used to query and report against predetermined data types

Some sources of structured data include:

- Relational databases (in the form of tables)
- Flat files in the form of records (like CSV and tab-separated files)
- Multi-dimensional databases (majorly used in data warehouse technology)
- Legacy databases
- Microsoft Excel

Table 2 shows a sample of the structured data in which the attribute data for every customer is stored in the defined fields:

TABLE 2: Sample of Structured Data

| Customer ID | Name | Product ID | City | State |
|---|---|---|---|---|
| 12365 | Smith | 241 | Graz | Styria |

| Customer ID | Name | Product ID | City | State |
|---|---|---|---|---|
| 23658 | Jack | 365 | Wolfsberg | Carinthia |
| 32456 | Kady | 421 | Enns | Upper Austria |

### 1.3.2 UNSTRUCTURED DATA

Unstructured data is a set of the data that might or might not have any logical or repeating patterns.

Unstructured data:

- Consists typically of meta data, i.e., the additional information related to data.
- Comprises inconsistent data, such as data obtained from files, social media websites, satellites, etc.
- Consists of data in different formats such as e-mails, text, audio, video, or images.

Some sources of unstructured data include:

- **Text both internal and external to an organisation:** Documents, logs, survey results, feedbacks, PowerPoint presentations and body of e-mail from both within and across the organisation.
- **Social media:** Data obtained from social networking platforms in the form of images and videos, such as YouTube, Facebook, X (formely Twitter), LinkedIn, etc.
- **Mobile data:** Data such as text messages and location information.

Unstructured data is generally deployed to:

- Gain considerable competitive advantage by organisations.
- Gain a clear, complete, and big picture of future prospects by organisations.

Working with unstructured data poses certain challenges which are as follows:

- Identifying the unstructured data that can be processed.
- Sorting, organising, and arranging unstructured data in different sets and formats.
- Combining and linking unstructured data in a more structured format to derive any logical conclusions out of the available information.
- Costing in terms of storage space and human resource (data analysts and scientists) needed to deal with the exponential growth of unstructured data.

### 1.3.3 SEMI-STRUCTURED DATA

Semi-structured data, also known as schema-less or self-describing structure, refers to a form of structured data that contains tags or markup elements in order to separate semantic elements and generate hierarchies of records and fields in the given data. To be organised, the semi-structured data should be fed electronically from the database systems, file systems and through data exchange formats including scientific data and XML (eXtensible Markup Language). XML enables data to have an elaborate and intricate structure that is significantly richer and comparatively complex.

Notes

Some sources for semi-structured data include:

- Database systems
- File systems like Web data and bibliographic data
- Data exchange formats like XML, X12, EDI, HL7, etc.
- Zipfile
- Binary executable file

### 1.3.4 DIFFERENCE BETWEEN STRUCTURED AND SEMI-STRUCTURED DATA

The differences between structured and semi-structured data are as shown in Table 3:

**Table 3: Differences between Structured Data and Semi-Structured Data**

| Structured Data | Semi-Structured Data |
|---|---|
| Structured data can be defined as the data that has a defined repeating pattern. | Semi-structured data refers to a form of structured data that contains tags or markup elements in order to separate semantic elements and generate hierarchies of records and fields in the given data. |
| Processing structured data is much easier and faster. | Processing semi-structured data is slower than processing of structured data. |
| Structured data is organised data in a predefined format. | To be organised, semi-structured data should be fed electronically from the database systems, file systems, and through data exchange formats. |
| Some sources of structured data include relational databases, flat files, multidimensional databases and legacy databases. | Some sources for semi-structured data include database systems, file systems like web data and bibliographic data, Data exchange formats like scientific data, etc. |

### 1.3.5 QUANTITATIVE DATA

Quantitative data refers to the data that is related to the quantities of real-world objects. This type of data can be measured and stated by using numbers. Some examples of quantitative data can be height of a person, age of a vehicle, money in your pocket, etc. The quantitative data can be of two categories:

- Discrete (finite values)
- Continuous (measurements)

An organisation might encounter different types of data while dealing with its clients or customers. By analysing whether the data is discrete or continuous, appropriate methods can be used for analysis of data and its reporting.

The data that is countable is known as discrete data. For example, number of questions attempted in a paper must be discrete as you can easily count them. Some more examples of discrete data could be number of food items in a catalogue, number of employees working in a company, number of feedbacks for a particular product, etc.

The continuous data is the data that can be measured in terms of fractions and decimals. For example, suppose there are five apples. The weight of the apples can be 3.323 kg. Some more examples of continuous data can be height of a tree, time taken to finish a race, dimensions of a mobile phone, etc.

### 1.3.6 QUALITATIVE DATA

Unlike quantitative data, qualitative data cannot be measured. Some examples of qualitative data are colour of the sky, softness of your voice, etc. The qualitative data is collected through the process of observation. The qualitative data cannot be stated in numbers. For example, consider a student giving a presentation in a class. The teacher has provided feedback to the student on the basis of his/her fluency, clarity in pronunciation and loudness of voice. This type of data cannot be measured in numbers. Qualitative data is all about the perception of people or what they feel about a particular thing. It is mainly related to the emotions of people. In case of qualitative data, the emotions of people are documented. Qualitative data is also important like quantitative data. For example, a market researcher collects qualitative data about products, such as issues people are facing with the product, good points about the product, suggestions about the product, etc., to make the product better. Qualitative data can be of three types:

- **Nominal:** The qualitative data that can be grouped into categories is known as nominal data. The nominal data can be grouped on the basis of gender of a person, such as male or female. It can also be grouped into classes, such as normal weight, overweight, obese, etc. These categories are countable, but their order is not fixed.
- **Ordinal:** The qualitative data that can be counted and ordered is known as ordinal data. The ordinal data mainly includes rankings, such as 1 = Excellent; 2 = Good; 3 = Fair.
- **Dichotomous:** The qualitative data that has precisely two distinct values is known as dichotomous. For example, alive/dead, high/low, sick/well, etc.

**Self Assessment Questions**

4. _____ data source provides structured or organised data that originates from within the enterprise and helps run business.
5. _____ data can be defined as the data that has a defined repeating pattern.
6. Which of the following is a source of unstructured data?
   a. Text both internal and external to an organisation
   b. Social media
   c. Mobile data
   d. All of these

**Activity**

Explore how banks use large collection of data of their customers for enhancing profitability. Prepare a report on the basis of your findings.

Notes

## 1.4 HOW TO MANAGE DATA?

An information system is based on identifying hidden patterns of data, a valuable resource, to explore information that is necessary for effective decision-making in an organisation. With the help of this information, the organisation keeps updating itself to remain competitive and prepare its growth path. With the help of data, organisations not only set their objectives, but also ensure control to attain them. Data is one of the main elements of any information system. Data can be collected from internal as well as external sources. Data helps users in effective decision making. Therefore, procurement, collection and preservation of data are very crucial for the success of an organisation. Data administration and database management help an organisation in effectively managing its data resources.

Administration of data involves monitoring of data resource requirement and its procurement. As data is a valuable asset for the organisation, proper follow-up of its availability and storage is required by data administrators. Data misuse can also be checked through data administration. The scope of data administration includes:

- Monitoring the creation of data
- Administrating the usage of data
- Matching with the industry standards in terms of procurement and utilisation of data
- Maintenance of data warehouse and storage
- Maintenance of data quality

### 1.4.1 DATABASE

Database refers to the collection of data elements in a logical and integrated manner. This collection of data forms the basis for data storage and access for information processing. This pool provides data for many business applications as and when required. In this way, database concept aims at the following things:

- Fulfilling data storage requirement
- Fulfilling data structuring requirement
- Providing bases for data organising
- Fulfilling data retrieval requirement
- Providing data processing base
- Effectively supporting information system process

The development in technology has contributed in changing forms of database. Let us understand the major categories of database:

- **Operational database:** Supports the data requirement in the operations of the whole organisation. An operational database stores detailed data necessary for operational support. These are also known as the Subject Area Databases (SADBs),

transaction databases and production databases. Examples of such databases include customer databases, personal databases, inventory databases, accounting databases, etc.

- **Analytical database:** Supports the summarised data and information which is mostly required by the management of an organisation and end users. It stores data and information from select operational databases. These are also known as multidimensional, management or information databases.
- **Distributed database:** Includes databases of local work-groups, departments at regional offices, branch offices, manufacturing plants and other work sites. This database consists of segments of both operational as well as analytical databases. It also includes data generated and used only at a user's own site. This database resides at network servers, such as the Internet, intranet and extranet. It aims at continuously updating an organisation's database.
- **End users' database:** Refers to the data files developed by end users at their workstations.
- **Hypermedia database on the Web:** Refers to the set of interconnected multimedia pages on a website. It consists of hyperlinked pages of multimedia or mixed media, such as text, graphic, photographic images, video and audio. Such database stores data on the websites consisting of hyperlinked pages of critical subjects included.
- **External database:** Provides access to external, privately owned online data which is available for a fee to end users and organisations from commercial services.
- **Navigational database:** Involves queries that find objects primarily by following references from other objects. Traditionally, navigational interfaces are procedural, though one could characterise some modern systems like XPath as being simultaneously navigational and declarative.
- **In-memory database:** Refers to such database that depends mainly on the main memory for computer data storage, unlike database management system which utilises a disk-based storage mechanism. It is also known as the main memory database. The main memory databases are faster than disk-optimised databases since the internal optimisation algorithms are simpler and execute fewer CPU instructions. Accessing data in memory provides faster and more predictable performance than in the disk. The main memory databases are used in applications where response time is critical, such as telecommunications network equipment that operates emergency systems.
- **Document-oriented database:** Refers to computer programs designed for document-oriented applications. These systems may be implemented as a layer above a relational database or an object database. Document-based databases do not store data in tables with uniform-sized fields for each record. But they store each record as a document that has certain characteristics. Any number of fields of any length can be added to a document. Fields can also contain multiple pieces of data.

- **Real-time database:** Refers to a processing system that is designed to handle workloads which change constantly. This differs from traditional databases containing continual data which remain unaffected by time. For example, a stock market changes rapidly and dynamically. Real-time processing means that a transaction is processed fast enough for the result to come back and be acted on right away. Real-time databases are useful for accounting, banking, law, medical records, multimedia, process control, reservation systems, and scientific data analysis. As computers are getting upgraded in terms of power and storage capacity, the real-time databases are getting integrated into society and getting employed in many applications.
- **Relational database:** Involves tables and graphs to structure information so that it can be readily and easily searched through. The relational databases are commonly used databases today.

### 1.4.2 DATABASE MANAGEMENT SYSTEM (DBMS)

The application that controls the creation, maintenance and use of a database is known as Database Management System (DBMS). In DBMS, data is stored centrally, which allows users to access easily and share data as a common resource. Some popular DBMS programs are Microsoft Access (MS Access), Microsoft SQL Server, Oracle, MySQL and Open Office Base. Some applications of DBMS are as follows:

- **Banks:** DBMS is used to store the accounts information and all types of transactions.
- **Universities:** DBMS is used to keep track of students, their registrations and grades.
- **Organisations:** DBMS is used to keep track of employees, their salaries, products, sales and purchases.
- **Airlines and Railways:** DBMS is used for reservations and schedules.

#### Need for DBMS

Suppose that one of the students living in the hostel informs the hostel warden to update his/her home address and contact number. The hostel warden informs the school authority and hostel authority about the change. The hostel authority makes the required changes in the student records, but the school authority forgets to implement the changes. In this scenario, two files have different data pertaining to the same student. It leads to data inconsistency. This problem can be solved easily by using DBMS as there is no need to update separate files. DBMS allows you to store data at a central location wherein modification can be made easily. The need for DBMS can also be understood on the basis of the following parameters:

- **Storage:** DBMS stores large amounts of data. Therefore, you can add more data to it in comparison to the traditional file processing system.
- **Sorting:** DBMS sorts the data into an organised manner.
- **Summarising:** DBMS allows you to summarise the data. Summarising refers to the procedure of retrieving the summary of the data based on some defined criteria.

- **Classifying:** DBMS classifies the data depending on the requirements. For example, in a bank, DBMS categorises the accounts in different categories, such as savings, current and salary accounts.
- **Retrieving:** DBMS allows you to fetch the data instantly from the database.

Retrieving is the process of fetching information from the database.

### Advantages of DBMS

Earlier, data was stored in a traditional file processing system. However, in this system, finding the relevant information was difficult. Also, there was no effective method to control data redundancy and data inconsistency. DBMS was introduced to overcome these problems. Let us discuss some advantages of the DBMS over the traditional file processing system.

- **DBMS reduces data redundancy:** Duplication or repetition of data is known as data redundancy. As DBMS stores the data at a central location, it is easy to access or modify the data. Also, any changes made in the data are reflected automatically and made available to all the users. As the changes are made at only one location, chances of data redundancy are greatly reduced.
- **DBMS reduces data inconsistency:** Data inconsistency occurs due to repetition of data (data redundancy). For example, you have two copies of the same record stored at different locations. A change has been made in the first copy with the other being kept intact. This leads to data inconsistency. In this case, it becomes difficult to know which record is accurate. DBMS helps you reduce data inconsistency as the entire data is stored at one central location.
- **DBMS allows data sharing:** The facility to use or share the same data or data resource with multiple users is known as data sharing. In DBMS, all data is stored at a central location from where users can easily access it simultaneously.
- **DBMS enforces database standards:** DBMS ensures that the data stored in it follows certain applicable standards. These standards are set by an organisation or a person who has created the database. Setting such standards helps in transferring the data from one system to another.
- **DBMS ensures data security:** The data that is stored in a database may contain valuable or sensitive information. The DBMS provides security to data by ensuring that only authorised users are able to access the database. In other words, authorised users have the rights or permission to access or change the database. This not only helps keep track of the users who access and make changes in the database, but also prevents others without right credentials to access the database.
- **Back-up and recovery:** DBMS provides the back-up and recovery facilities to protect data from hardware or software failures. The back-up is the copy of data that can be used for future references for the purpose of recovery of the database.

### Components of DBMS

The components of DBMS include—DBMS engine, data definition subsystem, data manipulation subsystem, application generation subsystem and data administration subsystem.

Notes

The main components of DBMS are as shown in Figure 2:

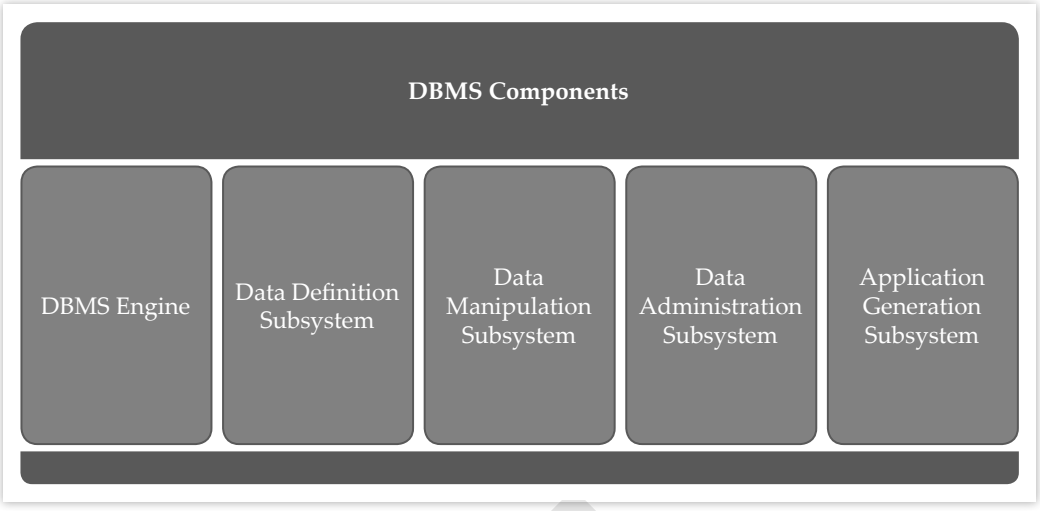



Figure 2: Components of DBMS

As shown in Figure 2, a brief discussion about these components is as follows:

- **DBMS engine:** Involves logical acceptance of commands from different DBMS users. It converts the commands into physical view to understand them. It can access database and data dictionary as these are available in the storage device.
- **Data definition subsystem:** Helps user create and maintain the data dictionary and define the structure of the files in a database.
- **Data manipulation subsystem:** Helps user add, change and delete information in a database and question it for gathering valuable information. Software tools within the data manipulation subsystem are most often the primary interface between user and information contained in a database. It allows user to specify its logical information requirements.
- **Data administration subsystem:** Helps users manage overall database environment by providing facilities for backup and recovery, security management, query optimisation, concurrency control and change management.
- **Application generation subsystem:** Contains facilities to help users develop transaction-intensive applications. It usually requires that users perform a detailed series of tasks to process a transaction. It facilitates easy-to-use data entry screens, programming languages and interfaces.

### 1.4.3 TABLES, KEYS AND DATA TYPES

In a database, data is stored in multiple tables. A table stores data about a single entity, such as employee and student. The data is organised in the format of vertical columns and horizontal rows. This structure helps represent data in an easily understandable and readable format. In a table, **a row is also known as record**, which represents a complete set of information. **A record consists of fields where each field contains one type of information.** For example, an Employee_Details table may contain records having four fields: an Emp_Code field, an Emp_Name

field, an Emp_Address field, and an Emp_Salary field. In this table, the record of Kavita, an employee, contains all information about her, such as her employee code, name, address, and salary, are shown in Table 4:

**Table 4: Employee Table**

| Emp_Code | Emp_Name | Emp_Address | Emp_Salary (₹) |
|---|---|---|---|
| Emp001 | Devansh | Preet Vihar | 15000 |
| Emp002 | Kavita | Vivek Vihar | 18000 |
| Emp003 | Karthik | Pandav Nagar | 20000 |

As you can see, Table 4 consists of a number of components which are explained as follows:

- **Field:** Refers to the smallest unit of information in a table. Each field in a table is given a unique name and a data type**.** For example, in a table named Employee_ Details, which contains information about the employees of an organisation, you can have a field named Emp_Name, containing the names of employees. It should be noted that data types are predefined. Several data types are available in DBMS that allow you to enter different types of data into the table. Data types can be broadly categorised into the following types:
  - Text data types
  - Date and time data types
  - Numeral data types
  - Boolean data types
- **Primary key and foreign key:** Primary key refers to a key that helps us uniquely identify a record in a table. The primary key is used to avoid duplicate data. In other words, a column with a primary key will not contain duplicate information in any of its records. For example, in the Employee_ Details table, we can set the primary key in the Emp_Code column, which contains the employee code. In this case, duplicate values are not possible in the column. The primary key is also used to connect the information in one table with the information in other tables. The primary key, when used in this way, is known as foreign key in the referring table.
- **Record:** Refers to a row of data that represents a complete set of information in a table. For example, the record of Kavita in Table 4 contains all information about her, such as her employee code, name, address and salary.

A Relational Database Management System (RDBMS) is based on 12 rules of relational model by E.F. Codd and uses tables for representing data elements as well as the relationships between them. The tables are usually identified by a primary key; whereas, foreign keys are used to create relationships between the tables. The RDBMS also allows the manipulation of data using various relational operators, which are based on relational algebra. You can perform various database operations on relational database using SQL as the query language.

Notes

### 1.4.4 ENTITY-RELATIONSHIP (E-R) MODEL

The basic E-R model represents a data model that defines the meaning associated with the business data. Therefore, the organisation usually maps the meanings and interactions associated with each data entity to design a conceptual or logical schema. The E-R diagram is a high-level modelling tool that helps in defining the relationships that exist among various entities. An entity is the basic unit of the E-R diagram composed of a set of attributes, whereas an entity set is a collection of similar types of entities, which share similar attributes. Attributes contain some values that can be used to identify individual entity in a large entity set. In other words, an entity may uniquely be identified by the values that a particular set of attributes takes. For example, every employee working in an organisation possesses unique employee IDs. In such a scenario, the employee ID acts as an attribute and the value it contains can be used to uniquely identify the employee. Therefore, attributes can be defined as a set of values that may uniquely identify each individual entity present in an entity set. An E-R diagram is basically composed of the following components:

- **Rectangles:** Represent entity sets
- **Ellipses:** Represent attributes
- **Diamonds:** Exhibit relationship sets
- **Lines:** Act as a link that is used to establish a connection between attributes and entity sets, as well as between entity sets and relationship sets
- **Double ellipses:** Represent multi-valued attributes
- **Dashed ellipses:** Represent derived attributes

A simple E-R diagram is shown in Figure 3:

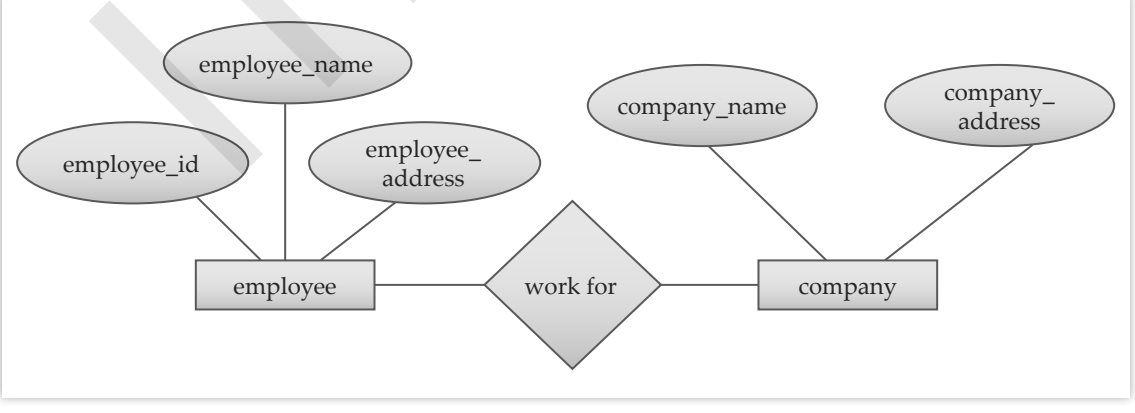



Figure 3: A Simple E-R Diagram

### 1.4.5 BRIEF INTRODUCTION TO SQL

Structured Query Language (SQL) is the heart of Relational Database Management System (RDBMS). It is the language used to perform all the operations in a relational database. The operations that are required for handling a database are creating tables, inserting data in tables, updating data in tables, deleting data from tables and retrieving data from tables. In addition to these operations, sometimes we may also

require a control to access data in a multi-user database. In a multi-user database, users may want to secure their data from other users; therefore, adequate security should be provided.

When you issue a command to the RDBMS in SQL, the RDBMS interprets your command and takes the necessary actions. Figure 4 shows how SQL commands are interpreted by the RDBMS:


Figure 4: Displaying the Role of SQL in RDBMS

SQL is a very powerful language and most of the RDBMSs use SQL as their language for specifying operations. The syntax of SQL changes very little from one RDBMS to another. Therefore, SQL for Oracle is similar to SQL for Ingress or Sybase.

An important feature of SQL is that it is a non-procedural language. In a non-procedural language, you have to describe what to do rather than how to do a job. How the job should be done is the responsibility of the underlying system, such as RDBMS. On the contrary, in the procedural way, you have to give the complete procedure of doing the job.

Suppose you ask your friend to reserve a railway ticket. The procedural way to do this is to give your friend complete stepwise instructions as follows:

- Take a bus
- Buy a ticket for traveling in a bus
- Go to the railway station
- Fill the reservation form
- Make the reservation
- Take a bus
- Buy a ticket for travelling in a bus
- Come back home

In this case, you are specifying a complete set of instructions. However, in a non-procedural way, you just need to specify your requirement to your friend as: make a railway reservation.

After you specify this, it becomes the responsibility of your friend to find out how to do it, i.e., how to go to the railway station and how to make the reservation.

Following are the advantages of using the non-procedural way:

- **Smaller code snippets:** Refers to the non-procedural way that allows you to use smaller code snippets to perform a task.
- **Simplicity:** Refers to the simplicity of code. In this method, you need not remember lengthy codes to perform a task.

Notes

- **Easy code writing:** Refers to the ease required in writing the code. This means that in a non-procedural way, you do not have to write complex programs to do the given task.
- **Reduction in the cost of maintaining software:** Reduces the cost of maintaining software. Though all the other advantages definitely reduce the time and effort you put in to perform a task, this is the most important advantage in real life.

**Self Assessment Questions**

7. A/An _____ system is based on identifying hidden patterns of data, a valuable resource, to explore information that is necessary for effective decision making in the organisation.
8. _____ database involves queries that find objects primarily by following references from other objects.
9. Real-time database refers to a processing system that is designed to handle workloads which change constantly. (True/False)
10. In a table, the intersection of a row and a column is known as _____.

**Activity**

Explore the concept of normalisation in database.

## 1.5 BUSINESS VIEW OF INFORMATION TECHNOLOGY APPLICATIONS

Information technology has changed the way we used to look at our world. Earlier, it was troublesome to do business from one country to another country predominantly because of distance issues.

Information technology has provided a solution to this mammoth problem by providing global access and usability of information through technology. Information technology has been defined as the use of technology for managing information. The task that earlier required days or even weeks to complete is a matter of just a few clicks today. The internet has undoubtedly facilitated real- time information sharing around the world that supports business operations worldwide.

According to the Information Technology Association of America (ITAA), *"IT (Information Technology) is the study, design, development, implementation, support or management of computer-based information system, particularly software applications and computer hardware."*

Business requires an interaction among stakeholders so that they can share information necessary to run the business effectively. Technology has brought revolutionary changes in the way business communication used to take place in the past. Earlier, sharing information between countries was a matter of one or more days, but now, it is possible in a few minutes.

IT has made information a scientific product and has materialised the field of Information Technology-Enabled Services, also known as ITES. Besides its

application in business, IT has contributed significantly in various other spheres, such as education, manufacturing, banking and telecommunication.

### 1.5.1 BUSINESS ORGANISATIONS AND THEIR FUNCTIONS

A business organisation can be defined as a group of individuals that is systematically structured to accomplish common goals and objectives. Organisations can be differentiated on the basis of following attributes:

- Structure
- Size
- Objectives
- Time frame
- Decision-making authority

Every organisation selects a unique path to achieve its long-term and short-term objectives. The vision statement of an organisation describes the objectives that an organisation aspires to achieve in the future; whereas, the mission statement defines the means to achieve the vision. Organisations conduct various activities or an activity in order to meet their objectives. These activities are known as business processes. A business process can be regarded as the core business process if it adds value to a product or service being provided by the organisation. These processes are related to the business strategy of an organisation and are important for its sustainability.

Business organisations have many functions, some of which are as follows:

- **Organising function:** It is one of the most crucial functions of management in which the primary resources, such as human, physical and financial resources, of an organisation are combined and synchronised together. The organising function of management is mainly concerned with the proper delegation of roles and responsibilities among individuals.
- **Financing function:** Finance is the backbone of any business. According to the size and nature of the business, proper capital structure is required to achieve organisational goals.
- **Production function:** The production function is also an important function in an organisation. In this function, the raw material is converted into finished or semi-finished products. The main objective of this function is to convert raw material into a product that suits customers' needs effectively.
- **Marketing function:** The marketing function also plays a vital role in business. After a successful production of the goods, it is needed to be transferred to the market from where the customer can buy the products easily. The tasks involved in this function are market information, customer demand, risk analysis, storage, buying-selling, transportation, etc.
- **Employment function:** In a business, an appropriate number of manpower is required for different departments. An organisation hired different types of people, mainly based on their skill level for performing technical and non-technical tasks in the organisation.

Notes

### 1.5.2 KEY PURPOSE OF USING IT IN BUSINESS

IT has provided such a network of communication worldwide that geographic limitations are no longer barriers in the expansion of business. Ideas, documents, products and services, all can be shared equally, regardless of any distinction derived by countries. Sharing information in the digital form has redesigned the world economy. The key purposes of using IT in business are as follows:

- For online advertising
- For gaining a competitive advantage
- For enhancing organisational efficiency
- For reaching more markets
- For knowing faster marketing response
- For better customer interaction and satisfaction

Technologies of fibre optics, microprocessors and telecommunications have contributed significantly in changing the way of organisational functioning. Fibre optics is the technology that supports the capturing of data including voices in digital forms to be changed into optical form as minute pulses of light. After this conversion, these pulses can be transmitted speedily. This technology is used in telecommunication network development.

The microprocessors are capable of developing brain-like structure in a system. These have continuously been updated in terms of its capacity. They have been successfully installed in a number of products of consumer durables, such as washing machines, cars and refrigerators.

Telecommunication devices have changed the way people used to interact with each other. One of the amazing factors of this system of information sharing is that the share of one person does not cause loss of information sharing for the other person. While about ten years back, organisations hardly utilised networks for business, it has become a necessity as of today for every business. As another trend in the use of information technology, the cost of using this technology has been reducing continuously. All this has caused dissolution of barriers made by territorial boundaries in the business world.

Such breakage of territorial constraints in business has emerged in the form of multinational corporations in different countries. According to the definition given by International Labour Organization (ILO), an MNC is *"a corporation that has its management headquarters in one country, known as the home country, and operates in several other countries, known as host countries"*.

As an evident example, while Mr. X wanted to book one ticket from US to China sitting in UK, he got assistance from a call centre executive in India. This example also clarifies that technology has unified the world business operations. The famous economist, Richard R. Nelson has pointed out in his work 'The Sources of Economic

Growth', "*We believe that the internationalization of trade, business and technology is here to stay. This means that national borders mean much less than they used to be regarding the flow of technology*." Among the examples of internationally established organisations, we have the name of business giants, such as Nokia, Microsoft, Accenture, HP, and IBM.

### 1.5.3 CHARACTERISTICS OF THE INTERNET-READY IT APPLICATIONS

The Internet-ready IT applications are also known as web-based applications. These kinds of applications are accessed using Web browsers (Internet Explorer, Google Chrome, Mozilla Firefox, etc.) present in your computer. The most common example of the Internet ready application is website. Some of the characteristics of Internet-Ready IT applications are as follows:

- These applications are capable of fulfilling the diversified requirements of a large number of users.
- These applications are accessible on various devices, such as mobile phones, tablets, laptops, etc.
- These applications are deployed on servers which are nothing but computers having large storage space and high processing capability.
- These applications use special authentication and authorisation mechanisms to verify users.
- These applications can be run on any operating system.
- These applications have the ability to get connected to any RDBMS for data storage.
- These applications are developed and implemented using multiple programming languages or their combinations as well.
- These applications are capable of supporting extensive connectivity to various types of networks and the internet services.

### 1.5.4 ENTERPRISE APPLICATIONS (ERP/CRM/SCM)

Enterprise Applications (EA) refer to a software solution that uses business logic and tools for modelling the entire business processes in organisations to enhance productivity and efficiency. Some commonly used enterprise applications are Enterprise Resource Planning (ERP), Supply Chain Management (SCM), Customer Relationship Management (CRM), etc.

An ERP system integrates all the core business processes, such as purchase management, inventory management, production and distribution management, human resource, finance, and sales. It produces a common database for all these business processes. An example of ERP system is SAP. Following are the major benefits of an ERP system:

- Supports the organisation with an integrated view of business processes.
- Provides quality information to support all the business processes.

- Provides enterprise-wide information to managers for effective decision-making.
- Supports organisational change and provides flexibility of business processes.

Many renowned organisations, such as Microsoft, Cisco and Coca-Cola, have adopted the ERP system to increase operational effectiveness.

Like ERP, an SCM system is also a cross-functional information system that is able to adjoin business processes from an organisation to suppliers on the one hand and to customers on the other hand. It helps in planning, organising, directing and controlling all the activities related to the procurement of the required materials in an organisation. The system also helps in coordinating and integrating the links involved in this process. These links include suppliers, transporters, wholesalers, retailers, and customers.

The SCM system coordinates the activities involved in the procurement of raw materials utilised in the production of finished goods. Supply chain management is a tactical task for every organisation since it has a direct effect on the efficiency of the business. Therefore, an organisation should have an effective SCM system. An example of SCM system is Epicor. The SCM system helps an organisation in:

- Providing effective and timely feedback and control
- Avoiding unnecessary delays in product delivery
- Preparing effective production planning
- Establishing effective delivery pattern
- Maintaining good relations with retailers, wholesalers, and suppliers

As we know, the success of an organisation directly depends on the level of customer satisfaction. Therefore, it is important for an organisation to build rapport with the customers, identify their needs and expectations, and fulfill them. Maintaining successful relationships with the customers helps organisations attract new prospects and gain market share. A CRM is a tool that helps in determining mutually satisfying goals between the organisation and its customers. The CRM system provides information related to the existing and potential customers of an organisation. It also provides information related to customers' preferences and their feedback and facilitates customer interaction. An example of CRM system is Salesforce.

### 1.5.5 INFORMATION USERS AND THEIR REQUIREMENTS

Information System (IS) has become more efficient and gives extremely high productive results to end users or information users. Essentially, an IS provides the necessary information for business operations to information users after processing data. The main requirement of information users is to acquire meaningful information so that they can reach to a certain conclusion.

Essentially, the IS is a set of several components that process data to derive information. This information supports decision-making and helps in performing control on various organisational departments and their functions. The IS can

be useful for individual departments and collectively for all departments. This interrelationship between the components of the IS is shown in Figure 5:

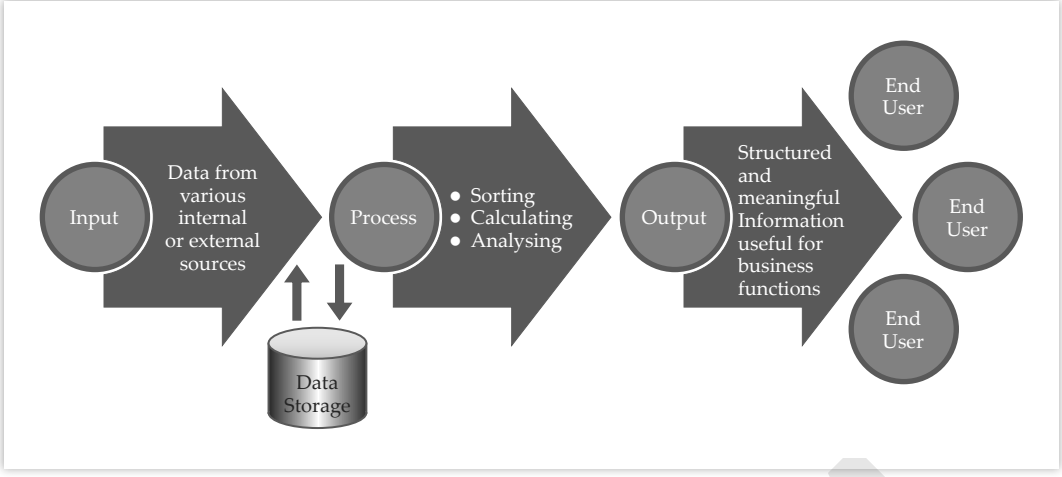



FIGURE 5: DISPLAYING THE IS MODEL

The data for an IS can be collected from various inter-organisation and intra-organisation sources, such as employees' personal details and competitors' performance. This data can be stored in paper or electronic format. The processing of this data includes calculations, logical analysis, and other statistical methods, depending upon the information to be derived. Once data is processed, it is transformed into information. This information is disseminated to various end users or information users to support their decision making, problem solving, strategy forming and controlling functions in an organisation.

For example, data about individual sales by each sales associate in a month, the general sales target for each associate and market sales trends for that month can be put to use. This data can be sorted, classified, calculated, and analysed to produce information. This will provide information about the sales trends for that time period, gaps between target and actual sales for individual associates, and aggregate sales based on teams. The same data can be utilised to predict the next month's sales, target setting for the next month, strategy building for future sales and the expected production level.

## SELF ASSESSMENT QUESTIONS

11. Organisations can be differentiated on the basis of structure, size, objectives, timeframe, and decision-making authority. (True/False)
12. A _____ can be regarded as the core business process if it adds value to a product or service being provided by the organisation.
13. _____ is a tool that helps in determining mutually satisfying goals between the organisation and its customers.

## ACTIVITY

Search and explore various components of SQL.

Notes

## 1.6 BUSINESS INTELLIGENCE DEFINED

The decision-making process is one of the most important functions of the organisations. In today's highly competitive business environment, decision-making is one of the most important functions the organisations have to perform. In fact, taking timely and logical decisions plays a big role in establishing and sustaining successful businesses. Increasing complexities of business and information overload have compelled many business organisations to employ sophisticated tools to assist in decision-making. Business intelligence (BI) is one such tool that helps organisations in analysing information and taking effective business decisions.

It comprises numerous activities, such as identifying, collecting, extracting, cleaning, and analysing data. BI helps organisations optimise resource usage, improve performance, and do predictive analysis.

At present, BI is implemented in various types of industries, such as retailing, marketing, manufacturing, insurance, healthcare and clinical research, sports, defence, weather, and agriculture. Most of the BI tools are available in the form of ready-to-use software, such as Micro Strategy, SSAS Compare, Jaspersoft, and Tableau. These software enable business analysts and decision-makers to analyse data and reach a conclusion rather than depending on probabilities.

Although BI holds enormous potential, its real-world implementations are determined by technical and organisational challenges. In such a scenario, clean and consistent data become imperative for an organisation. Data for BI can be collected from different sources, such as point of sale (POS), Webstores, inventory, call centres, social media, and mobile data.

### 1.6.1 DEFINITIONS AND EXAMPLES IN BUSINESS INTELLIGENCE

In 1958, IBM researcher, Hans Peter Luhn defined BI as *the ability to apprehend the interrelationships of presenting facts in such a way as to guide action towards a desired goal*. Moving ahead, in 1989, Future Gartner Group Analyst, Howard Dresner defined BI as *the concepts and methods to improve business decision-making by using fact-based support systems"*.

Let us consider a few examples of some organisations that have benefitted from BI implementation.

Bituach Haklai, an insurance company from Israel, had successfully implemented InsFocus BI, an insurance BI system by InsFocus. The general manager of Bituach Haklai, Mr. Arieh Herman, states that *"InsFocus BI had been tested and found to be the most suitable system for handling our company's needs at Bituach Haklai, both today and in the future to come. This system will allow us to continue to maintain our grasp in the market, improve our profitability and bring up-to-date and accurate data to the decision-makers in the company, which will improve and streamline our organisation"*.

Rubio's Restaurants, Inc. is a chain of restaurants from Mexico, which extends across six western states from California to Colorado. The director of IT for Rubio's Restaurants, Paul Nishiyama, says that *BI has been extremely important to us. Finance is getting a whole series of more robust reports that it did not have before. Producing those*

*reports without a business-intelligence system would be a manual process that would drain our small staff.*

Chase-Pitkin Home & Garden is a chain of hardware and garden-supply stores in upstate New York. The business was facing a problem of goods disappearance, that were ordered from suppliers, but did not reflect sales in the cash registers. So, the company decided to resolve this 'shrink' issue by using BI. For the last seven years, the company had been using SPSS Inc. Showcase Essbase, a data- presentation tool for collecting point-of-sale data on all store items. This task was executed by the CIO, T. Christopher Dorsey. Two-and-a-half years ago, Dorsey implemented SPSS's Showcase Analyser, a BI tool, to dig into that data. The Showcase Analyser surprised Dorsey by disclosing 15 store names and 16 items that disappeared and contributed to almost half of the chain's losses. Dorsey says, *When we discovered this, we put policies and procedures in place to halt the losses*. These measures generated $200,000 as savings in the first year.

The preceding examples show that the major benefit of BI is to help an organisation in taking better decisions, which can have a major impact on the organisational performance.

## 1.6.2 INTRODUCTION TO DATA MINING, ANALYTICS, MACHINE LEARNING, AND DATA SCIENCE

Data mining in the information system is like the mining of the earth. While the hidden valuables in the earth can be found through the mining process, data mining not only finds out, but also provides analysis of hidden patterns of data in a data warehouse. Data mining aims at exploring knowledge from the data warehouses. Data warehouse organises data in such a manner that it can derive the inherent meaning to contribute in the knowledge base. This process is represented by Figure 6:

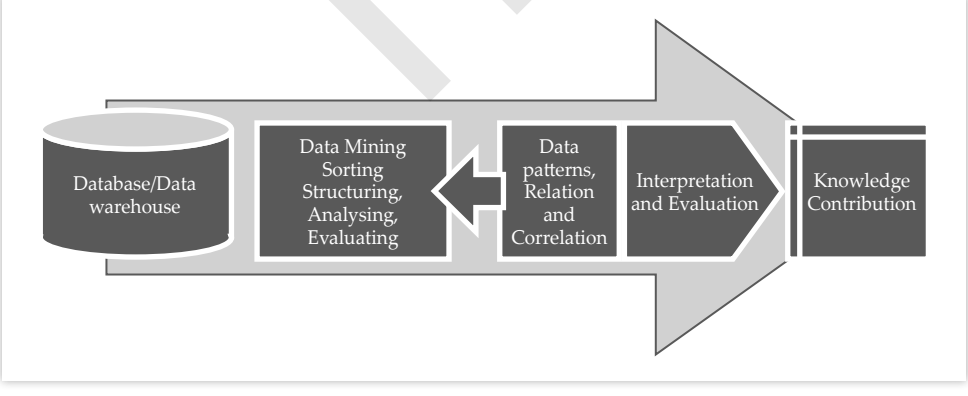



Figure 6: Data Mining Structure

As shown in Figure 6, the data from a database or a data warehouse is sorted to prepare the target data and then analysed. Data mining has its application in almost every business area. For example, from a marketing database, a data warehouse can explore the consumer behaviour which will help in preparing marketing strategies and product promotion. In this way, advertising department is benefitted with data mining.

Notes

Data without any analysis or interpretation is useless. At the same time, there is no 'easy' button to generate a great analysis in one simple step. Analysing statistics, writing a report, and using a modeling algorithm are each a combined step of several steps required to obtain a great analysis. Not understanding the aim of the analysis and focussing only on these individual tasks without keeping the end goal or problem to be solved in mind can lead to wrong decisions and generate a lot of extra work. Therefore, instead of working on individual tasks, the data should be analysed thoroughly by keeping the goal or objective in mind.

For example, sales data can be analysed to decide whether or not a product should be sold in a particular month based on the product's sales trend in that month. Generally, organisations require the information that comes after analysing the data. As you know, data is available in structured and unstructured formats. So, it becomes a challenge for organisations to perform data analysis. The data of different formats is collected, combined and then analysed.

The importance of data and its processing leading to information demanded for proper analysing and decision making. The humans started developing algorithms and programs so as to analyse big chunks of data with ease and take proper decisions based on the inferences. There came a need to innovate a process where the computing systems started learning on their own based on the data and processing results without any human intervention. This process is known as Machine Learning (ML). We are well aware of the fact that machines can outperform humans when it comes to scientific calculations and numerical processing. Thus, the need arose where machines were programmed in such a manner to learn from their previous data processing experiences and progress further so as to make it easy for the human beings to make proper decisions and increase profitability.

Another field using which the valuable insights of large amount of data can be extracted is known as data science. Nowadays, huge volumes of data are generated from various sources known as Big Data. Once the data is generated, it is very essential to structure it in a systematic process and store it properly. This demand of structuring data in a systematic process led to the evolution of various analytical tools, such as R-programming, Spark, Python, SaaS, etc. These days, data science is viewed as an umbrella term at the intersection of machine learning, artificial intelligence, data mining, computing, and statistics in the context of all forms of data including Big Data.

### 1.6.3 EVOLUTION OF BI

Business intelligence as a concept has evolved over the last few years. Earlier, the organisations used several software platforms for BI, such as Micro strategy, Hyperion, Cognos, etc. These BI environments were implemented on a large scale to deliver a wide range of standard reports. They share a number of characteristics between almost every organisation, such as the configuration, maintenance, and management of software by IT, generation of the logic behind the reports using standard SQL constructs and flexibility for few customers to create or customise their own reports. Though there was nothing wrong with this system, but with the evolution of BI, its scope also has expanded.

The scope of BI has also increased over the years, as it includes not just the traditional reporting and analytics tools, but also a wide range of functions and features including predictive modelling, business performance measurement, new Extraction Transformation and Loading (ETL) processes, budgeting and forecasting, and new interfaces.

### 1.6.4 MIS, DSS, EIS, AND DIGITAL DASHBOARDS

A management support system (MSS) provides useful information to managers for decision-making and control. Following are the different types of MSS:

- **Management Information System (MIS):** Provides information on various business aspects to managers. It generates information for monitoring performance and maintaining coordination. For example, a production manager can check the report of cost and time of production for the previous year, so that he/she can take effective production decisions for the current year.
- **Decision Support System (DSS):** Supports managerial decision-making. For example, a sales manager can set sales targets for the coming year by considering the current market conditions.
- **Executive Information System (EIS):** Provides critical information, such as market trends, sales, ratings of products, etc., to the executives and top-level managers for making strategic decisions. It provides statistical representation of information. For example, an executive can check for a brand image, by looking at the graphs representing customers' preferences and the sales of competitive brands in the last five years.

Another important decision-making tool for managers is dashboard. This tool collects data from data warehouses and other data sources, and represents the output through simple visual aids, such as graphs, charts and tables on a single screen, as shown in Figure 7:

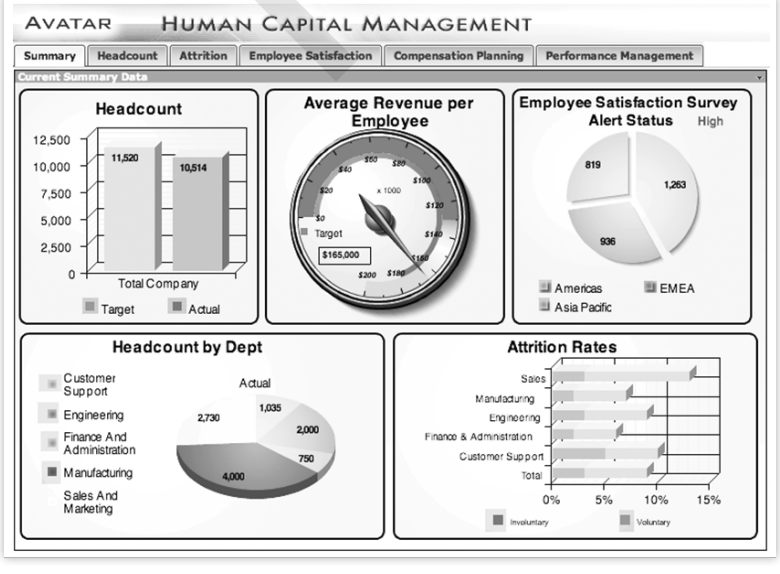



Figure 7: Dashboard Design

**Source:** *http://businessintelligencetalk.blogspot.in/2010/04/executive-dashboard_16.html*

Notes

The information presented in such a way is easy to understand and monitor, and is also actionable. Dashboards provide an insight into various KPIs of an organisation. This kind of information is very useful for the organisation in strategic decision-making. The information is updated regularly to keep it in sync with the real-time data.

Digital dashboards were introduced by Microsoft in 1991. These dashboards are available in distributed desktop applications or on the organisation's Intranet site. Dashboards that have the functionality to meet BI requirements are known as business intelligence dashboards or enterprise dashboards. Some examples of business intelligence dashboard vendors are Microsoft, IBM, and Oracle. Microsoft's Digital Dashboard tool includes Web-based elements (such as news and stock quotes) and corporate elements (such as e-mail applications) in Outlook.

### 1.6.5 NEED FOR BI

BI is the art of making decisions based on information, knowledge and experience. With the advancement and involvement of computers in our daily life, various computer-based techniques have improved the BI processes. The BI tools turn 'data' into 'information' and the 'information' further aids in taking 'decisions' on time. This results in data transparency, consistency, and information reliability.

The decision-making process requires evaluating performance (what happened), testing hypotheses (why things happened and exploring relationships) and predicting the future events (what may happen). This implies that the strategies of an organisation are sound and they meet the required purpose. In a nutshell, the BI system allows the user to answer the following questions:

1. What happened?
   - Did 'what happened' align with 'what you expected to happen'?
   - Assuming X produces Y, are you really executing X?
2. How did it happen?
   - How did X produce Y?
   - Can you be certain that X will produce Y, or is it Z that is actually producing Y?
3. Why did it happen?
   - Did X cause Y to happen?
   - If we execute X, then will Y happen?
4. What may happen?
   - If X occurs in the future, then will Y also occur?
   - Assuming you executed X and it produced Y, can you assume that continuing to do X will continue to produce Y?

An effective BI solution must:

- Present a holistic picture and an insight into tactical and strategic efforts
- Assist in fact-based decision-making

- Accept or reject assumptions
- Discover non-intuitive relationships
- Provide prompt feedback regarding actions

**Self Assessment Questions**

14. Data mining aims at exploring knowledge from the _____.
15. Management Information System (MIS) provides information on various business aspects to managers. (True/False)
16. _____ helps organisations optimise resource usage, improve performance and do predictive analysis.

**Activity**

Search and enlist the differences between scorecard and dashboards.

## 1.7 SUMMARY

- Data, simply put, is the raw material that does not make any definite sense unless you process it to any meaningful end.
- Information is the result that we achieve after the raw data is processed.
- Knowledge is something that is inferred from the data and information.
- Explicit knowledge and its offspring can be kept in a certain format, e.g., encyclopaedias and textbooks.
- The second type is termed as the tacit knowledge referring to the type of knowledge that is complex and intricate.
- Structured data can be defined as the data that has a defined repeating pattern. This pattern makes it easier for any program to sort, read, and process the data.
- Unstructured data is a set of data that might or might not have any logical or repeating patterns.
- Semi-structured data, also known as schema-less or self-describing structure, refers to a form of structured data that contains tags or markup elements in order to separate semantic elements and generate hierarchies of records and fields in the given data.
- Quantitative data refers to data that is related with quantities of real-world objects.
- Unlike quantitative data, qualitative data cannot be measured.
- An information system is based on identifying hidden patterns of data, a valuable resource, to explore information that is necessary for effective decision making in the organisation.
- Data administration and database management help an organisation in effectively managing its data resources.

Notes

- Database refers to the collection of data elements in a logical and integrated manner.
- The application that controls the creation, maintenance and use of a database is known as Database Management System (DBMS).
- Duplication or repetition of data is known as data redundancy.
- The facility to use or share the same data or data resource with multiple users is known as data sharing.
- DBMS provides the backup and recovery facilities to protect data from hardware or software failures.
- In a table, the intersection of a row and a column is known as a cell. In a table, a row is also known as a record which represents a complete set of information.
- A record consists of fields where each field contains one type of information.
- Entity-Relationship (E-R) modelling concept was introduced to facilitate database designing by specifying an organisation schema which defines the logical structure of the database.
- Structured Query Language (SQL) is the heart of Relational Database Management System (RDBMS).
- Business intelligence is an umbrella term that covers the skills, processes, methods, technologies, applications, and practices that are used to collect, process, and analyse heterogeneous and distributed data.

## 1.8 KEY WORDS

- **Data redundancy:** It refers to duplication or repetition of data.
- **Database Management System (DBMS):** The application that controls the creation, maintenance and use of a database.
- **Database:** It refers to the collection of data elements in a logical and integrated manner.
- **End Users' Database:** It refers to the data files developed by end users at their workstations.
- **Field:** It refers to the smallest unit of information in a table.
- **Knowledge:** It signifies a design that links and usually provides a high-level view and likelihood of what will happen next or what is described.
- **Primary key:** It refers to a key that helps us uniquely identify a record in a table. The primary key is used to avoid duplicate data.
- **Record:** It refers to a row of data that represents a complete set of information in a table.
- **Structured data:** It can be defined as the data that has a defined repeating pattern.
- **Unstructured data:** It is a set of data that might or might not have any logical or repeating patterns.

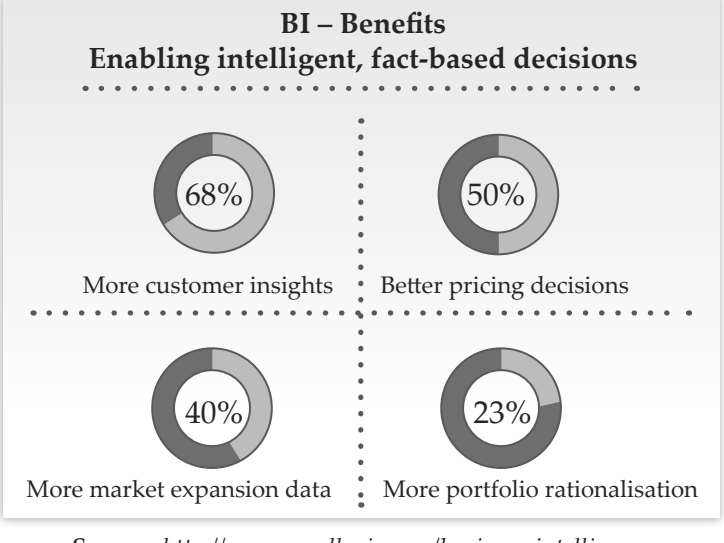
## 1.9 CASE STUDY: BENEFITS OF BUSINESS INTELLIGENCE IN RETAIL INDUSTRY

In today's business environment, a lot of data is produced on a daily basis. This is true for all business industries. In case of the retail industry, the data produced relates to emerging trends and sales, changes in the global market, etc. This data can be used effectively by gathering, measuring, and reporting on it. This data can be stored and analysed effectively by using a Business Intelligence system. The concerned retail outlet may choose to implement a traditional sales and trends analytics BI system or may use an advanced BI system that provides a predictive outlook.

The BI tools help retail outlets increase their sales by understanding customer needs, optimising prices and current trends, predicting upcoming trends, etc. It has been predicted that by 2020, the sales of e-Commerce retail will reach $4 trillion. A large number of retail brands are now increasingly using BI tools, such as Artificial Intelligence (AI) and customer service bots. Major challenges faced by retailers include tracking merchandise, organising shipping, managing suppliers, and identifying customer behaviour.

Fiverr is an e-Commerce shopping site that sells services. In order to boost its sales, they decided to use Sisense's BI tool. This tool helped them combine data from Google Docs, spreadsheets, and analytics to track events on their website and mobile app. With an increase in the user base, Fiverr's needs also increased. Fiverr required the BI tool to track the user behaviour in real time. Before implementing the BI tool, Fiverr was gathering user data from various sources. Now, Fiverr is using Sisense's BI tool to analyse data and to determine away forward. The use of BI tools helped Fiverr in creating and maintaining an agile environment that could support speedy and smart decision-making based on the user behaviour data.

E-Commerce retailers need to optimise their products and their prices based on the trends and customers' past purchases. Therefore, it becomes important to generate trend and price reports on a real-time basis using BI tools. Such practices help a retailer in improving its market positioning. This claim can be validated by a CIO study, according to which 19% of executives claimed that real-time data analysis helped them immensely.




**Source:** *http://www.quadlogix.com/business-intelligence*

Notes

Retailers also use BI for fulfilling their in-store needs and make decisions. Some of the e-Commerce platforms, such as Amazon, Flipkart, etc., offer a variety of products all over the world in multiple markets. In such circumstances, things often get complicated because these platforms need to gain an understanding of what products are selling and where they are selling. Also, since these platforms sell in multiple markets, these become harder to track.

Let us explain this situation by taking the example of Kargo Card. Kargo Card is a Chinese gift card company, which supplies gift card walls to convenience and retail stores in various Chinese cities. Since the number of locations they send their products to is quite large, it becomes difficult to track and measure metrics. They implemented a BI system to take care of this situation. Using the BI tools, it was able to put together partner data in Excel and CSV files to track inventory possession data. It means that Kargo Card could track the inventory of cards across the entire network. The brands of cards that were out of stock could also be identified. It helped Kargo Card in retaining its customers.

The use of BI helped Kargo Card in increasing its revenue by making effective decisions, restocking stores and re-ordering cards when required. Apart from this, key advantages of BI for the retail sector include improvement in customer experience, optimisation of prices, etc.

## QUESTIONS

1. What are the major challenges faced by retailers?

   (**Hint:** Major challenges faced by retailers include tracking merchandise, organising shipping, managing suppliers, and identifying customer behaviour.)

2. How do BI tools help retail outlets?

   (**Hint:** Retail outlets and platforms use BI tools in increasing their sales by understanding customer needs, optimising prices, current trends, predicting, and upcoming trends, etc.)

3. Why did Fiverr feel the need to use BI tools?

   (**Hint:** Fiverr required the BI tool to track the user behaviour in real time.)

4. Why has generation of trend and price reports on a real-time basis using BI tools become important for e-commerce retailers?

   (**Hint:** E-Commerce retailers need to optimise their products and their prices based on the trends and customers' past purchases. Such practices help a retailer in improving its market positioning.)

5. What is the purpose of using business intelligence system?

   (**Hint:** To store and analyse collected data effectively.)

## 1.10 EXERCISE

1. Discuss the linking between data, information, and knowledge.
2. What do you understand by structured, unstructured, and semi-structured data? Explain with suitable examples.
3. Explain the following components of a database:
   a. Table
   b. keys
   c. Fields
4. Describe the importance of ERP, CRM and SCM.
5. Elucidate the importance of business intelligence for an organisation.

## 1.11 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Data, Information, Knowledge and Wisdom | 1. | Knowledge |
| | 2. | a. Explicit |
| | 3. | True |
| Types of Data | 4. | Internal |
| | 5. | Structured |
| | 6. | d. All of these |
| How to Manage Data? | 7. | information |
| | 8. | Navigational |
| | 9. | False |
| | 10. | cell |
| Business View of Information Technology Applications | 11. | True |
| | 12. | business process |
| | 13. | Customer Relationship Management (CRM) |
| Business Intelligence Defined | 14. | data warehouses |
| | 15. | True |
| | 16. | Business Intelligence |

Notes

## 1.12 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Vercellis, C. (2013). *Business Intelligence*. Hoboken, N.J.: Wiley.
- Scheps, S. (2008). *Business Intelligence for Dummies*. Hoboken, N.J: Wiley.


### E-REFERENCES


- Search Business Analytics. (2018). *What is business intelligence (BI)?* – Definition *from WhatIs.com*. [online] Available at: https://searchbusinessanalytics. techtarget. com/definition/business-intelligence-BI [Accessed 24 Nov.2018].
- Pratt, M. (2018). *What is BI? Business intelligence strategies and solutions*. [online] CIO. Available at: https://www.cio.com/article/2439504/business- intelligence/ business-intelligence-definition-and-solutions.html [Accessed 24 Nov.2018].
- OLAP.com. (2018). *What is Business Intelligence? BI Definition*. [online] Available at: http://olap.com/learn-bi-olap/olap-bi-definitions/business- intelligence/[Accessed 24 Nov.2018].

CHAPTER 2

# Introduction to Business Analytics

## Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the types of business analytics
- Explain the relation between BI and BA
- Describe the business analytics in the context of strategy
- Discuss the importance of business analytics
- Elucidate the emerging trends in BI and BA

## 2.1 INTRODUCTION

In the previous chapter, you have studied about linking between data, information, and knowledge. The chapter also explained about the different types of data and its management. Further, the chapter has described about the impact of information technology in business and the concept of data analytics and business intelligence.

Business Analytics (BA) is a group of techniques and applications for storing, analysing, and making data accessible to help users take better strategic decisions. Business analytics is a subset of Business Intelligence (BI), which enhances companies' capabilities of contesting in the market efficiently and is likely to become one of the main functional areas in most companies.

Analytics companies develop the ability to support decisions through analytical perception. Analytics certainly influences the business by acquiring knowledge that can be helpful to make enhancements or bring changes. BA can be segregated into many branches on the basis of organisational domains, such as financial service analytics, fraud analytics, health care analytics, marketing analytics, retail sales analytics, pricing analytics, supply chain analytics, etc.

For example, a sales and advertising company uses marketing analytics to understand which marketing tactics and strategies click best with customers. With the performance data of the marketing branch in hand, business analytics became an essential way for measuring the overall impact on the organisation's revenue chart. These understandings direct the investments in areas like media, events, and digital campaigns. These allow us to understand customer results clearly, such as lifetime value, acquisition, profit and revenue driven by our marketing expenditure.

This chapter begins by explaining the types of business analytics that are generally used to aid businesses. Next, the chapter describes the importance of business analytics. Towards the end, the chapter discusses the business strategy analysis models.

## 2.2 TYPES OF BUSINESS ANALYTICS

Going purely by the linguistic definition, there may be multiple elucidations of the term 'business analytics'. However, in practical terms, there are four types of

business analytics that help an organisation in gauging customers' sentiments and then take respective actions:

- **Descriptive analytics:** It refers to 'what is happening?' or 'what happened?' type analytics that is based on the incoming data. Such analytics is better studied by dashboards and reports which provides summaries of data.
- **Diagnostic analytics:** It refers to the analysis of the past figures and facts to derive the scenarios about 'what happened' and 'why it happened'. The result of this analysis is often a predefined reporting structure, such as the root cause analysis (RCA) report. For example, in a hospital, a doctor monitors a patient's response and changes the medication according to the situation for a positive response.
- **Predictive analytics:** It tries to forecast on the basis of previous data and scenarios. For example, a hotel chain owner might ramp down promotional offers during a restive season of rains in a coastal area. This is based on the predictions that there are going to be fewer footfalls due to heavy rain.
- **Prescriptive analytics:** It guides you about the actions you should take. This is the most essential analysis and typically forms the standards and recommendations for the next phase. For example, a doctor prescribes medicines to the patient after researching, studying, evaluating, and diagnosing the cause of the problem of the patient. Similarly, after drawing out the resultants or conclusions, organisations will also take a step in ensuring that the factors affecting the growth charts positively continue to exist.

**SELF ASSESSMENT QUESTIONS**

1. Which type of analytics tries to forecast on the basis of previous data and scenarios?
   a. Descriptive analytics
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics
2. The _____ analytics is the study of the past figures and facts to derive the scenarios about what happened and why it happened.
3. In diagnostic analytics, RCA refers to _____.

**ACTIVITY**

How can business analytics bring a change for a newspaper hawker? Think it out.

## 2.3 RELATION BETWEEN BI AND BA

BA is a subset of BI. BI, at the root level, is the skill of converting business data into knowledge to aid the decision-making process. The conventional method of doing this includes logging and probing the data from past and using the overall outcome from the reading as the standard for setting future benchmarks.

BA emphasises on data usage to get new visions, while conventional BI uses a constant, recurring metric sets to drive strategies for future business on the basis of historical data.

Notes

With the help of BA, you get to know the pain points or vulnerabilities of your business, your product's standing in the market, your strengths related to business that put you ahead of the competition and the opportunity which you are yet to explore. BA helps you in knowing your business thoroughly. BI helps in bridging that gap between ground reality and management perspective.

BI helps you in aggregating your strong points, weeding out the weakness in an efficient manner and managing the organisational business more efficiently. It helps you capitalise on the lessons learned from the BA findings about the organisation. For example, by analysing the complaints of customers and their requests for refunds because of product delivery or quality issues, an e-commerce company can drop poor performing suppliers and ensure that their clients will not face these types of issues in future.

Table 1 shows some main points about BI and BA:

**Table 1: Main Points about BI and BA**

| BI | BA |
|---|---|
| Uses current and past data to optimise the current age performance for success. | Utilises the past data and separately analyses the current data with past data as reference to prepare the businesses for the future. |
| Informs about what happened. | Tells why it happened. |
| Tells you the sales numbers for the first quarter of a fiscal year or total number of new users signed up on our platform | Tells you about why your sales numbers tanked in the first quarter or the effectiveness of the newly launched user campaign |
| Quantifiable in nature, it can help you in measuring your business in visualisations, chartings and other data representation techniques. | More subjective and open to interpretations and prone to changes due to ripples in organisational or strategic structure. |
| Studies the past of a company and ponders over what could have been done better in order to have more control over the outcomes. | Predicts the future based on the learning gained from the past, present and projected business models for a given term in the near future. |

Another new trend is the skill to combine multiple data projects in one, while making it useful in sales, marketing and customer support. That concept is also called CRM – Customer Relationship Management software, which sources raw data from every division and department, compiles it for a new understanding that otherwise would not have been visible from one point alone.

All these boil down to the interchangeable usage of the term 'business intelligence' and 'business analytics' and its importance in managing the relationship between business managers and data. Owners and managers now, as a result of such accessibility, need to be more familiar with what data is capable of doing and how they need to actively produce data to create lucrative future returns. The significance of data has not changed, its availability has.

**Self Assessment Questions**

4. BI, at the root level, is the skill of converting business data into knowledge to aid the decision-making process. (True/False)

5. BA emphasises on data usage to get new visions, while conventional BI uses a constant, recurring metric sets to drive strategies for future business on the basis of the historical data. (True/False)
6. _____ is more subjective while _____ is quantifiable in nature.

## 2.4 ROLE OF BUSINESS MODELS IN ANALYTICS

Business analytics (BA) frequently utilises numerous quantitative tools to convert Big Data into meaningful information for making informed business decisions. These tools can be further categorised into tools for data mining, operations research, statistics, and simulation. Statistics for instance, can be helpful in gathering, articulating, and understanding Big Data as part of the descriptive analytical model. A BA model assists organisations in making a move which yields fruitful results. Here, we will discuss two most commonly used analytical models by the analysts across the globe as standard analysis factors–SWOT and PESTEL analysis.

### 2.4.1 SWOT ANALYSIS MODEL

SWOT stands for Strengths, Weaknesses, Opportunities, Threats. As evident from the abbreviation, an organisation uses SWOT analysis to figure out its greatest extremes–strengths to help it stand even in the toughest of times, weaknesses that may lead to its failure, opportunities that may help in realising its full potential and finally the threats to the businesses that may end up exploiting its weaknesses and may turn its strengths into weaknesses.

Figure 1 shows the SWOT diagram:

| **Strengths** | **Opportunities** |
|---|---|
| • What does your organisation do better than others?<br>• What are your unique selling points?<br>• What do your competitors and customers perceive as your strengths?<br>• What is your organisation's competitive edge? | • What political, economical, socio-cultural, or technological (PEST) changes are taking place that could be favourable to you?<br>• Which areas have gaps or unfulfilled demands?<br>• What new innovations can your organisation bring to the market? |
| **Weaknesses** | **Threats** |
| • What do other organisations do better than you?<br>• What elements of your business add little or no value?<br>• What do competitors and customers perceive as your weakness? | • What political, economical, socio-cultural or technological (PEST) changes are taking place that could be unfavourable to you?<br>• What restraints do you face?<br>• What is your competitor doing that could negatively impact you? |

FIGURE 1: SWOT Diagram

**Source:** *https://s-media-cache-ak0.pinimg.com/736x/88/b0/1a/88b01aa805648a30 4c0a3bbd954c1a5e.jpg*

Notes

Businesses that have been in market for long should conduct SWOT analysis periodically to evaluate the impact of the changing situations in the market, getting around the newer business models and respond actively. On the other hand, new starters should include SWOT as their planning process. SWOT is not necessarily a pan-organisation process; rather each of the organisation's departments can have their own dedicated SWOT, such as Marketing SWOT, Operational SWOT, Sales SWOT, etc.

Consider an example of the implementation of SWOT analysis in the organisation, Apple Inc. Apple was incorporated in 1995 after a long battle with the existing stakeholders who had control over the shares and stocks. Post return to the computing market, facing a mighty challenger in Microsoft, Apple did not take them head-on as most would have expected.

Rather, it realised the opportunities and laid back on the threats part since they had 'nothing to lose'. Apple identified opportunities in newer areas of the technology, while the world was considering computers as the lone IT revolution torch-bearer.

## 2.4.2 PESTEL OR PEST ANALYSIS MODEL

PESTEL stands for Political, Economic, Social, Technological, Legal and Environmental. PESTEL analysis is a method for figuring out external impacts on a business. In some countries, legal and environmental parts are combined in the social, legal, political, and economic parts. Hence, they use PEST. PESTEL analysis is an examination of the external environment in which an organisation currently exists or is going to enter. The sample PESTEL analysis is shown in Figure 2:

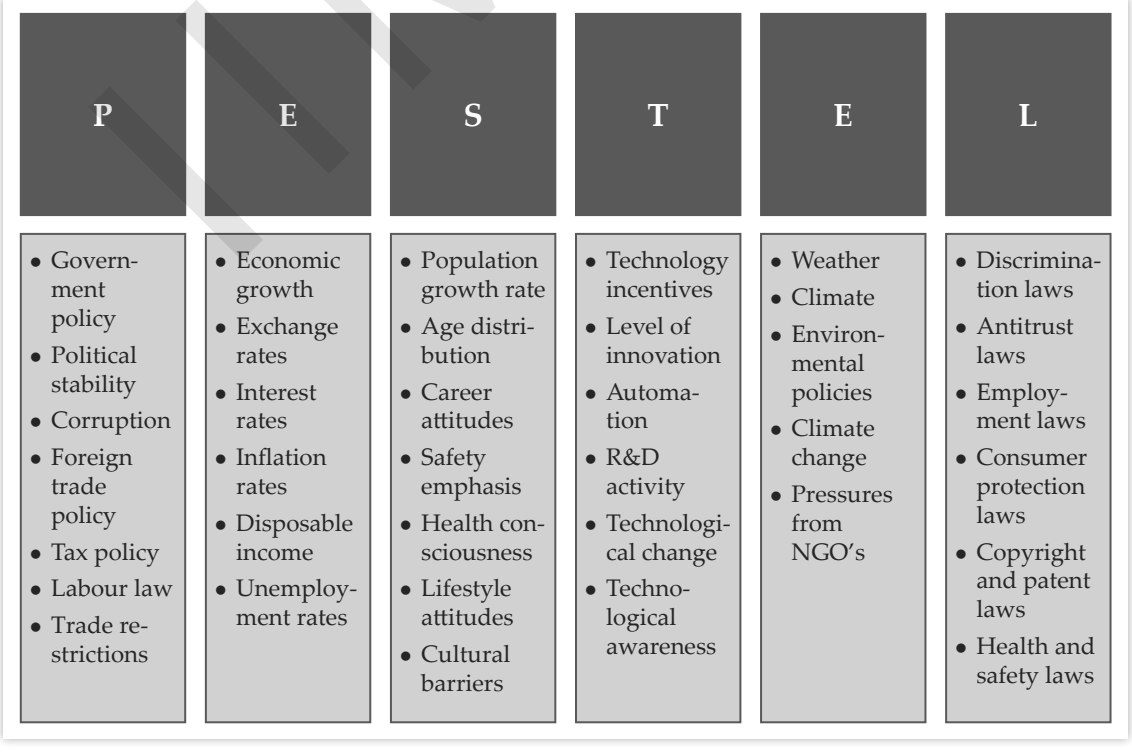



Figure 2: PESTEL Analysis

**Source:** *https://www.business-to-you.com/scanning-the-environment-pestel-analysis/*

The benefits of PESTEL analysis are as follows:

- **Political factors:** These are government regulations in different countries related to employment, tax, environment, trade, and government stability.
- **Economic factors:** These factors affect the purchasing power and cost of capital of a corporation, such as economic growth, inflation, currency exchange, and interest rates.
- **Social factors:** These influence the consumer's requirement and the possible market size for an organisation's products and services. These factors include age demographics, population growth, and healthcare.
- **Technological factors:** These influence the barricades to entry, investment decisions related to buying and innovation, such as investment incentives, automation and the adaptability quotient for the technology.
- **Environmental factors:** These influence mainly the marketers with respect to various environmental factors and policies of a specific country.
- **Legal factors:** These influence the business decisions of an organisation with respect to various legal factors, such as discrimination laws, anti-trust laws, employment laws, consumer protection laws etc., of a specific country.

Also, it is a point worth noticing that the six components of the PESTEL model vary in meaning on the basis of business type. For example, social factors are more important to a consumer-oriented business at the customer's side of the supply chain. On the other hand, political factors play their role more for an aerospace manufacturer or a defence contracting firm.

**Self Assessment Questions**

7. In the context of analytics, SWOT stands for _____.
   a. Strong, Weak, Opportunities, Threats
   b. Strengths, Weaknesses, Opportunities, Threats
   c. Strengths, Weak, Opportunities, Threats
   d. Strengths, Weaknesses, Openness, Theoretical
8. SWOT is often considered as a 180-degree tool to measure the pulse and vitals of an organisation. (True/False)
9. PESTEL stands for Political, Economic, Society, Technological, Environmental and Legal. (True/False)
10. It is ideal to complete a _____ analysis before SWOT.

**Activity**

Search a real-life example on PEST analysis.

## 2.5 IMPORTANCE OF BA

The need of analytics arises from our day-to-day life. An average person analyses the time factor right from getting up from the bed to getting ready for office. It also includes analysing the best possible route to avoid traffic and save more time in order to have an extra cup of coffee for the day! All this analysis and planning to make things smoother is even more applicable in a business. BA helps organisations understand leads, audience, prospects, and visitors. Moreover, it also helps in understanding, improving and tracking the method that can be used to impress a valuable customer.

The importance of BA is as follows:

- **To get visions about customers' behaviour:** The prime advantage of financing a BI software application in analytics is that it increases your skill to examine the present customer-purchasing trend. Once you know what your customers are ordering, you can create products matching the present consumption trends of customers, and, thus, improve your profitability since you now know how to attract more valued customers.
- **To improve visibility:** BA helps you in getting a better visibility of the processes and recognising any parts requiring a fix or improvement.
- **To convert data into worthy information:** BA uses the BI system as a logical tool that enables you to create result-oriented strategies for your corporation. Since such a system identifies patterns and key trends from your corporation's data, it makes it easier for you to connect dots between different points of your business that may seem otherwise disconnected. Such a system also helps you comprehend the inferences drawn from the multiple structural processes better and increase your skill to recognise the right opportunities for your organisation.
- **To improve efficiency:** One critical reason to consider a BI system in analytics is an increase in the efficacy of the organisation, thus leading to increased productivity. The BI helps in sharing information across multiple channels in the organisation, saving time on reporting analytics and processes. This ease of sharing information reduces redundancy of duties or roles within the organisation and improves the precision and practicality of the data produced by different divisions.

Consider a typical website that relies on visitor footfall and subsequent click-based advertising revenues. Such an organisation needs analytics more often than other organisations which have a dedicated business running in brick and mortar stores and which use their websites only for marketing purposes.

BA is an important area that equips you with correct weapons to take correct business decisions. BA arms you with situational arsons – you get a machine gun in the form of viral marketing campaigns when you are targeting a mass audience for a given product, whereas in case of customer withdrawal or ramp-up, you can have your sniper ready to specifically target them out.

From the preceding discussion, we can conclude that analytics play a key role in viewing performance, managing logistics and analysing production quality in real time.

**Self Assessment Questions**

11. BA does not help organisations understand leads, audience, prospects and visitors. (True/False)
12. BA helps you in getting a better visibility of the _____ and recognising any parts requiring a fix or improvement.

## 2.6 EMERGING TRENDS IN BI AND BA

Following are some contemporary trends in BI and BA fields:

- **More power and monetary impact for data analysts:** Analysts are consistently creating demand charts across many industries. All thanks to the demand-driven analytical bandwagon that has made the industry take cognizance of the data analysts and led to a spike in other roles, like Information Research Scientists and Computer Systems Analysts.
- **Location analytics:** Another major business driver in 2016 was related to location and geospatial analytical tools that gave organisations better market intelligence and placements in terms of effective campaigns; for example, a company aiming geocentric campaigns for specific customers.
- **Data at the rough edge:** Businesses must look beyond the usual sources of data besides their data centres since the data flows now initiate outside the data from multiple sensor devices, and servers—e.g., a spatial satellite or an oil rig in the sea.
- **Artificial Intelligence (AI):** This is a top trend as per multiple studies with scientists targeting to make machines or software think intelligently like humans. The analytical work on such programmes is exponentially growing with AI and machine learning transforming the way we relate with the analytics and data management.
- **Predictive analytics and impact on data discovery:** By gathering more information, organisations will have the capacity to build more detailed visual models that will help them act in more accurate ways. For instance, having better information models shows organisations more about what clients are purchasing, and even what they are possibly going to purchase in the future.
- **Cloud computing:** Cloud computing is a technique that makes it possible for organisations to dynamically regulate the use of computing resources and access them as per the need while paying only for those resources that are used. Cloud computing is being absorbed into many systems and will continue to grow. We have witnessed the division of cloud into multiple vendor systems and many companies are utilising cloud services to host the powerful data analytics tools. A lot of customers are already using Microsoft Azure and Amazon Redshift along with cloud resources that provide flexible handling and scalability for the data.
- **Digitisation:** It is a process of turning any analogue image, sound or video into a digital format that is understandable by the electronic devices and computers. The gains from digitising the data-intensive processes are great: with up to 90% cost cut and much faster turnaround times than before. Creating and utilising software over manual processes allows businesses to gather and screen the data in real time, which assists the managers to tackle issues before they turn critical.

Notes

**Self Assessment Questions**

13. Name the technique that makes it possible for organisations to dynamically regulate the use of computing resources and access them as per the need while paying only for those resources that are used.

    a. Business Analytics
    b. Business Intelligence
    c. Cloud Computing
    d. None of these

14. _____ is a process of turning any analogue image, sound or video into a digital format that is understandable by the electronic devices and computers.

15. Using _____, scientists make machines or software to think intelligently like humans.

**Activity**

Search and enlist the implications while implementing the cloud deployment model in an organisation.

## 2.7 SUMMARY

- Business analytics is a subset of business intelligence, which enhances companies' capabilities of contesting in the market efficiently and is likely to become one of the main functional areas in most companies.
- It refers to 'What is happening?' or 'What happened?' type analytics that is based on the incoming data. Such analytics is better studied by dashboards and reports which provides summaries of data.
- Diagnostic analysis refers to the analysis of the past figures and facts to derive the scenarios about what happened and why it happened.
- Predictive analysis refers to the analysis of probabilities. It tries to forecast on the basis of previous data and scenarios.
- Prescriptive analysis tells you about the actions you should take.
- BA uses the BI system as a logical tool that enables you to create result- oriented strategies for your corporation.
- One critical reason to consider a BI system is an increase in the efficacy of the organisation, thus leading to increased productivity.
- BI helps in sharing information across multiple channels in the organisation, saving time on reporting analytics and processes.
- BI at the root level is the skill of converting business data into knowledge to aid the decision-making process.
- BA emphasises on data usage to get new visions, while conventional BI is constant, recurring metric sets to drive strategies for future business on the basis of historical data.
- BI helps you in aggregating your strong points, weeding out the weakness in an efficient manner and managing the organisational business more efficiently.

- Artificial Intelligence (AI) is a top trend as per the multiple studies with scientists targeting to make machines or software think intelligently like humans.
- SWOT analysis is amongst the most popular method of gauging the organisational and corporate nerve of an organisation.
- PESTEL analysis is the method for figuring out external impacts on a business.

## 2.8 KEY WORDS

- **Business analytics:** It is a subset of business intelligence, which enhances companies' capabilities of contesting in the market efficiently and is likely to become one of the main functional areas in most companies.
- **Cloud computing:** It refers to a technique that makes it possible for organisations to dynamically regulate the use of computing resources and access them as per the need, while paying only for those resources that are used.
- **Descriptive analytics:** It refers to 'What is happening?' or 'What happened?' type analytics that is based on the incoming data and is better studied by dashboards and reports which provides summaries of data.
- **Digitisation:** It is a process of turning any analogue image, sound or video into a digital format understandable by the electronic devices and computers.
- **PESTEL analysis:** It is a method for figuring out external impacts on a business.
- **SWOT analysis:** It is a popular method of gauging the organisational and corporate nerve of an organisation.

## 2.9 CASE STUDY: ROLE OF BUSINESS ANALYTICS IN MID-SIZED ORGANISATIONS

Businesses today face the challenge of increasing revenues and reducing costs and the associated business risks amidst the increasingly volatile environment. Businesses need to be responsive towards all their stakeholders, which include suppliers, customers, shareholders, regulators, etc.

In various surveys, such as the one conducted by IBM, it has been established that large organisations are now increasingly using business analytics and look upon it as a means to attain a competitive advantage. In addition, the same survey also established that the organisations that have fully adopted business analytics are comparatively better and are likely to outperform their rivals who have not adopted BA or are in the process of adopting it.

Until recently, it was believed that only large organisations can use and reap benefits by implementing business analytics. IBM conducted a survey of C-level executives of mid-sized organisations and found that these executives were very well aware of the benefits their organisations could achieve by implementing business analytics, but a large section of the business decision-makers in these organisations did not have a correct understanding of the benefits. Some of the most visible benefits that accrue to an organisation include greater data visibility, greater analysing capability, measuring and monitoring financial and operational business performance, predicting outcomes, etc.

Notes

Despite having the knowledge of BA and its related benefits, many mid-sized organisations are hesitant to implement it because of certain perceived barriers. Mid-sized organisations believe that business analytics solutions are very expensive to implement. They also believe that it is difficult to implement business analytics such as in cases where new technologies need to be integrated in older or existing applications. Some other challenges in the implementation of business analytics include integrating structured and unstructured data in business strategies to achieve the required goals. Lastly, it is also difficult to integrate the existing CRM infrastructure with business data and analysis.

In reality, business analytics solutions are much more accessible and affordable than perceived by mid-sized organisations. Now, business analytics solutions providers have recognised the needs of mid-sized and small-sized organisations. They offer solutions for different businesses to provide business analytics capabilities to all the interested parties at the prices they can afford.

Aberdeen Group has identified certain trends that put up a strong case for use of business analytics solutions by mid-sized organisations. They include the following:

- **Big data:** A lot of data is generated by internal and external sources in an organisation. The volume, nature and the complexity of data leads to a collection of data known as big data which is an important factor in business decision-making.
- **Meaningful analysis:** In mid-sized organisations, the number and role of executives who need direct access to data for decision-making or to co-ordinate with other executives/departments is increasing. Therefore, it becomes inevitable to use business analytics.
- **Critical decisions and little time:** Executives of these mid-sized organisations are also responsible for making important or critical business decisions.

As a part of its study, Aberdeen Group found that many mid-sized organisations are increasingly resorting to business analytics alongwith formal data management practices to increase their market relevance and market value.

In the IBM's survey, it was found that mid-sized organisations were well aware of the associated benefits and about 64% of the respondents who were surveyed accepted that there was a need for greater visibility of their data and the associated analysis capabilities. 31% of the respondents recognised that they required business analytics capabilities, but had plans to implement it somewhere in future. Lastly, 5% of the respondents said that they were satisfied with their current approach and do not need any business analytics capabilities. The survey also revealed that a large percentage of mid-sized organisations viewed business analytics as a technology investment.

Investment in business analytics is highly recommended for mid-sized organisations because the most important functional data of organisations is stored in the departmental spreadsheets and business analytics can be used conveniently to retrieve data from various disparate sources and then combine in a meaningful manner. This data can then be used by business analytics solution for reporting and analysis. Most business analytics solutions provide features, such as dashboard, scorecard, planning, budgeting, and forecasting.

Business analytics solution providers are now providing affordable solutions targeted at mid-sized organisations. These solutions come with various capabilities and features, such as the following:

- Pre-configured solution to ensure ease of installation and integration with the existing applications.
- Common Web portal
- Plug and play compatibility
- Centralised Web console
- Minimum pressure on the IT staff
- No additional investment in IT resources
- Provides consistent and reliable information

**Source:** *https://www-03.ibm.com/innovation/us/engines/assets/ibm-business-analytics-case-study-1-22-13. pdf*

## QUESTIONS

1. What kind of challenges businesses are experiencing today?

   (**Hint:** Increasing revenues and reducing costs and the associated business risks amidst the increasingly volatile environment)

2. Why do you think mid-sized organisations are reluctant to implement business analytics?

   (**Hint:** Many mid-sized organisations are hesitant to implement business analytics because of certain perceived barriers. For example, mid-sized organisations believe that business analytics solutions are very expensive to implement.)

3. Assume that you are the CIO of a mid-sized organisation. Prepare a case for adopting a business analytics solution for your organisation. You have to present this case to your management. You can briefly list the important points.

   (**Hint:** There are pre-configured solutions to ensure ease of installation and integration with the existing applications.)

4. What are the most visible benefits that accrue to an organisation?

   (**Hint:** Some of the most visible benefits that accrue to an organisation include greater data visibility, greater analysing capability, measuring, and monitoring financial and operational business performance, predicting outcomes, etc.)

5. How many trends have been identified by Aberdeen group?

   (**Hint:** Aberdeen Group has identified certain trends that put up a strong case for use of business analytics solutions by mid-sized organisations:

   - Big Data
   - Meaningful analysis
   - Critical decisions and little time)

## 2.10 EXERCISE

1. Explain the concept of business analytics.
2. Enlist the different types of business analytics.

Notes

3. Describe the importance of business analytics.
4. Write a short note on SWOT analysis model.
5. Discuss PEST analysis model.

## 2.11 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Types of Business Analytics | 1. | c. Predictive analysis |
| | 2. | diagnostic |
| | 3. | root cause analysis |
| Relation between BI and BA | 4. | True |
| | 5. | True |
| | 6. | BA, BI |
| Role of Business Models in Analytics | 7. | b. Strengths, Weaknesses, Opportunities, Threats |
| | 8. | False |
| | 9. | False |
| | 10. | PEST |
| Importance of BA | 11. | True |
| | 12. | processes |
| Emerging Trends in BI and BA | 13. | c. Cloud Computing |
| | 14. | Digitisation |
| | 15. | Artificial Intelligence (AI) |

## 2.12 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Liebowitz, J. (2013). *Big Data and Business Analytics*. Boca Raton (FL): CRC Press.
- Laursen, G. H., & Thorlund, J. (2017). *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*. Hoboken, NJ: John Wiley & Sons, Inc.


### E-REFERENCES


- SearchBusinessAnalytics.(2018).*Whatisbigdataanalytics?-Definition*from*WhatIs.com*. [online] Available at: https://searchbusinessanalytics.techtarget. com/definition/big-data-analytics [Accessed 27 Nov.2018].
- Search Business Analytics. (2018). *What is business analytics (BA)? - Definitionfrom WhatIs.com*. [online] Available at: https://searchbusinessanalytics. techtarget.com/definition/business-analytics-BA [Accessed 27 Nov.2018].
- Simplilearn.com.(2018).*DataSciencevs.BigDatavs.DataAnalytics*.[online]Available at:https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article [Accessed 27 Nov.2018].

CHAPTER 3

# Resource Considerations to Support Business Analytics

## Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Discuss the role of business analytics personnel
- List the required competencies for an analyst
- Recognise the challenges of business data analytics
- Explain the technology used for business analytics

## 3.1 INTRODUCTION

In the previous chapter, you studied about explaining the types of Business Analytics (BA). Next, the chapter described the importance of BA. You have also studied about the business strategy analysis models.

Business analytics is a process to filter and analyse sets of data which might be small bits of data, a file containing the data or a large collection of data generally known as a database. With the growth in the data, a need of storing it at some appropriate location arises from where it can be easily accessed and modified irrespective of the geographical location. Unlike small datasets, which are useful only for individual organisations, Big Data is useful for various organisations. To store Big Data, companies use cloud technology, data warehousing, etc. This data is further retrieved from its storage and analytics is applied on it to derive useful information. The analytics involves the use of various statistical methods, such as measures of central tendency and graphs, to derive significant information from the data. This useful information is further used in businesses for decision-making, growth, planning, creating action plans and increasing overall profitability. The way of sorting the data to derive useful information has given a new purpose to business analytics.

In this chapter, you will first study about business analytics personnel and their roles. Further, the chapter discusses the required competencies for an analyst. Next, the chapter details upon business analytics data and the techniques used in the analytical processes.

## 3.2 ANALYTICS PERSONNEL AND THEIR ROLES

Both business analytics professionals and business analysts work closely till the finalisation of a project in a company, and, thus, are often considered to belong to the same profile. However, both these profiles have their individual roles that are very different from each other. Business analytics professionals focus on data and statistical analysis. They use data to predict future states of a business and, thus, help an organisation take important business decisions. To perform these tasks, a business analytics professional has to be an expert in statistics, mathematics and programming. Some additional skills, such as critical thinking, interpretation skills, communication skills, domain expertise, vision to picturise (think bigger), hands-on practices and knowledge of real-world situations, are also required to become a good business analytics professional.

With an increase in the adoption of analytics, many new job roles have emerged, some of which are shown in Figure 1:

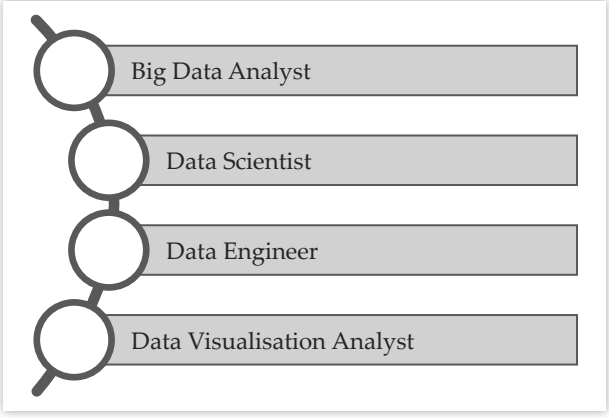



Figure 1: Types of Business Analytics Professionals

The responsibilities of business analytics professionals are as follows:

- **Big data analyst:** A big data analyst must have strong data mining skills and the basic working skills of Big Data platforms, such as Hadoop, MapReduce, Pig, Hive, etc. He/she can use scripting languages, such as R and Python, for processing raw data to generate business insights. The role of a Big Data analyst also involves communication and working with IT teams for fulfilling project requirements. Sometimes, the role of Big Data analyst also requires statistical analysis capabilities.
- **Data scientist:** The role of a data scientist involves deconstruction of structured and unstructured data as well as exploration of data by using different analytic techniques, such as predictive analytics and prescriptive analytics. The insight obtained after exploring data is then related to the objectives of an organisation and the conclusion is further communicated to different departments, such as marketing, IT, operations, etc. To become a data scientist, you require inter-disciplinary skills and sound knowledge of analytical languages, such as SAS/R/Python and Big Data platforms.
- **Data engineer:** The job role of a data engineer involves designing, building and managing of information or infrastructure related to Big Data. The main role of a data engineer is to build the architecture and systems used for analysing and processing data and ensure smooth functioning of these systems.
- **Data visualisation analyst:** The role of a data visualisation analyst is to analyse huge amounts of data or Big Data using various BI and visual analytical tools, such as Tableau and QlikView. The insight gained from data analysis is then presented using different visual formats, such as infographics, maps, multi-dimensional charts and dashboards. Due to the competitive business environment, it becomes necessary to track a company's performance in an efficient way or in a visual form so that it can be assessed easily and in less time. Therefore, visual analysts are in great demand these days.

On the other hand, a business analyst is anyone who has the key domain experience and knowledge related to the paradigms being followed. He/she often needs to sport multiple hats related to the field he/she is in.

NOTES

A business analyst can be anyone, from an executive to a top-level project director, given that they have grasp of the system, its techniques and functionality – since all they represent is the business their organisation is offering to customers.

Requirements are the essential part of creating successful IT solutions. Defining, documenting and analysing requirements that are developed from a business analyst's perspective help in demonstrating what a system can do. The skills of a business analyst are shown in Figure 2:

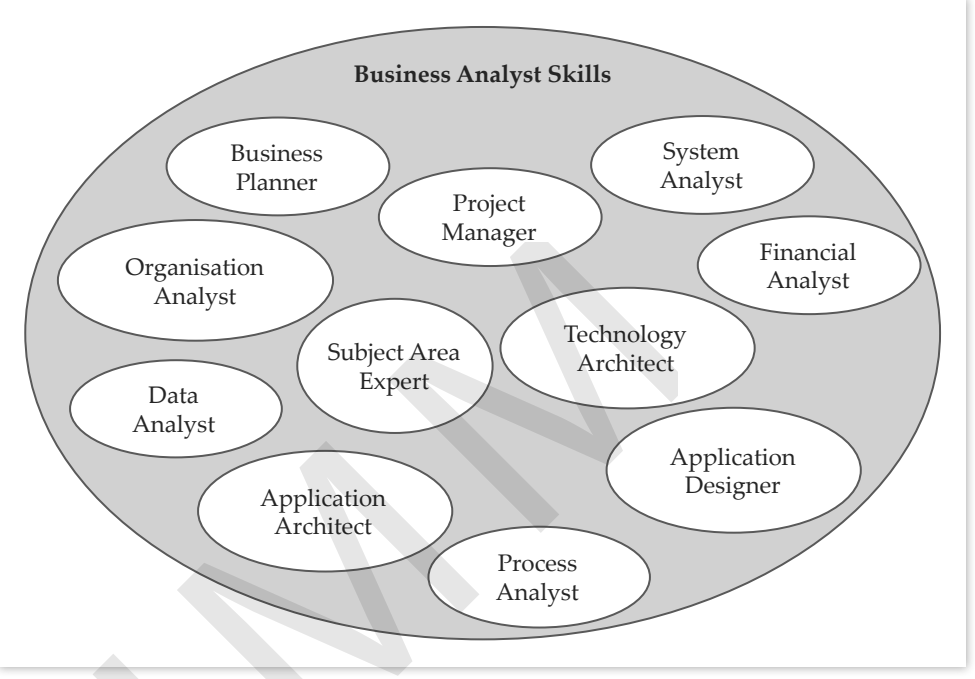



FIGURE 2: Skills of a Business Analyst

Described below are a few of the key requirements and responsibilities of a business analyst in managing and defining requirements:

- **Gathering the requirements:** Requirements are a key part of IT systems. Inadequate or unfitting requirements often lead to a failed project. The business analyst fixes the requirements of a project by mining them from stakeholders and from current and future users, through research and interaction.
- **Expecting requirements:** A business analyst who has expertise in his/her field knows that in the dynamic world of IT, things can change quickly even before they can expect the change. Plans developed at starting are always subject to alteration, and expecting requirements that might be needed in the future is key to successful results.
- **Constraining requirements:** While complete requirements are a must for a successful project, the emphasis should be the essential business needs, and not the personal user preference, functions based on the outdated processes or trends, or other unimportant changes.
- **Organising requirements:** Requirements often come from multiple sources that sometimes may contrast with other sources. A business analyst must segregate

requirements into associated categories to efficiently communicate and manage them. An ideal organisation averts project requirements from being overlooked, and, thus, leads to an optimum use of budgets and time.

- **Translating requirements:** A business analyst must be skilled at interpreting and converting the business requirements effectively to the technical requirements. It involves using powerful modelling and analysis tools to meet planned business goals with real-world technical solutions.
- **Protecting requirements:** At frequent intervals in a project's life cycle, the business analyst protects the user's and business needs by confirming the functionality, precision and inclusiveness of the requirements developed so far compared to the requirements gathered in the initial documents. Such protection reduces the risk and saves considerable time by certifying that the requirements are being fulfilled before devoting further time in development.
- **Simplifying requirements**: The main role of a business analyst is to simplify tasks and maintain easier functionality. Completing the business objective is the aim of every project; a business analyst recognises and evades unimportant activities that are not helpful in resolving the problem or achieving the objective.
- **Verifying requirements**: Business analysts are the most informed people in a project about the use cases. Hence, they frequently validate the requirements and discard implementations that do not help in growing the business objective to culmination. Requirement verification is completed through test, analysis, inspection and demonstration.
- **Managing requirements**: Usually, an official requirements presentation is followed by the review and approval session, where project deliverables, costs and duration estimates and schedules are decided and the business objectives are rechecked. Post approval, the business analyst shifts to requirement managing events and activities for the rest of the project life cycle.
- **Maintaining system and operations:** Once all the requirements are completed and the solution is delivered, the business analyst's role shifts to post-implementation maintenance. This shifting ensures that defects, if any, do not occur or are resolved in the agreed SLA timelines; any enhancements that are to be made to the project, or performing change activities to make the system yield more value; similarly, the business analyst is also responsible behind many other activities post implementation, such as operations and maintenance, or giving system authentication procedures, deactivation plans, maintenance reports and other documents like reports and future plans. The business analyst also plays a great role in studying the system to regulate when replacement or deactivation may be required.

**Self Assessment Questions**

1. A top-level project director cannot be a business analyst. (True/False)
2. Defining, _____ and analysing requirements that are developed from a business analyst's perspective help in demonstrating what a system can do.

NOTES

3. Which of the following are the key responsibilities of a business analyst?
   a. Gathering the requirements
   b. Expecting requirements
   c. Constraining requirements
   d. All of these
4. A _____ is the most informed person in a project about the use cases.

**ACTIVITY**

As a business analyst, prepare a report on your analytical study of Sony Corporation, currently undergoing turmoil for serving too many areas in business fields.

## 3.3 REQUIRED COMPETENCIES FOR PERSONNEL IN ANALYTICS

In the previous section, you have learned about the roles performed by the analytics personnel in an organisation. Analytics personnel need to possess various types of skills to perform their roles effectively. Some of these skills are discussed in Figure 3:

| | |
|---|---|
| **Statistical Skills** | Good knowledge of statistical skills or tools such as regression, Bayesian statistics, machine learning and descriptive statistics enable an analytics professional to perform efficient analysis of data. |
| **R or Python or Statistics Software Skill** | Good knowledge of languages and statistical software help an analytics professional in analysing Big Data. Analysation of large amounts of data is not possible without the knowledge of these tools. |
| **Visualisation Skills** | Data visualisation helps an analytics professional to communicate the results of data analysis across organisation. Visualisation skills are not about creating elegant charts. It is about visualising data in a simple, accurate and concise manner so that the targeted audience can understand and act on the information. |
| **SQL Skills** | SQL is the coding language used for performing query on databases. Different types of variations exist for SQL, such as Presto SQL, Hive and Server SQL, which are commonly used by organisations these days. Knowledge of SQL skills helps in fetching and analysing data. |
| **Writing Skills** | Good writing skills help an analyst to provide insights of data in an efficient way. |

FIGURE 3: Skills Required for Personnel in Analytics

Besides preceding skills, a business analytics professional should have the knowledge of handling real-world problems existing in the field of analytics. A business analyst is also an analytics professional. However, the skills required for the role of a business analyst differ, but not to great extent.

Business analysts need to be great in verbal and written communication, diplomatic, experts with problem-solving acumen, theorists with the ability to involve with

stakeholders to comprehend and answer to their needs in a dynamic business environment. This includes dealing with senior members of management and challenging interrogation sessions to confirm that the time is well spent and value for money development can commence.

Business analysts need not necessarily be from IT background, although it certainly helps having the basic understanding of the IT systems and how they work. Sometimes, business analysts come from a programming or other technical background often from within the business – carrying thorough information of the business field, which can be likewise very useful. Following are some of the most common skills that a decent business analyst should have:

- **Understanding the objectives:** Ability to understand the path and commands is important. If you cannot understand what and, more significantly, why you are assigned to do something, chances you cannot deliver what is required are high. Do not hesitate in asking questions or additional information if you have any doubts.
- **Having good communication skills:** Sounds obvious, but it is necessary to have good verbal communication skills, preferably in a global environment, where multitudes of stakeholders, management and resources from diverse backgrounds will collaborate on a single platform to discuss, debate and finalise the requirements which would incidentally be captured by you. It is necessary for you to have that comprehension level along with the eloquence to deliver your conceptions or clear any doubts which you have. You should be able to make your point evidently and explicitly.
- **Manage stakeholder meetings:** While email also acting as an audit trail is a fair method to facilitate communication, sometimes it turns out to be not enough. Old school F2F discussions and meetings for detailed deliberation over the problems and any queries are still a popular way of carrying out effective analysis. Most of the time, you end up discovering more about your project from the physical presence of all stakeholders where all collaborators tend to be open about debating circumstances.
- **A good listener:** You are better off listening to more than you speak and jotting down the notes and takeaways from the meetings. Having good listening skills requires patience and virtue to understand and listen to the stakeholder, which gives them a feeling of being heard and not being overlooked or overpowered by a dominating analyst. Such projects often end up in mess sooner than they should be. Your listening and information-absorbing skills are important to make you an effective analyst. Not only listen, but also understand the situation, question only where you think you are being condescended upon by the stakeholders passing off unnecessary off-business requirements and ignoring the actual requirements that can help in the making of an efficient system. You can attend personality development training to get the control over voice modulation, and dialect and pitch moderation along with an effective body language with business presentation skills.
- **Improving the presentation skills:** As a business analyst, you are supposed to be presentable at any time round the clock. As a business analyst, you will often lead

Notes

workshops or pitch a workpiece to the stakeholders, or to the internal project team. It is important to give due consideration to the content of your presentation and ensure that it matches the objectives to be met – since there is no point of presenting the implementation methods if the meeting is about gathering requirements. These presentations not only represent information, but also act as a good way to get more clarity or information from stakeholders in case you are looking for further details on a specific part of the project.

- **Time manager:** A business analyst is responsible for maintaining the timeframes of the project as well as the corporate schedules. He should ensure that the project meets the pre-agreed project milestones along with daily tracking schedules being fulfilled by the development team. He should prioritise activities separating critical ones from the others that can wait, and focus on them.
- **Literary and documenting skills:** Being a business analyst, you are supposed to deliver numerous types of documentations, such as requirements documents, specifications, reports, analysis and plans that will go on to become projects and legal documents later on. So, you need to ensure that your documents are created concisely, and at a comprehensible level for the stakeholders. Avoid specific jargons to a particular field as they may not be understood by all stakeholders and, later, may create confusion or other complexities with their interpretations. Starting as an inexperienced business analyst, you will gradually learn to write requirement documentations and reports, but having strong writing skills is enough to give you a head start over the others since it will lead to unambiguous requirements documentation.
- **Stakeholder management**: It is important that you know how to deal with stakeholders and know how much power and impact they have on your project. They can either be your best friends/supporters or your greatest critics. An accomplished business analyst will have the skill to investigate the degree of management every stakeholder needs and how they ought to be independently dealt with.
- **Develop your modelling skills:** As the expression goes, a photo paints a thousand words. Procedures (such as process modelling) are compelling tools to pass on a lot of data without depending on the textual part. A visual portrayal enables you to get an outline of the issue or project so that you can see what functions well and where the loopholes lie.

**Self Assessment Questions**

5. The role of a business analyst is considered as a bridge between business and _____.
6. Business analysts need not necessarily be from IT background, although it certainly helps having the basic understanding of the IT systems and how they work. (True/False)
7. Communicating the _____ and the _____ at the appropriate level is important – as some stakeholders require more detailed information than others due to the varying levels of understanding.

8. Which of the following statements is false about the role of a business analyst?
   a. To be called a successful business analyst, you ought to be a multi-skilled person who is adaptable to an ever-changing environment.
   b. Your listening and information-absorbing skills are important to make you an effective analyst.
   c. As a business analyst, you are supposed to be presentable at any time round the clock.
   d. A business analyst is not responsible for maintaining the time frames of the project.

**Activity**

You are a veteran business analyst, responsible for coaching a new batch of management trainees in an organisation. Layout the course plans and methods you will utilise to train them about the standards and the knowledge.

## 3.4 BUSINESS ANALYTICS DATA

Any approach for analytics must adjust to changes in the way people work inside their business settings, particularly with the developing size of data volumes. Arranging data that is redone in a way that bodes well for every business customer requires infusing content with context before augmenting the estimation of relevant filtering and representation. Enhancing the enormous amounts of data and making a presentation of significant learning for every business consumer's needs shows up with many difficulties. We will segregate those problems as data analytics challenges—creating algorithms that will gather, analyse, group, channel, categorise and, at last, filter the meaning and also persistently retrain the machine, cutting and dicing this data in view of the individual needs and conveying it in a way that is most useful relying upon a person's perspective (area, time, gadget, and so on). Some of the data analytics challenges are as follows:

- **Content variety and quality:** Information sources are no longer entirely organised. Business folks depend on a pool of information objects that mix customarily structured information with various types of artefacts, for example, transactional system databases and, in addition, Web-based social networking channels, like Facebook, Twitter, LinkedIn, Web journals, wikis, etc., each of which must be surveyed for logical importance and incorporated inside different data models. For quality, the bits of information that can be mined from an information source like a database or an online networking Web page may have distinctive levels of relevance for various sorts of data consumers in different places of an organisation. One example is information gathered for announcing the item launched for senior officials. A moved-up lookout of positive or negative beliefs might be adequate, while the product manager may search for insights with respect to potential item defects that can be quickly remediated.

Notes

- **Content organisation:** Forming the data inputs begins with a set of meaning and semantics, but business requirements change over time. So, the models need to be flexible with capacity to provide allowances in relation to taxonomic models, tag inputs and match them based on incidental content.
- **Connectivity:** Any information source may have different levels of importance inside a wide range of business settings. For instance, remarks about a bike's drivability might be more important coming from a vehicle-enthusiast blog owner, which can be checked through Twitter. That poses two difficulties – firstly, linking information artifacts to various business domains, while the second includes deriving dynamic linkages, connections and relevance beyond settled ordered models. The last challenge likewise implies striving to advance an understanding of how data sets are utilised by various people and adjusting analytical models, respectively.
- **Personalisation challenges:** More important than separating through substantial volumes of data resources taken from a variety of sources is that a wide range of channels must be set up to recognise different filters of business value relying on who the customers are. For instance, a sales delegate may be informed about a few particular contacts from their client base to help in generating leads. Similar data sources can be refined to provide sales and marketing executives with subjective information about their top clients, help recognise potential threats from competitors and inform about techniques for continuing with expansion inside vertical markets.
- **Finding correlations in a dynamically changing business world:** Pattern detection in data correlations may specify developing trends. For example, investigating the correlation between Web searches about influenza symptoms, and medicines and geographical places over a period can help in forecasting the patterns for influenza infections.

### Self Assessment Questions

9. Which of the following is an example of social networking?
   a. Facebook
   b. Twitter
   c. LinkedIn
   d. All of these
10. Dissimilar levels of information sparseness, density, freshness and quality affect the capability to unify the data and require increased sophistication. (True/False)
11. _____ detection in data correlations may specify developing trends.

### Activity

If a raw sample data from a research institute lands at your department, what will be your first reaction in order to polish up the data?

## 3.5 TECHNOLOGY FOR BUSINESS ANALYTICS



As a push to make analysis more significant and unmistakable to the business client, solutions are concentrating on particular vertical applications and customising the outcomes and business audience interfaces. For usability, less complex and compelling arrangement, and ideal value, analytics are being installed in bigger systems. Therefore, issues like information gathering, storage and processing related to analytics are overall increasingly viewed as critical issues in system design. In an endeavour to expand the capability of analytics in a business procedure, provisions are being developed that go beyond the client facing applications, working in background to applications in sales, supply chain perceptiveness, advertising, value improvement, and workforce analysis.

Technologies and trend variations in technologies are possibly the most noticeable BI components in the IT industry which are used by organisations for the analysis of business information. These technologies can also handle structured or unstructured data in large amounts to determine, develop or create new business opportunities.

We might think that the need of data volume to make a specific decision has decreased over time either due to the overall shift in management or due to assumptions with terms with higher significance, such as insight, knowledge and ideas.

While taking the human factor in mind, the change between reactive and proactive decision-making is defined by the complexity level of the fields between advanced analytics and BI. Summary reports, statistics and queries, and low-latency dashboards are built on chronological information. There is a mid-ground for simple analytics, e.g., algebraic or trending predictions that give estimated answers about expectations in terms of sales, production, etc.

Advanced analytics are much more refined, support techniques, such as statistical analysis, forecasting, prediction and correlation, whereas trend analysis simply infers the existing data to project the next quarter.

Let us take a look at decision-making from another point of view. Say, we want to examine our brain while taking a decision. From a logical viewpoint, when our brain encounters a task it has no idea about, it attempts to create rational assumptions guessing the input, likely outcomes vs. actions to be taken, and attempts to find the best answer. When the brain encounters the same level of problem again, it reimagines the outcomes and methods deployed as in the old task, before trying to figure out the right answer to the current problem, assesses what worked earlier and what did not. After being subjected to a certain amount of similar or varying tasks, brain becomes familiar to cracking a specific type of task. Consequently, the time of re-examining the older solutions and finding the right solution for the new task reduces significantly.

Alongside the issues of supportive human policy-making patterns, the structural set-up of a BI system should be prudently measured. A number of sources with printed study specify that intelligence works finest when planned as a joint effort by involving people. This effort needs to be correctly synchronised in terms of urgencies, responsibilities, procedures and, at the same time, intelligence set-up should backup

and inspire an effective flat exchange of data among contributors. There are multiple cases where a state-of-the-art business intelligence technology failed to deliver on the expectations because of the unwillingness of the persons to take care of the data-hungry system and accomplish additional actions that are required from time to time. Though taking the learning curve into account, capabilities and patterns intricate ways of human capacity and learning still exceed machine learning in countless areas. People have never been more able to understand, and use specific technologies. The next generations are finding technologies less intimidating and assume the techno-human connection as regular and undisputable.

The business analytics functions are compilation of the most commonly utilised efficient practices in business analysis across the globe as per the standards prescribed by the BABOK (Business Analysis Body of Knowledge). These standards keep on evolving and incorporate new changes dynamically in the form of versions. It is a framework that describes the knowledge, skills and capabilities required to accomplish business analysis efficiently. Software development methodologies like Agile and SCRUM are commonly occurring standards that help in creating an iterative informative solution for the system which is composed of several layered steps of dealing with the SDLC and associated phases. Coming to application tools, business analysts across the world utilise applications, like MS Word, Excel, Visio, PowerPoint and Project, and many such tools in order to put their best foot forward. These tools are effective and clear in presenting information closest to depiction as wanted by the analyst, and, hence, elate the overall levels of analytical operational standards.

**Self Assessment Questions**

12. A refined _____ takes seasonality, correlations between strong and weak quarters, and historical sales outlines into account.
13. For usability, less complex and compelling arrangement, and ideal value, analytics are being installed in _____ systems.
14. Summary reports, statistics and queries, and low-latency dashboards are built on a _____ information.

## 3.6 SUMMARY

- Business analytics is a process to filter and analyse sets of data which might be small bits of data, a file containing the data or a large collection of data generally known as a database.
- A business analyst is anyone who has the key domain experience and knowledge related to the paradigms being followed.
- A business analyst can be anyone, from an executive to a top-level project director, given that they have grasp of the system, its techniques and functionality – since all they represent is the business their organisation is offering to customers.
- Defining, documenting and analysing requirements that are developed from a business analyst's perspective help in demonstrating what a system can do.

- The business analyst fixes the requirements of a project by mining them from stakeholders and from current and future users, through research and interaction.
- Plans developed at starting are always subject to alteration, and expecting requirements that might be needed in the future are key to successful results.
- A business analyst must segregate requirements into associated categories to efficiently communicate and manage them.
- A business analyst must be skilled at interpreting and converting the business requirements effectively to the technical requirements.
- Post approval, the business analyst shifts to requirement managing events and activities for the rest of the project life cycle.
- Once all the requirements are completed and the solution is delivered, the business analyst's role shifts to post-implementation maintenance.
- The business analyst also plays a great role in studying the system to regulate when replacement or deactivation may be required.
- To be called a successful business analyst, you ought to be a multi-skilled person who is adaptable to an ever-changing environment.
- Communicating the data and the information at an appropriate level is important – as some stakeholders require more detailed information than others due to the varying levels of understanding.
- As a business analyst, you will often lead workshops or pitch a workpiece to the stakeholders, or to internal project team.
- A business analyst is responsible for maintaining the time frames of the project as well as the corporate schedules.
- Pattern detection in data correlations may specify developing trends.
- A refined predictive model takes seasonality, correlations between strong and weak quarters, and historical sales outlines into account.
- Any information source may have different levels of importance inside a range of business settings.

## 3.7 KEY WORDS

- **Business Analysis Body of Knowledge (BABOK):** It is a compilation of most commonly utilised efficient practices in business analysis across the globe.
- **Business analyst:** Anyone who has the key domain experience and knowledge related to the paradigms being followed.
- **SCRUM:** It is a software development methodology that helps in creating an iterative informative solution for the system which is composed of several layered steps of dealing with the SDLC and associated phases.
- **Stakeholder management:** It is a process of dealing with stakeholders and understanding how much power and impact they have on your project.
- **Stakeholders:** These can either be your best friends/supporters or your greatest critics.

Notes

## 3.8 CASE STUDY: THE BLUEPRINT OF A BUSINESS ANALYST

Michael Rodriguez was a lead software developer and specialised in gathering requirements for various software. He was a competent professional and was satisfied with his career, but he wanted to further develop his business analysis skills. To fulfil this desire, he joined the Business Analyst Blueprint programme.

In an interview with Laura Brandenburg in a Business Analyst Blueprint session, Michael shared his story as a Business Analyst and gave a few useful tips as well. Let us go through a synopsis of Michael's story.

Michael started off as an application developer and worked his way up as a lead developer. For the most part of his 15-year-long career, he served as a lead developer of a small team, or was the primary developer of small projects.

During any maintenance or development projects for software, it is important to meet the end users and gather their requirements. Michael understood the importance of this phase and started getting into a role that dealt solely with requirements gathering and eliciting. He believed that if requirements are gathered properly, the software solution can be developed effectively. He stated that during this 10-year-long phase, when Michael got into the requirements-gathering phase, he had never undergone any formal business analysis course. Only when he started interacting with such clients, he started realising that he should probably go for some business analytics courses. Therefore, he started searching the Internet and visited the Bridging the Gap website.

Michael stated that he moved from development to lead development to requirements analysis and started taking on more complex projects. Michael told the interviewer that he started experiencing various challenges during his career while handling complex projects, which further motivated him to take up Business Analytics study. To explain this further, he illustrated one of his project experiences, wherein he was serving as a lead developer and the project director asked him to start talking with various organisations and gather their requirements that would help them in improving their processes. At that point, Michael did not have any idea on how to start this work.

Michael's director asked him that he should start asking deeper questions to determine what exactly the clients want. Michael stated that the process of determining requirements comprises two parts – 80% of the entire process relates to requirements gathering, and the rest 20% of the process relates to the manner in which the requirements are gathered. According to Michael, "Once you gather your requirements, once you learn the software, 20% is the mechanics of understanding requirements and putting that to play into the software that you're developing."

During the interview, Michael stated that he wanted to gain some BA skills through the BA Blueprint course and club it with his own experiences to determine how things flow and work during a BA environment. Michael also told the interviewer that he had set a certification goal for himself.

Michael started with the Business Process Analysis module. He also stated that he has worked as a business analyst on a big project. He said that he was looking into various aspects of the project that had not been built into the software. He added that the course study materials, meeting agenda, opening scripts and the assignments helped him tremendously. To illustrate this point, he stated that he had organised a meeting. He wanted to review the process flow with the stakeholders. He had the flow diagram ready with him along with handy notes. The meeting went very well. He told that opening scripts, use cases and wireframes are really helpful. He used to start the requirements gathering session with a wire-framing tool. He also stated that he started this practice of having a wire-framing tool handy before the requirements-gathering session.

Similar to the Business Process Analysis module, Michael also did trainings on data modelling module. The interviewer closed the session by listing down the main points of the interview, which are as follows:

- Communication skills play a huge role in asking deeper questions regarding stakeholder assumptions and thinking.
- Business processes can be used to raise an analyst's level of thinking specifically when an analyst is focussed on software requirements.
- Expanding and growing one's career in Business Analytics, one must be sure that he is applying all the business analysis techniques in the right manner.

**Source:** *https://www.bridging-the-gap.com/michael-rodriguez/*

## QUESTIONS

1. What according to Michael is the most important task during software development?

   (**Hint:** Requirements elicitation is the most important task during software development.)

2. List down the features of Business Analytics that Michael found beneficial.

   (**Hint:** Opening scripts, use cases and wireframes, etc.)

3. Name the tool used by Michael to start requirements-gathering session.

   (**Hint:** He used to start the requirements gathering session with a wire-framing tool.)

4. Which type of skills play a huge role in asking deeper questions regarding stakeholder assumptions and thinking?

   (**Hint:** Communication skills play a huge role in asking deeper questions regarding stakeholder assumptions and thinking.)

5. Which type of processes can be used to raise an analyst's level of thinking specifically when an analyst is focussed on software requirements?

   (**Hint:** Business processes can be used to raise an analyst's level of thinking specifically when an analyst is focussed on software requirements.)

Notes

## 3.9 EXERCISE

1. Enlist and explain the roles and responsibilities of a business analyst.
2. Describe the required competencies for an analyst.
3. What do you understand by business analytics data?
4. Write a shot note on technologies used for business analytics.

## 3.10 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Analytics Personnel and their Roles | 1. | False |
| | 2. | documenting |
| | 3. | d. All of these |
| | 4. | business analyst |
| Required Competencies for Personnel in Analytics | 5. | stakeholders and IT |
| | 6. | True |
| | 7. | data; information |
| | 8. | d. A business analyst is not responsible for maintaining the time frames of the project |
| Business Analytics Data | 9. | d. All of these |
| | 10. | True |
| | 11. | Pattern |
| Technology for Business Analytics | 12. | predictive model |
| | 13. | bigger |
| | 14. | chronological |

## 3.11 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Laursen, G. and Thorlund, J. (2013). *Business Analytics for Managers.* Hoboken, N.J.: Wiley.
- Noble, B. (2018). *The Enterprise Business Analyst: Developing Creative Solutions to Complex Business Problems*. [online] Barnes & Noble.


### E-REFERENCES


- Brandenburg, L. (2018). 5 Steps to Becoming a Business Analyst. [online] Bridging-the-gap.com. Available at: https://www.bridging-the-gap.com/becoming-a-business-analyst/[Accessed 29 Nov. 2018].
- Businessanalystsolutions.com. (2018). What is a Business Analyst?. [online] Available at: http://www.businessanalystsolutions.com/what_is_a_business_analyst.html [Accessed 29 Nov. 2018].

CHAPTER 4

# Introduction to Statistics

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the concept of probability theory
- Describe the significance of measures of central tendency
- Define the concept of probability theory
- Discuss the significance of statistical inference
- Elucidate the meaning and importance of hypothesis testing
- Define the importance of Correlation
- Explain the concept and types of regression analysis

## 4.1 INTRODUCTION

In the previous chapter, you studied the basic concept of resource considerations to support business analytics. The chapter explained in detail about various business analytics personnel and their roles, required competencies for an analyst, business analytics data and technology for business analytics, which directly or indirectly influence an organisation's decision-making process. You also studied about different managing changes in the previous chapter.

Ever since the evolution of Big Data, the data storage capacities have grown a lot. It becomes increasingly difficult for the organisations to process such huge amounts of data. Here comes the role of data science. Data science is a multi-disciplinary subject that has developed as a combination of mathematical expertise (data inference and statistics) and algorithm development, business acumen and technology in order to solve complex problems.

At the core of business operations is data. An overwhelming amount of data is stored in the enterprise data warehouses and a lot of value can be derived from it by mining the data. Data warehouse can be used to discover the data and development of a data product that helps in generating value.

Data science helps in uncovering findings from data. Discovering data insights involves mining data at granular level and understanding complex behaviours, trends and inferences which can be used by the organisations to make better business decisions. For example, the P&G Company makes use of time series models to understand the future demand and plan for production levels more optimally.

A data product is a technical asset that takes in data as input and processes the data to return algorithm-based results. An appropriate example of a data product is the recommendation engine that takes in user data and makes personalised recommendations based on the data. For example, e-commerce websites, such as Flipkart and Amazon mine data to understand the buying pattern of the consumers and then, based on its analysis, it recommends other similar products that may interest the buyers.

Notes

Field of data science involves use of techniques, such as machine learning, statistical skills, cluster analysis, data mining, algorithms and coding and visualisation. Note that statistics plays a central role in the data science applications.

Data science involves use of statistical techniques for data collection, visualising the data and deriving insights from them, obtaining supporting evidence for data-based decisions and constructing models for predicting future trends from the data.

In this chapter, you will learn about various techniques of statistics that are used frequently in data sciences. You will learn about the important concepts, such as probability theory, statistical inference, sampling theory, hypothesis testing and regression analysis.

## 4.2 MEASURES OF CENTRAL TENDENCY

There are three main measures of central tendency. These are as follows:

- **Arithmetic Mean:** The mean of a variable represents its average value. It can be calculated by using the following formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum f_i}$$

where, X represents the sample mean and $f_i$ represents the frequency of an $i^{th}$ observation of the variable. One of the problems with arithmetic mean is that it is highly sensitive to the presence of outliers in the data of the related variable. To avoid this problem, the trimmed mean of the variable can be estimated. Trimmed mean is the value of the mean of a variable after removing some extreme observations (e.g., 2.5 per cent from both the tails of the distribution) from the frequency distribution. Mean is the hypothetical value of a variable. It may or may not exist in the dataset.

- **Median:** Median is known as the 'positional average' of a variable. If we arrange the observations of a variable in an ascending or descending order, the value of the observation that lies in the middle of the series is known as median. The value of the median divides the observations of a variable into two equal halves. Half of the observations of the variable are higher than the median value and the other half observations are lower than the median value. The extensions of median are quartiles, deciles and percentiles.

- **Mode:** The mode of a variable is the observation with the highest frequency or the highest concentration of frequencies. Let us take an example of the sales figures of an enterprise to better understand the concept of mean, median and mode. The monthly sales figures (in crores) of an enterprise for 20 consecutive months are given in Table 1:

**Table 1:** Monthly Sales Figures (in crores)

| Month | Sales |
|---|---|
| 1 | 60 |
| 2 | 70 |

| Month | Sales |
|---|---|
| 3 | 45 |
| 4 | 90 |
| 5 | 110 |
| 6 | 40 |
| 7 | 90 |
| 8 | 50 |
| 9 | 70 |
| 10 | 65 |
| 11 | 54 |
| 12 | 72 |
| 13 | 45 |
| 14 | 24 |
| 15 | 12 |
| 16 | 8 |
| 17 | 15 |
| 18 | 40 |
| 19 | 54 |
| 20 | 56 |

The formula required for estimating mean in MS Excel is shown in Figure 1:

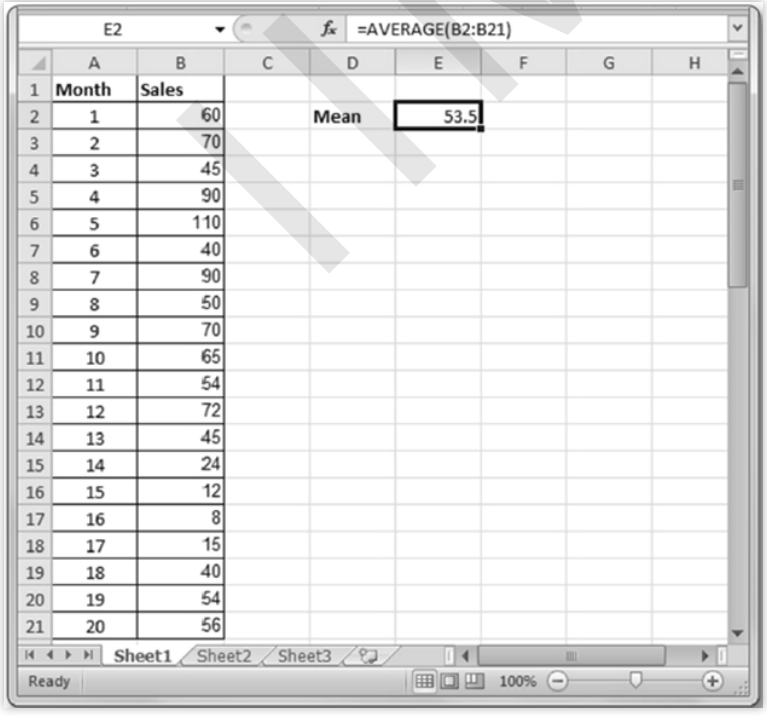



Figure 1: Estimating Mean in MS Excel

Notes

The formula required for estimating median in MS Excel is shown in Figure 2:

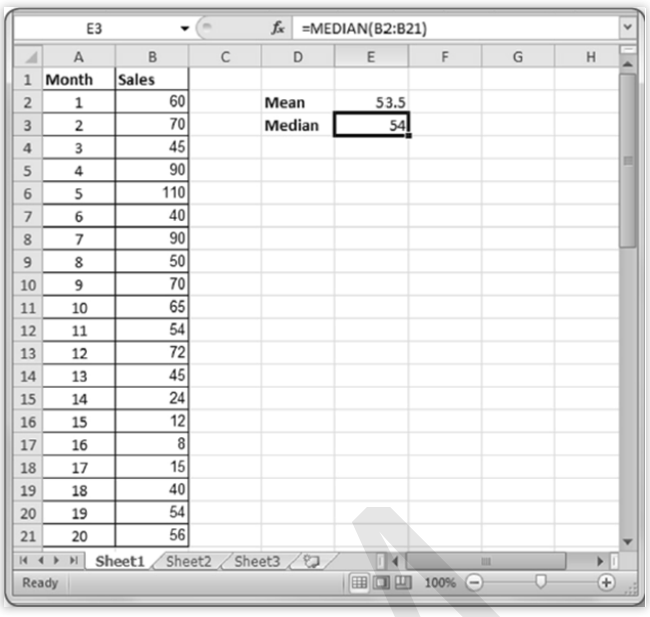



Figure 2: Estimating Median in MS Excel

The formula for estimating mode in MS Excel is shown in Figure 3:

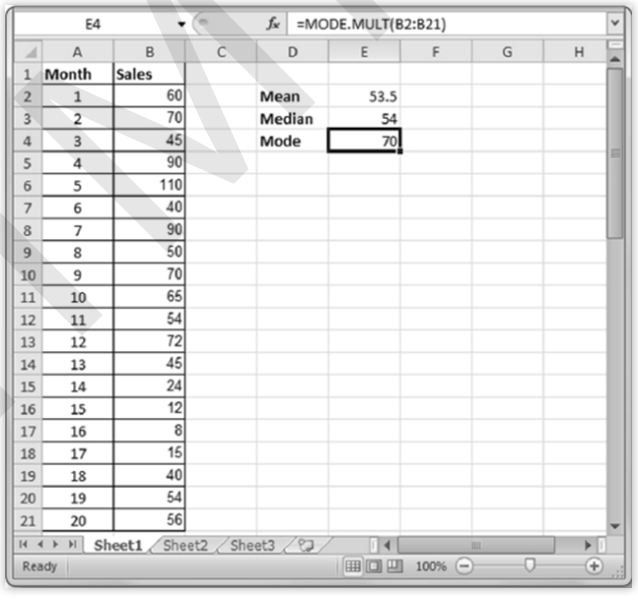



Figure 3: Estimating Mode in MS Excel

## Self Assessment Questions

1. The _____ of a variable represents its average value.
2. _____ is known as the 'positional average' of a variable.
3. The _____ of a variable is the observation with the highest frequency or highest concentration of frequencies.
4. _____ is the hypothetical value of a variable.

**ACTIVITY**

Calculate the mean, median, mode and range for the following list of values:

12, 17, 12, 13, 12, 15, 13, 20, 12

## 4.3 PROBABILITY THEORY

Probability theory deals with concepts by expressing them in the form of axioms which formalise in terms of probability space. The probability may take any value between 0 and 1. The probability space assigns a value between 0 and 1 to a set of outcomes which are called sample space. If a subset of the sample space is taken, it is called an event.

The probability theory involves use of discrete and continuous random variables and probability distributions. The distributions provide mathematical abstractions of non-deterministic or uncertain processes or measured quantities which may occur as a single occurrence or over time. Random events cannot be predicted perfectly. However, their behaviour can be analysed. The law of large numbers and the central limit theorem are used to describe the behaviour of such random events.

Study of probability theory is essential for human activities involving quantitative data analysis. It acts as a mathematical foundation for the concepts, such as uncertainty, confidence, randomness, variability, chance and risk. Probability theory is also used by various experimenters and scientists who make inferences and test hypotheses based on uncertain empirical data. Probability theory is also used to build intelligent systems. For example, techniques and approaches, such as automatic speech recognition and computer vision, which involve machine perception and artificial intelligence, are based on probabilistic models.

In the study of probability theory, a probability distribution is a mathematical function. This mathematical function or probability distribution is used in experiments to provide the probabilities of occurrence of different outcomes (events). For example, if X is a random variable which denotes the outcome in a (fair) coin toss experiment, then probability distribution of X will be X = 0.5 for heads and X = 0.5 for tails.

Let us now study about the continuous probability distributions.

### 4.3.1 CONTINUOUS PROBABILITY DISTRIBUTIONS

A continuous random variable is a random variable having an infinite and uncountable range. If the random variable is continuous, its probability distribution is called continuous probability distribution.

A probability distribution can be described using an equation called Probability Density Function (PDF). The area under the curve of a random variable's PDF shows the probabilities of the continuous random variables. Here, it must be remembered that a range of values can have a non-zero probability. For example, we can calculate the probability that a student has scored marks between 80 and 90. Probability of a continuous random variable having some value is zero. Due to this reason, a continuous probability function cannot be expressed in tabular form.

Notes

A continuous probability distribution is described using an equation or a formula. For a random variable Y, PDF y = f(x) means that y is a function of x. For all values of x, the value of y will be greater than or equal to zero. Also, the total area under the curve of function is equal to one.

For example, the PDFs of men's height are shown in Figure 4 as follows:

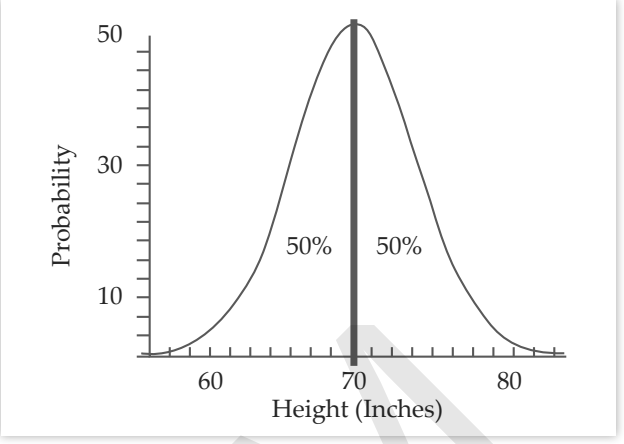



Figure 4: Continuous Probability Distribution Function of Men's Heights

For a continuous probability distribution for men's heights, we cannot measure the exact probability that a man will have a height of exactly 70 inches. It only shows that an average man has a height of 70 inches. It is not possible to find out the probability that any one person has the height of exactly 70 inches.

Examples of continuous distribution include uniform distribution, normal distribution and exponential distribution.

### 4.3.2 DISCRETE PROBABILITY DISTRIBUTIONS

Random events generally lead to discrete random variables whose values are not fixed. Usually, the discrete random variables are denoted as X and their probability distribution is denoted as P(X). For example, if a coin is tossed n times and the number of times tails comes up are counted, then it is a discrete random variable.

Here X may take values 0, 1 or 2 if a coin is tossed two times. Some of the most common discrete probability distributions used in statistics include binomial distribution, geometric distribution, hypergeometric distribution, multinomial distribution, negative binomial distribution and Poisson distribution.

The frequency distribution table for the probability of rolling a dice is shown in Table 2 as follows:

Table 2: Frequency Distribution Table for the Probability of Rolling a Dice

| Roll | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Odds | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

Consider an example to calculate the frequency distribution in Excel from the scores obtained by 20 students in a competitive exam. The scores of students are as follows:

97, 102, 123, 133, 141, 156, 97, 90, 100, 112, 114, 117, 123, 125

Perform the following steps to calculate the frequency distribution in Excel:

1. *Enter* the data in the Excel sheet along with its upper limits, as shown in Figure 5:

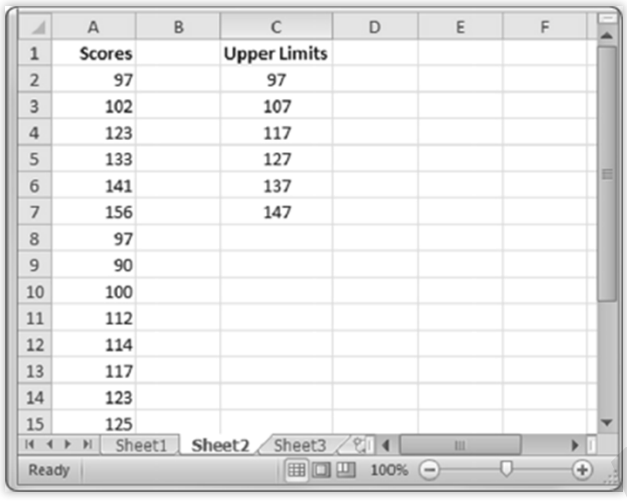



FIGURE 5: Entering Data in Excel

2. *Click* **File → Options** in the File menu. The Excel Options dialog box appears (Figure 6).
3. *Select* the Add-Ins option in the left pane. Ensure that the Excel Add-ins is selected in the Manage drop-down list (Figure 6).
4. *Click* the **Go** button, as shown in Figure 6:

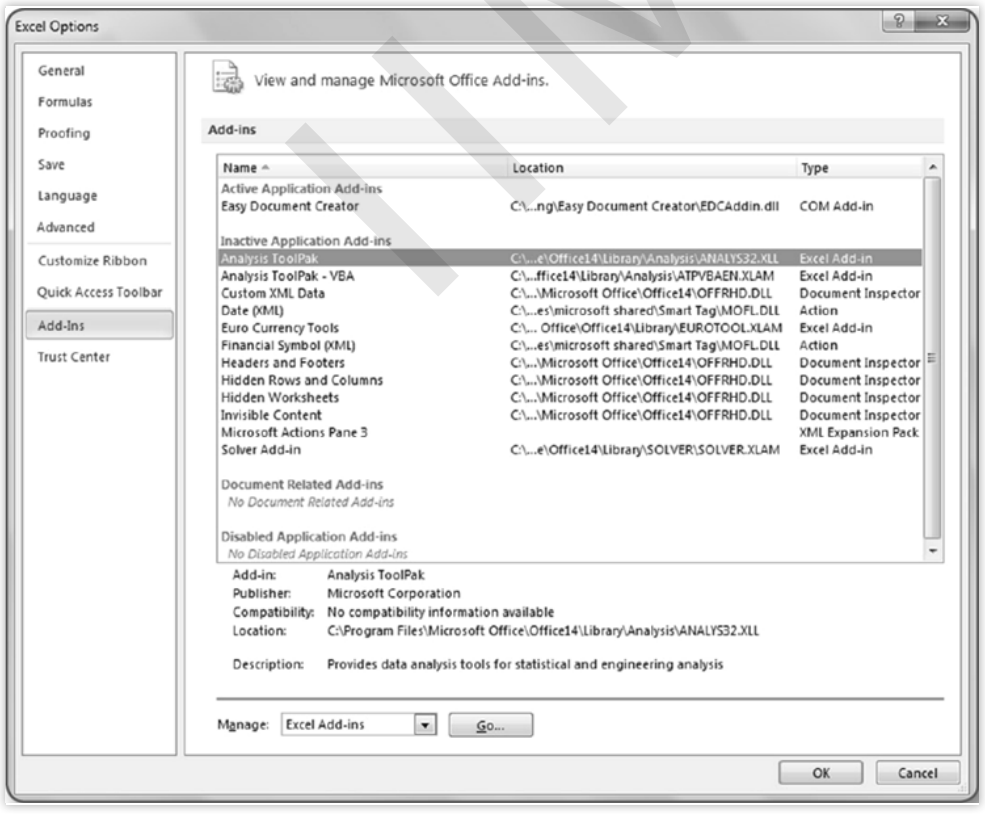



FIGURE 6: The Excel Options Dialog Box

NOTES

The **Add-Ins** dialog box appears.

5. *Select* the **Analysis ToolPak** option in the Add-Ins dialog box (Figure 7).
6. *Click* the **OK** button, as shown in Figure 7:


FIGURE 7: INCLUDING THE ANALYSIS TOOLPAK ADD-IN

7. *Click* the Data Analysis option in the Data tab. The Data Analysis dialog box appears (Figure 8).

   This will add the Data Analysis Add-ins in the Data tab (Figure 8).

8. *Click* the **Data Analysis** button in the Data tab, as shown in Figure 8:

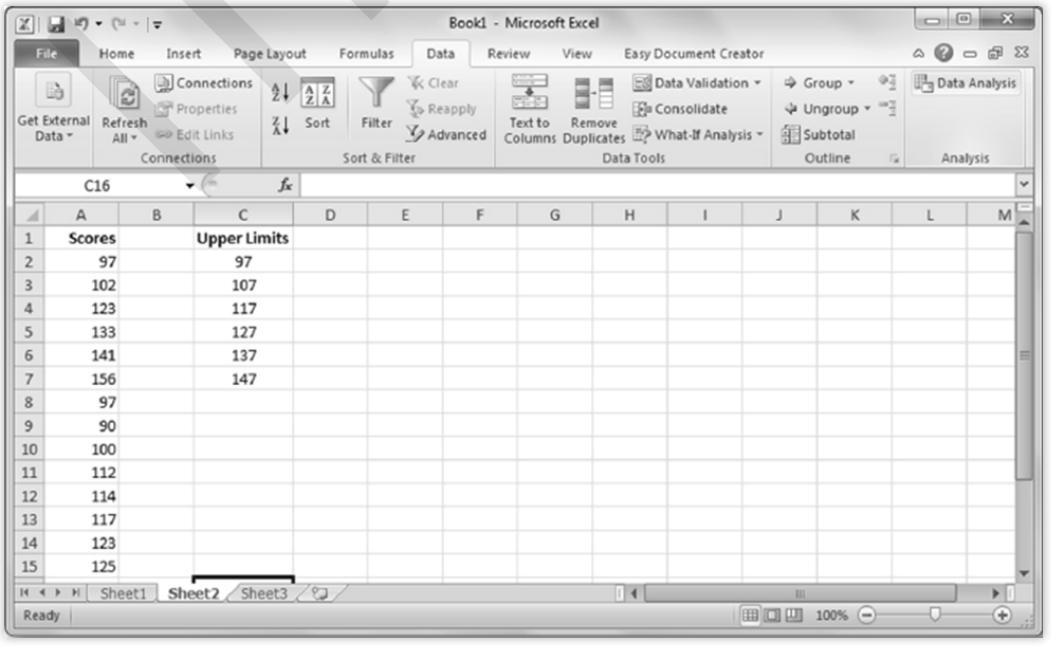



FIGURE 8: USING THE DATA ANALYSIS FEATURE

The Data Analysis dialog box appears (Figure 9).

9. *Select* the **Histogram** option in the dialog box (Figure 9).
10. *Click* the **OK** button, as shown in Figure 9:

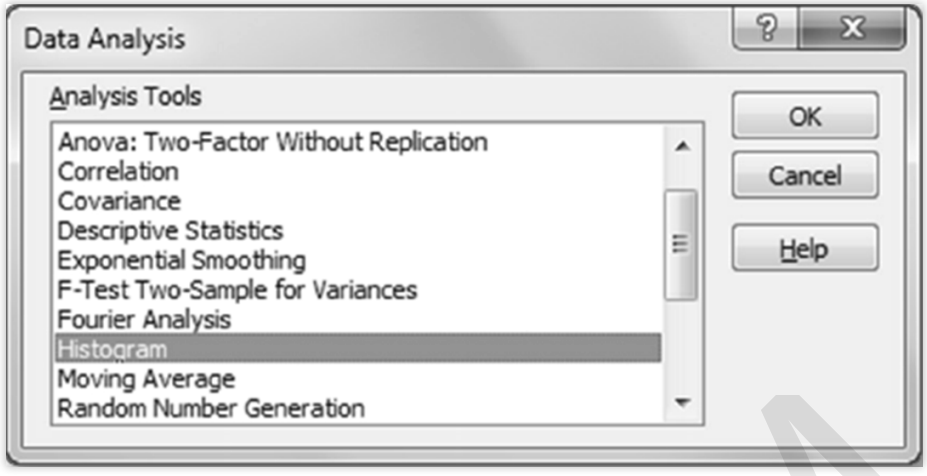



Figure 9: Using Histogram

The Histogram dialog box appears (Figure 10).

11. *Enter* the Input range from cells **A2:A15** (Figure 10).
12. *Enter* the Bin Range from cells **C2:C7** (Figure 10).
13. *Select* the **Chart Output** option (Figure 10).
14. *Click* the **OK** button, as shown in Figure 10:

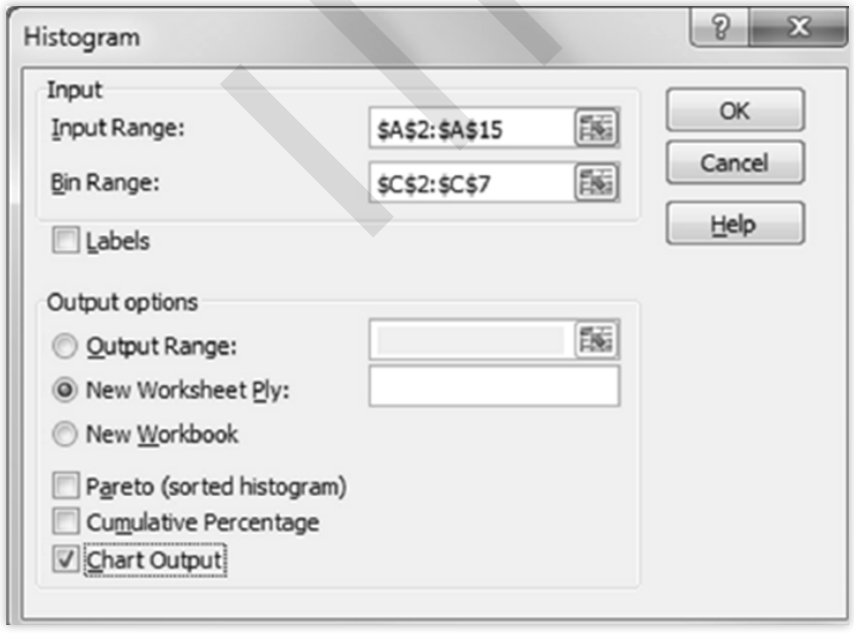



Figure 10: Setting Options in the Histogram Dialog Box

The histogram appears in the new sheet, as shown in Figure 11:

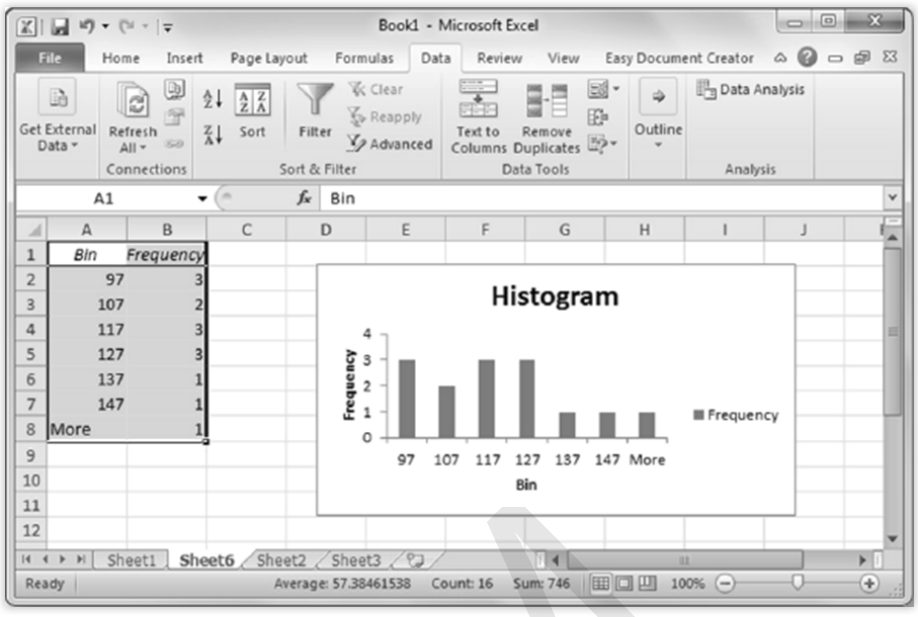



Figure 11: Displaying Frequency Distribution Using Histogram

Let us now discuss some of the most important probability distributions.

## 4.3.3 CLASSICAL PROBABILITY DISTRIBUTIONS

Some of the most prominent distributions that are used frequently are discussed as follows:

### Binomial Distribution

An experiment with only two possible outcomes repeated n number of times is called binomial. For example, if a coin is tossed once, it may result in heads or tails. If the coin is tossed for a second or third time also, it may result in heads or tails. The result of second or third trial does not depend on the result of the previous trials. The success or failure may also be defined as gain or loss, or winning or losing.

Assume that we assign a random variable X to the number of times a person wins a toss. Then, if a coin is tossed n times, X may be 0, 1, 2, 3, ……, n – 1, or n. X can be any number depending upon the number of times the coin is tossed. Each coin toss may result in only two possible outcomes – head or tail. Any outcome can be considered as a gain or loss. For example, if the coin turns up a head, it may be considered as a win.

Probability of heads (win), p = Probability of tails (failure), and q = 0.5.

It must, however, be remembered that when there are only two outcomes, the probability of both the outcomes may or may not be equal.

Major characteristics of a Binomial Distribution are:

- There are 'n' different trials and each trial is independent of the other.
- For each trial, there are only two outcomes, success or failure.

- All the trials are identical and the probability of success and failure is the same for all the trials.
- There are three parameters in a binomial distribution, namely n, p and q, where n = number of trials; p = probability of success, and q = probability of failure.

Mathematically, a binomial distribution is represented as:

$$\frac{n!}{x!(n-x)!}p^x q^{n-x}$$

The mean and variance of a binomial distribution are calculated as:

Mean, $\mu = n*p$

Variance, $V(X) = n*p*q$

When the probability of success is equal to probability of failure, the binomial distribution curve looks as shown in Figure 12:

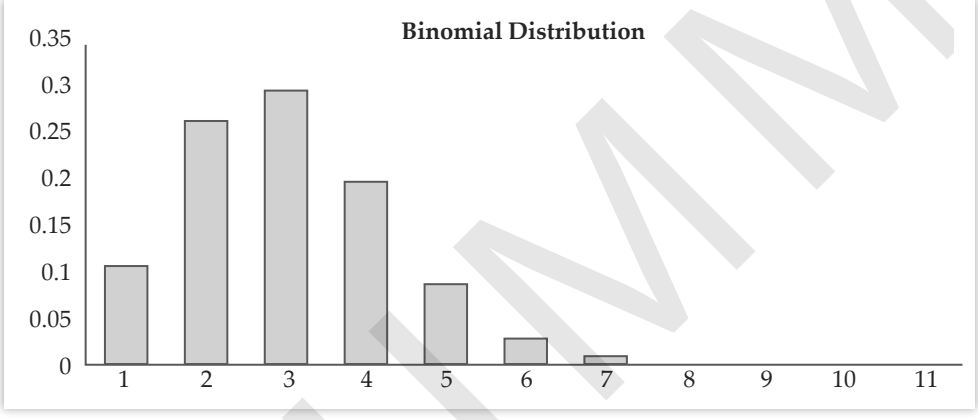



FIGURE 12: Binomial Distribution Curve when Probability of Success is Equal to Probability of Failure

When the probability of success is not equal to probability of failure, the binomial distribution curve looks as shown in Figure 13:

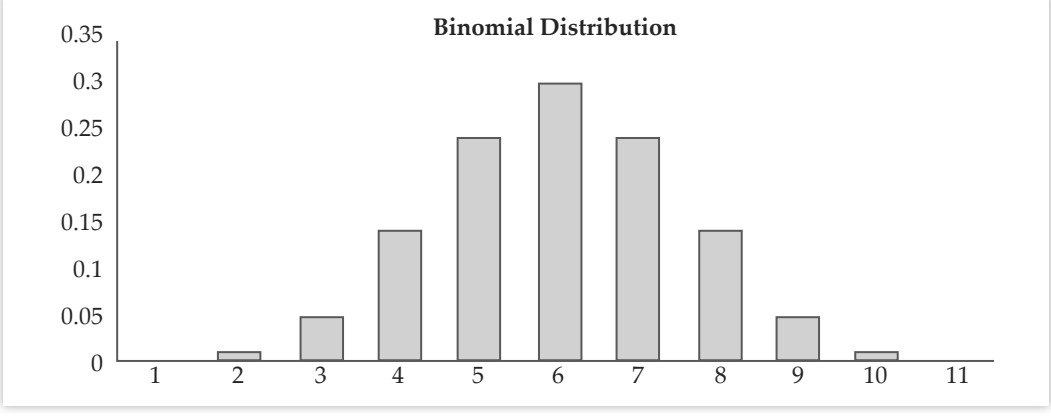



FIGURE 13: Binomial Distribution Curve when Probability of Success is not Equal to Probability of Failure

### Bernoulli Distribution

In a Bernoulli distribution, a random variable X can take value 1 (success) with probability p or can take value 0 with probability q (= 1 – p). For example, the toss of a coin once may result in heads or tails. In case of a fair coin, probability of heads = probability of tails = 0.5. The probability function for Bernoulli distribution = $p^x(1–p)^{1–x}$, where $x \in (0, 1)$.

Alternatively,

$$P(x) = p \text{ if } x = 1;$$

and

$$P(x) = q = 1 – p \text{ if } x = 0$$

In a Bernoulli trial, it is not necessary that both the outcomes will have equal probability like in case of a fair coin toss. For example, in a Karate competition between a karate green belt holder and a black belt holder, it is highly likely that the black belt holder would win. You may assume that probability of winning of black belt holder (success) = 0.9 and the probability of winning of green belt holder (failure) = 0.1.

It is known that the expected value of any distribution is the mean of the distribution. Therefore, for a Bernoulli distribution, the expected value of a random variable X is found as follows:

$$E(X) = 1*p + 0*(1 – p) = p$$

The variance of a random variable X from the normal distribution is calculated as:

$$V(X) = E(X^2) – [E(X)]^2$$

$$= p – p^2$$

$$= p(1 – p)$$

Some other examples of Bernoulli distribution include winning or losing in a game of chance, whether it will rain or not, whether an earthquake would happen tomorrow or not, etc.

### Uniform Distribution

In a uniform distribution, there may be any number of outcomes and the probability of getting any outcome is equally likely.

For example, when a fair dice marked A, B, C, D, E, and F on 6 of its sides is rolled:

Probability of getting A = Probability of getting B = Probability of getting C = Probability of getting D = Probability of getting E = Probability of getting F = 1/6.

Assume that in one trial, n outcomes may turn up. Then, all the n number of possible outcomes of a uniform distribution are equally likely. The probability function for a uniform distribution is written as:

$$f(x) = \frac{1}{b – a}; \text{ for } -\infty < a \le x \le b < \infty$$

In a uniform distribution, a and b are parameters. A uniform distribution curve is shown in Figure 14:


Figure 14: A Uniform Distribution

In Figure 14, we observe that the uniform distribution is rectangular in shape. Due to this reason, it is also called rectangular distribution.

Assume that a cake shop sells 50–80 cakes everyday. Let us calculate the probability that the daily sales is between 65 and 75 cakes.

Probability that the daily sales will fall between 65 and 75 = (75 – 65)*[1/(80 – 50)] = 0.33

Probability that the daily sales is greater than 60 = (80 – 60)*[1/(80 – 50)] = 0.67

For a uniform distribution, mean and variance are calculated as:

$$E(X) = (a + b)/2$$

$$V(X) = (b - a)^2/12$$

A standard uniform distribution has parameters a = 0 and b = 1. The probability distribution function for a standard uniform distribution is written as:

$$f(x) = 1 \text{ if } 0 \leq x \leq 1$$

$$f(x) = 0 \text{ for all other cases}$$

### Normal Distribution

This distribution occurs naturally in many situations. For example, if an examination is conducted, most of the students would pass with average marks, a few will score extremely high and a few will score extremely low. If this is shown on a graph, half of the data will fall on the left of the average marks and half of the data would fall on the right side of the average marks. Many situations follow a normal distribution. Due to this reason, normal distribution is widely used in businesses. Some of the situations that follow normal distribution include heights of people, measurement errors, test scores, IQ scores, salaries, etc. Under normal distribution, we have an empirical rule that tells us what percentage of the data falls within a certain number of standard deviations from the mean. They are:

- 68% of the data falls within the range of mean ± standard deviation (SD)(σ)

Notes

- 95% of the data falls within the range of mean ± 2 σ
- 99.7% of the data falls within the range of mean ± 3 σ

This empirical rule can be represented as shown in Figure 15:

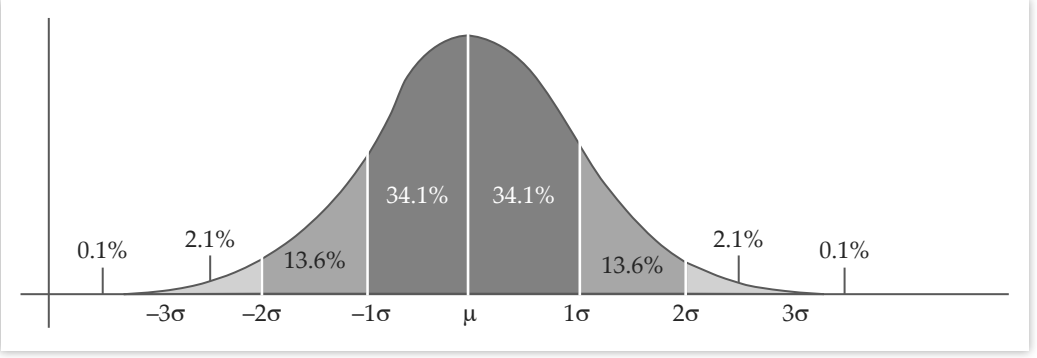



Figure 15: Empirical Rule for Normal Distribution

The spread of the normal deviation depends upon the standard deviation. If the standard deviation is low, much of the data will be accumulated around the mean and the distribution would appear taller. On the contrary, if the standard deviation is greater, the data would be spread out from the mean position and the normal distribution would appear flatter and wider.

Consider an example of a student who has secured the following marks in different subjects out of 100:

80, 82, 81, 81, 83

Perform the following steps to calculate the standard deviation:

1. *Enter* the marks in the Excel sheet (Figure 16).
2. *Calculate* the mean of the marks (Figure 16).
3. *Calculate* the standard deviation, as shown Figure 16:

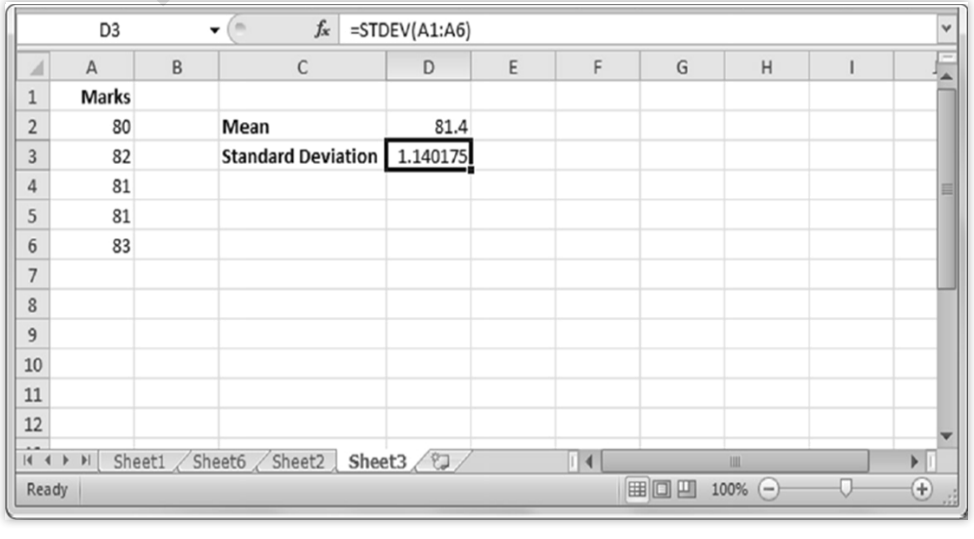



Figure 16: Calculating the Standard Deviation

In Figure 16, you can see that the marks are closer to their mean; therefore, the standard deviation is low. If the mean is not closer to the associated values, then the standard deviation will be high.

Important characteristics of a normal distribution are as follows:

- Mean, median and mode of normal distribution are equal.
- Normal distributions are symmetric at the centre around the mean.
- From among the entire data set, exactly half lie to the left of the centre and the remaining half lie to the right of the centre.
- Curve is bell-shaped.
- Total area under the curve is 1.

If the number of trials in a binomial distribution reaches infinity, the shape of the binomial distribution would start appearing similar to the normal distribution curve. The probability distribution function for a random variable X for a normal distribution is given as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left\{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}} \text{ for } -\infty < x < \infty$$

For a random variable that is normally distributed, the mean and variance are given as:

Mean, $E(X) = \mu$

Variance, $V(X) = \sigma^2$

A standard normal distribution is defined as a distribution with mean 0 and deviation 1.

The PDF of a standard distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{-x^2}{2}\right\}} \text{ for } -\infty < x < \infty$$

A standard normal distribution is shown in Figure 17:

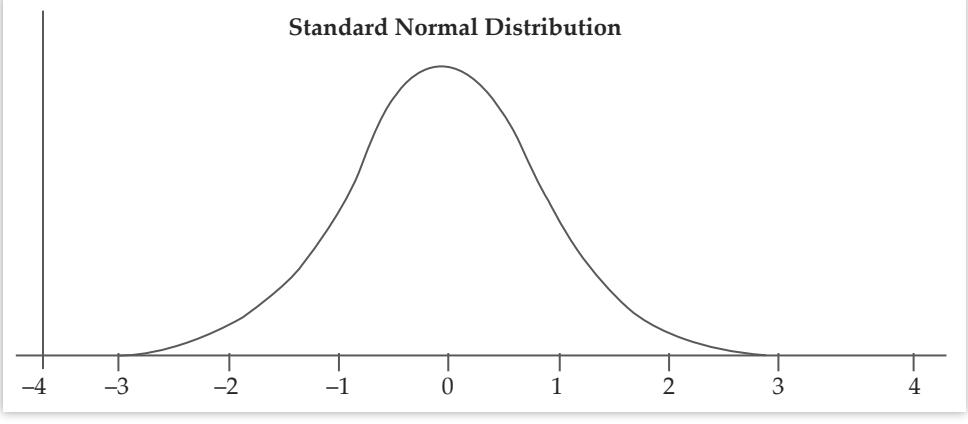



Figure 17: A Standard Normal Distribution

Notes

**Self Assessment Questions**

5. The probability space assigns a value between 0 and 1 to a set of outcomes which are called _____.
6. Which one of the following refers to the set of probabilities of the possible values of a continuous random variable?
   a. Probability distributions
   b. Discrete probability distributions
   c. Bernoulli distribution
   d. Continuous distribution
7. The probability function for Bernoulli distribution is _____, where $x \in (0, 1)$.

**Activity**

Make a PowerPoint presentation on probability distributions.

## 4.4 STATISTICAL INFERENCE

Statistical inference is the process of drawing conclusions or making predictions about a population based on a sample taken from that population. It involves using statistical techniques to analyze data and make inferences or generalizations about parameters, trends, or relationships within the larger population. The fundamental goal is to utilize the information obtained from a limited sample to make educated assessments or decisions about the entire population from which the sample was drawn.

Central to statistical inference is the consideration of uncertainty. As sample data are prone to randomness and variability, statistical inference aims to quantify this uncertainty and provide measures of confidence or probability associated with the conclusions drawn. Techniques such as hypothesis testing, estimation, and regression analysis are employed to derive meaningful insights from the data and make informed judgments about the population characteristics or future outcomes.

Statistical inference is crucial in various fields, including science, economics, social sciences, and healthcare, where decision-making often relies on drawing reliable conclusions from limited information. It serves as a powerful tool in extracting meaningful insights from data, aiding in understanding phenomena, making predictions, and guiding evidence-based decision-making.

### 4.4.1 TYPES OF INFERENCE

Statistical inference involves drawing conclusions or making predictions about a population based on a sample taken from that population.

There are two main types of statistical inference as shown in Figure 18:

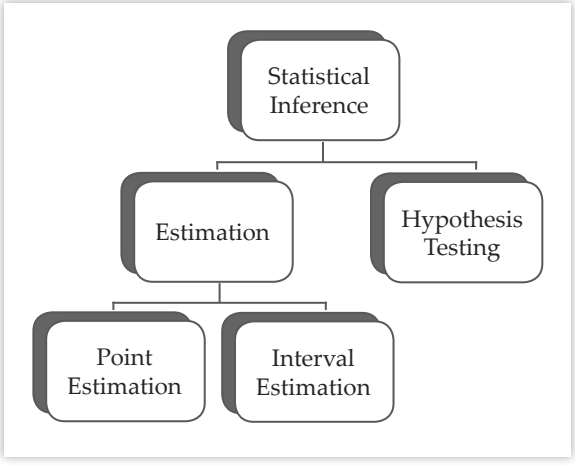



Figure 18: Types of Inference

Estimation in statistics involves determining unknown characteristics or parameters of a population based on sample data. There are two primary types of estimation: Point Estimation and Interval Estimation.

- **Point Estimation**: Point estimation involves using sample data to estimate a single value of an unknown population parameter. For instance, estimating the population mean, variance, proportion, etc., based on the sample data.

  **Example**:

  Given a sample of size n from a population, the point estimate of the population means (μ) is given by the sample mean ($\bar{x}$):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  **where,**

  - $\bar{x}$ is the sample mean
  - $x_i$ represents individual observations in the sample
  - n is the sample size

- **Interval Estimation**: Interval estimation provides a range of values within which the population parameter is likely to fall. It gives a measure of the uncertainty associated with the estimate by providing a confidence interval. Interval estimation involves constructing confidence intervals within which the population parameter is expected to lie.

  **Example**:

  For estimating the population mean μ with a confidence interval, it's often done using the sample mean ($\bar{x}$) and the standard error (SE).

Notes

The formula for a confidence interval for the population mean (μ) with a confidence level of $1 - \alpha$ is:

$$x \pm Z\frac{\sigma}{\sqrt{n}}$$

where,

- $\bar{x}$ is the sample mean
- Z is the Z-score corresponding to the desired confidence level
- σ is the Population standard deviation
- n is the sample size

### 4.4.2 PARADIGMS OF INFERENCE

Statistical inference operates within distinct paradigms, each offering unique perspectives on drawing conclusions from data. While these paradigms—such as classical, bayesian, likelihoods, and those based on the Akaike Information Criterion (AIC) —have their individual principles, they can often provide complementary insights

Bandyopadhyay & Forster delineate four major paradigms: classical (frequentist), Bayesian, likelihoods, and the Akaike Information Criterion-based paradigm. These paradigms aren't exclusive, and methodologies that excel in one paradigm can often offer compelling explanations within others.

❍ **Frequentist Inference**

- Frequentist inference assesses propositions by envisioning repeated samplings from a population distribution similar to the observed data. It quantifies statistical properties through these hypothetical repeated samplings, although practical quantification might pose challenges.
- **Mathematical Basis**: In frequentist inference, probabilities relate to the frequencies of events in the long run.
- **Key Concepts**: Probability is associated with the relative frequency of events. It involves using sample statistics to estimate population parameters.
- **Representation**:

  ✓ **Parameter Estimation**: For parameter θ in a statistical model: $\widehat{\theta_{MLE}} =$, $\text{argmax}_\theta L(\theta|x)$, where $L(\theta|x)$ is the likelihood function.

  ✓ **Confidence Interval**: $\theta \in [\hat{\theta} - z \cdot SE, \hat{\theta} + z \cdot SE]$, where $\hat{\theta}$ is the point estimate and SE is the standard error.

  ✓ **Hypothesis Testing**: Using the p-value to evaluate the strength of evidence against a null hypothesis.

❍ **Bayesian Inference**

- Bayesian inference employs probability to express degrees of belief, treating beliefs as probabilities obeying certain axioms. It hinges on posterior beliefs to formulate statistical propositions, using various justifications for its approach.

- **Mathematical Basis**: Bayes' Theorem links prior beliefs P(θ) with observed data (P(x|θ)) to yield posterior beliefs (P(θ|x)).
- **Key Concepts**: Probability represents degrees of belief, updating prior beliefs with observed data to obtain posterior beliefs.
- **Representation**:
  - ✓ **Posterior Estimation:** $P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$ using Bayes' Theorem.
  - ✓ **Credible Interval**: An interval containing a specified probability mass in the posterior distribution.
  - ✓ **Bayes Factor**: $BF = \frac{P(x|M_1)}{P(x|M_2)}$ compares the likelihood of data under different models ($M_1$ and $M_2$).

❍ **Likelihood-Based Inference**

- Likelihood-based inference revolves around estimating model parameters from observed data using the likelihood function. This function quantifies the probability of observing the given data under specific parameter values, aiming to maximize this likelihood.
- **Mathematical Basis**: Focuses on maximizing the likelihood function.
- **Key Concepts**: Likelihood represents the probability of observed data given parameter values.
- **Representation**:
  - ✓ **Likelihood Function**: $L(\theta|x) = P(x|\theta)$, where x is the observed data and θ are the parameters.
  - ✓ **Maximum Likelihood Estimation (MLE)**: $\widehat{\theta_{MLE}} = \text{argmax}_\theta L(\theta|x)$, maximizes the likelihood function.

❍ **Akaike Information Criterion (AIC) Based Inference**

- The AIC evaluates the relative quality of statistical models for a dataset, balancing model fit and complexity. It aids in model selection by estimating the information loss when using a particular model to represent the data-generating process.
- **Mathematical Basis**: AIC evaluates the relative quality of models based on information theory.
- **Key Concepts**: Balances model fit and complexity to select the best-fitting model.
- **Representation**:
  - ✓ **AIC Calculation**: $AIC = -2 \cdot \ln(\hat{L}) + 2k$, where $\hat{L}$ is the maximum likelihood of the model and k is the number of parameters.

Notes

Beyond these established paradigms, other approaches like Minimum Description Length (MDL) and Fiducial inference offer alternative perspectives on statistical inference. MDL focuses on maximal data compression without assuming specific data-generating mechanisms, while Fiducial inference, once prominent, has been challenged for its limited applicability. Structural inference, inspired by Fisher and Pitman's ideas, emphasizes invariant probabilities within group families for making statistical inferences.

**Self Assessment Questions**

8. Which paradigm of statistical inference employs probability to express degrees of belief, treating beliefs as probabilities obeying certain axioms?
   a. Frequentist Inference
   b. Bayesian Inference
   c. Likelihood-Based Inference
   d. AIC-Based Inference
9. In point estimation, what does the confidence interval provide?
   a. A single value estimate of the population parameter
   b. A range of values likely to contain the population parameter
   c. A measure of uncertainty associated with the estimate
   d. Both a single value estimate and a range of values
10. Likelihood-based inference primarily focuses on maximizing the posterior distribution. (True/False)
11. The AIC evaluates the relative quality of statistical models for a dataset by balancing model fit and _____.

**Activity**

Make a PowerPoint presentation on statistical inference.

## 4.5 HYPOTHESIS TESTING

There are basically two types of statistical inferences. One is estimation and the other is hypothesis testing. Before we discuss hypothesis testing, let us first discuss what a hypothesis is.

A hypothesis statement is usually associated with the population parameters. A hypothesis can be tested using a research method. For example, a school principal may want to test the hypothesis that the average number of leaves (per month) of Class 8$^{th}$ students is two. Hypotheses are statements that may be evaluated by appropriate statistical techniques. In hypothesis testing, there are two types of hypotheses, namely null hypothesis and alternate hypothesis. The null hypothesis ($H_0$) is the hypothesis to be tested. Alternate hypothesis ($H_A$) is the hypothesis that must be accepted if the sample data leads to rejection of $H_0$.

Hypothesis testing, also called significance testing, is a method which is used to test the hypothesis regarding the population parameters using the data collected from a sample.

In hypothesis testing, we test a hypothesis by determining the likelihood that a sample statistic would be selected if the hypothesis was true. For example, assume that a study published in a journal claims that Indians aged between 25 and 40 years of age sleep for an average of 6 hours. To test this claim made in the study for Indians aged between 25 and 40 years of age living in Bengaluru, we may first record the average sleeping time of 100 (sample size) Bengaluru-based Indians aged between 25 and 40 years of age. The average value of sleeping hours calculated for these 100 people is the sample mean. Next, we can compare the sample mean with the population mean.

### 4.5.1 FOUR STEPS TO HYPOTHESIS TESTING

The process of hypothesis testing consists of four steps as follows:

**Step 1: Identify the hypothesis to be tested.**

For example, the experimenter may want to test if the Bengaluru-based people in the age group of 25 and 40 years sleep for 6 hours on an average.

$H_0$; $\mu = 6$

**Step 2: Set the criterion upon which the hypothesis would be tested.**

For example, if we consider the hypothesis that Bengaluru-based people in the age group of 25 and 40 years sleep for 6 hours on an average, then the sample so selected should have a mean close to or equal to 6 hours. However, if the Bengaluru-based people in the said age group sleep for more than or less than 6 hours, the sample should have a mean distant from 6 hours.

However, here it is important to describe how much difference or deviation from 6 hours would make the experimenter reject the hypothesis. For example, the mean of $6 \pm 0.5$ hours may be acceptable, whereas the mean of $6 \pm 0.51$ hours onwards may not be acceptable.

**Step 3: Select a random sample from the population and measure the sample mean (Compute the test statistic).**

For example, a sample of 100 Bengaluru-based people in the age group of 25 and 40 years is selected at random and the mean time of their sleep is measured.

**Step 4: Make a decision – Compare the observed value of the sample to what we expect to observe if the claim we are testing is true.**

For example, if the sample mean calculated is approximately 6 hours with a small discrepancy between the population and sample mean, then the experimenter may decide to accept the hypothesis, else if the discrepancy is large, the hypothesis would be rejected.

Notes

## 4.5.2 HYPOTHESIS TESTING AND SAMPLING DISTRIBUTIONS

The sampling distributions are the basis of hypothesis testing. In other words, hypothesis testing is another use of sampling distributions. For understanding the relation between hypothesis testing and sampling distributions, two important characteristics of mean are as follows:

- Sample mean is an unbiased estimator of the population mean. If a sample is selected at random, it would have a mean equal to the mean of the population. The null hypothesis is the beginning of hypothesis testing. If the null hypothesis is true, then random sample selected from a given population will have sample mean equal to the value stated in null hypothesis.
- The sampling mean is usually normally distributed regardless of the type of distribution. This sample distribution can be used to state an alternate hypothesis to locate the probability of obtaining sample means with less than 5% chance of being selected if the value stated in null hypothesis is true.

For calculating the probability of obtaining sample mean in a sampling distribution, the population mean and the Standard Error of the Mean (SEM) must be known.

These values are input in the test statistic formula calculated in step 3. The notations used to describe populations, samples and sampling distributions are shown in Table 3:

**Table 3: Notations Used for the Mean, Variance and Standard Deviation in Populations, Samples and Sampling Distributions**

| Characteristic | Population | Sample | Sampling Distribution |
|---|---|---|---|
| Mean | $\mu$ | M or $\bar{x}$ | $\mu_M = \mu$ |
| Variance | $\sigma^2$ | $s^2$ or $SD^2$ | $M^2 = \frac{\sigma^2}{n}$ |
| Standard Deviation | $\sigma$ | s or SD | $\sigma_M = \frac{\sigma}{\sqrt{n}}$ |

We must know the important differences between population, sample, and sampling distributions.

They are described in Table 4:

**Table 4: Differences between Population, Sample and Sampling Distributions**

| Population Distribution | Sample Distribution | Distribution of Sample Means |
|---|---|---|
| Scores of all persons in a population | Scores of a select number of persons from the population | All the possible sample means that can be selected given a certain sample size |
| They are generally not accessible | They are accessible | They are accessible |

| Population Distribution | Sample Distribution | Distribution of Sample Means |
|---|---|---|
| It could be distributed in any shape | It could be distributed in any shape | It is usually distributed normally |

## 4.5.3 TYPES OF ERRORS

You studied that there are four steps in hypothesis testing. In the fourth stage, the experimenter decides whether to accept or reject the null hypothesis. Since a sample is used to observe the population, the decision taken regarding the null hypothesis may be wrong. When a decision is taken regarding a sample, there may be four decision alternatives as follows:

- Decision to retain null hypothesis is correct.
- Decision to retain null hypothesis is not correct.
- Decision to reject null hypothesis is correct.
- Decision to reject null hypothesis is not correct.

It is impossible to know the truth regarding a population in the absence of a census. Therefore, in the presence of the sample, we assume that we know about the exact status of the population. We label it as truth in Table 5 which depicts the four outcomes of making a decision:

**Table 5: Four Outcomes of Making a Decision**

| Truth in the Population | Decision to Retain Null Hypothesis | Decision to Reject Null Hypothesis |
|---|---|---|
| True | Correct $(1 - \alpha)$ | Type I Error $\alpha$ |
| False | Type II Error $\beta$ | Correct $(1 - \beta)$ Power |

The four decisions may be: correctly retaining the null hypothesis, correctly rejecting the null hypothesis, incorrectly retaining the null hypothesis and incorrectly rejecting the null hypothesis.

In the context of hypothesis testing, there are usually two types of errors as follows:

- **Type I Error:** It is the probability of rejecting a null hypothesis that is actually true. This error is depicted using symbol $\alpha$. Researchers directly control the probability of committing this type of error by stating an alpha level.
- **Type II Error:** It is the probability of incorrectly retaining a null hypothesis. This error is depicted using symbol $\beta$.

Let us now analyse the two decision types, namely retaining the null hypothesis and rejecting the null hypothesis.

### Retaining the Null Hypothesis

When the researcher decides to retain a null hypothesis, the decision may be correct or incorrect. The correct decision here is to retain a true null hypothesis (null result). In this case, we are retaining what we had already assumed. At times, the researcher

may make an incorrect decision to retain a false null hypothesis. This is a Type II (β) error. In most tests, there is a probability of making Type II error because the experimenters do not reject the previous notions of truth that are, in fact, false. Type II error is less problematic than Type I error, but it might be problematic in fields, such as medicine and defence. Testing of defence equipment or medicine may involve accepting null hypothesis that should have been rejected. This may even put a risk on the lives of the patients and other individuals.

### Rejecting the Null Hypothesis

When the researcher decides to reject a null hypothesis, the decision may be correct or incorrect. In the first case, a true null hypothesis may be rejected, leading to a false positive finding. This is a Type I (α) error. In most tests, there is a probability of making Type I error because the experimenters reject the previous notions of truth that are, in fact, true. For example, finding an innocent person guilty is a Type I error. To minimise this error, the burden is placed on the researcher to demonstrate that the null hypothesis is false. Since here the researcher assumes that the null hypothesis is true, he controls for Type I (α) error by stating a level of significance α. To minimise the Type I (α) error, a lower value of α should be used. While making a decision, the α value can be compared to the p value (likelihood of obtaining a sample mean if the null hypothesis was true). When p value is less than α value of 0.05, the experimenters reject the null hypothesis, else the null hypothesis is retained. Here the correct decision is to reject a false null hypothesis or deciding that the null hypothesis is false when it is actually false.

## 4.5.4 EFFECT SIZE, POWER AND SAMPLE SIZE IN HYPOTHESIS TESTING

In hypothesis testing, three terms, namely power, effect size and sample size are often used. Let us now study about these.

Whenever an experimenter wants to use inferential statistics to analyse the evaluation results, first of all, he/she should conduct a power analysis to determine the required size of sample. Whenever we are conducting an inferential statistics test, we are basically comparing two hypotheses, i.e., the null hypothesis and the alternate hypothesis. For example, a null hypothesis may state that when a group of students are taken for an environment scanning and conservation trip, the attitude towards environment conservation before and after going to the trip will remain the same. On the contrary, the alternate hypothesis may state that there is a significant difference between the attitude of students before and after going to the trip. Usually, the statistical tests look for tests that can allow you to reject the null hypothesis and conclude that the programme had an effect. In any statistical test, there exists a possibility that there will be a difference between groups when there exists none. This is type I error. Similarly, there is a possibility that the test will not be able to identify a difference when it does exist. This is type II error.

Power is the probability that a statistical test will find a significant difference when such difference exists. In general, a power of 0.8 or more is considered as a standard. It means that there should be an 80% or more chance of finding a statistically significant difference when the difference actually exists.

The experimenter can use power calculations to determine the sample size. There is a relation between sample size and power. As the size of sample increases, the power of test also increases. This is so because when a large sample is collected, more information is available and it makes it easier to correctly reject the null hypothesis. In order to ensure that a sample size is sufficiently large, power analysis calculation should be conducted. For a power calculation, following must be known:

- Type of inferential test that must be used
- Alpha value or significance level being used
- Expected effect size
- Sample size you are planning to use

These values are input in statistical software to calculate the value of power. The power value comes out to be between 0 and 1. In case the power is less than 0.8, it is recommended that the sample size is increased.

On entering these values, a power value ranging between 0 and 1 will be generated. If the power is less than 0.8, you will need to increase your sample size.

### Statistical Significance

Let us continue our previous example of a group of students going to an environment scanning and conservation trip. In such examples, there is a possibility that the students' knowledge, attitude and behaviour might change due to chance rather than the trip itself. Testing for statistical significance helps the experimenter estimate how likely it is that these changes occurred randomly and not due to the programme. To determine whether the difference is statistically significant or not, the p-value must be compared with the critical probability value or the alpha value. If the p-value is less than the alpha value, then it can be concluded that the difference is statistically significant. The p-value is the probability that the results so obtained were due to chance and not due to the programme. The p-value can be in the range of 0 and 1. If the p-value is lower, it has higher probability that the difference has occurred as a result of the program. Alpha ($\alpha$) level (Type I error) refers to the error rate that an experimenter is willing to accept. Usually, the alpha level is set at .05 or .01. An alpha of 0.05/0.01 means that the experimenter is willing to accept a 5%/1% chance that the results are due to chance and not due to the programme.

0.05 is the most common alpha level chosen by experimenters for hypothesis testing in social science field. It is considered as statistically significant.

### Effect Size

A statistically significant difference does not mean that it is big or helpful in decision-making. It only means that there exists a difference. For example, assume that a group of students from a population are selected and a pre-programme test is conducted. The mean score obtained by students is 85. After an improvement programme is implemented, the students are again tested on a post-programme test. The mean score obtained by the students is 85.5. Here the difference is statistically significant due to large sample size, but the difference in scores is very low, which indicates that the improvement programme did not lead to a meaningful increase in

Notes

the knowledge of students. It can be concluded that the difference among the means must be statistically significant and also meaningful.

To determine whether or not an observed difference is statistically significant and meaningful, its effect size must be calculated. Effect size is a standardised measure and it is calculated on a common scale, which allows for comparing the effectiveness of different programmes on the same outcome. Let us now see how we can calculate the effect size depending on the evaluation design. To calculate the effect size, the difference between the test and control groups is taken and is divided by the standard deviation of one of the groups. For example, in a medical hypothesis testing, the difference of the means of the test and control groups is calculated and divided by the standard deviation of the control group.

$$\text{Effect Size} = \frac{\text{Mean of treatment group} - \text{Mean of control group}}{\text{Standard deviation of control group}}$$

Now the value of effect size is calculated as follows:

- Effect size < 0.1 ⇒ Trivial effect
- Effect size between 0.1 and 0.3 ⇒ Small effect
- Effect size between 0.3 and 0.5 ⇒ Moderate effect
- Effect size > 0.5 ⇒ Large difference effect

The effect size can only be calculated after collecting data from all the objects or subjects in the sample. Therefore, an estimate for the power analysis must be derived. Most commonly used value for moderate to large difference is 0.5.

## 4.5.5 | t-TEST

In many research studies, researchers may want to find out differences between various calculated values. For example, there may be situations where the researcher may need to find out the difference between the sample mean and the population mean (one-sample t-test). Similarly, in other situations, a researcher may need to find out the difference between two independent sample means (independent samples t-test) or the difference between pre- and post-event outcomes (paired samples t-test). Let us now discuss the three different types of t-tests as follows:

- **One-Sample t-test:** To test the difference between sample mean and population mean
- **Independent-Sample t-test:** To test the difference between two independent sample means
- **Paired-Sample t-test:** To test the difference between pre- and post-event outcomes

### One-Sample t-Test

In many situations, we come across claims made by the marketeers about their products. For example, a car manufacturer may claim that the average mileage of a car is, say, 19.9 kmpl or a business school may claim that the average package offered to its students is ₹ 12 lakhs per annum. A researcher may be interested in analysing the truthfulness of these claims. For this analysis, the researcher needs to randomly

pick a small sample from the population and compare its mean with the claimed population mean. The sample mean and the population mean may be different from each other. In order to test whether this difference is statistically significant, we should apply one-sample t-test.

The null hypothesis of one-sample t-test is:

'$H_0$: There is no significant difference between sample mean and population mean.'

The t-statistic in one-sample t-test can be estimated by using the following formula:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma}{N-1}}}$$

where, X = sample mean, μ = population mean, σ = standard deviation of sample mean and N = sample size.

### Independent-Sample t-Test

When we want to test the difference between two independent sample means, we use independent-sample t-test. The independent samples may belong to the same population or different population. Some of the instances in which the independent-samples t-test can be used are as follows:

1. Testing difference in the average level of performance between employees with the MBA degree and employees without the MBA degree.
2. Testing difference in the average wages received by labour in two different industries.
3. Testing difference in the average monthly sales of the two firms.

The null hypothesis of independent-sample t-test is:

'$H_0$: There is no significant difference between sample means of two independent groups.'

The t-statistic in the case of independent-sample t-test can be calculated by using the following formula:

$$t_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

where, $N_1$ and $N_2$ are the sample sizes of two independent samples.

In SPSS, the independent-sample t-test is conducted in two stages. At stage one, SPSS software compares variances of two samples. The statistical method of comparing two sample variances is known as Levene's homogeneity test of variance. The null hypothesis of this test is 'equal variances assumed', i.e., there are no significant differences between the sample variances of two independent samples. In other words, the two samples are comparable. On the basis of Levene's test of homogeneity,

the SPSS gives two values of t-statistic. In case of equal variances, both the values are the same. In case the sample variances are different, the lower t-statistic value should be considered for final analysis.

### Paired-Sample t-Test

A paired-sample t-test should be used when we want to test the impact of an event or experiment on the variable under study. In this case, the data is collected from the same respondents before and after the event. After this, means are compared.

The null hypothesis of paired-sample t-test is that the means of pre-sample and post-sample are equal. Some of the instances where paired sample t-test can be applied are as follows:

- Analysing the effectiveness of training programme on the performance of employees of a business enterprise.
- Analysing the impact of a new advertisement on the sales of a product.
- Analysing the impact of a policy on the volatility in the stock market.
- Analysing the difference of responses of the same group to the two different treatments.

## 4.5.6 | ANALYSIS OF VARIANCE (ANOVA)

Independent sample t-test can be applied to situations where there are only two independent samples. In other words, we can use independent-sample t-tests for comparing the means of two populations (such as males and females). When we have more than two independent samples, t-test is inappropriate. The Analysis of Variance (ANOVA) has an advantage over t-test when the researcher wants to compare the means of a larger number of population (i.e., three or more). It helps in explaining the amount of variation in the dataset. In a dataset, two main types of variations can occur. One type of variation occurs due to chance and the other type of variation occurs due to specific reasons. These variations are studied separately in ANOVA to identify the actual cause of variation and help the researcher in taking effective decisions.

In case of more than two independent samples, the ANOVA test explains three types of variance. These are as follows:

- Total variance
- Between group variance
- Within group variance

The ANOVA test is based on the logic that if the between group variance is significantly greater than the within group variance, it indicates that the means of different samples are significantly different.

There are two main types of ANOVA, namely one-way ANOVA and two-way ANOVA. One-way ANOVA determines whether all the independent samples (groups) have the same group means or not. On the other hand, two-way ANOVA

is used when you need to study the impact of two categorical variables on a scale variable.

In case of more than two independent samples, the sample means can be compared with the help of multiple t-tests. However, still ANOVA is preferred over multiple t-tests. The basic reason of this preference of ANOVA test over multiple t-tests is the presence of a family-wise error in case of multiple t-tests. Suppose that we are interested in comparing the sample means of three independent samples A, B and C. If we are interested to apply t-test, it requires three independent sample t-tests:

- Between A and B
- Between B and C
- Between A and C

If the level of significance in each test is 5 per cent, the confidence level is 95 percent. If we assume that the three independent-sample t-tests are independent, an overall confidence level of all t-tests together will be:

$$\text{Overall confidence level} = 0.95 \times 0.95 \times 0.95 = 0.857$$

Hence, the combined probability of committing type I error in multiple t-tests is = 1 – 0.857 = 0.143 or 14.3 per cent. Therefore, the probability of making type I error increases from 5 per cent to 14.3 per cent in multiple t-tests. This error is known as the family-wise error rate. The family-wise error rate can be calculated using the generalised method in which n represents the number of tests carried out in data:

$$\text{Family-wise error} = 1 - (0.95)^n$$

Because of the presence of a family-wise error, the ANOVA test is always preferred to multiple t-tests.

Various examples where one-way ANOVA test can be used are as follows:

- To test the difference in the level of product usage among the citizens of four different cities
- To test the difference in the performance level among the respondents of different educational backgrounds
- To test whether the average income of different professionals is different

In case of t-test, the null hypothesis is that there is no difference between two samples' means, that is, the two samples' means are equal. Similarly, in case of ANOVA test, the null hypothesis is that all group means are equal.

### F-Statistics

Similar to t-statistics in t-tests, the ANOVA procedure calculates F-statistics, which compare the systematic variance in the data (between group variance) to the unsystematic variance (within group variance). As F-distribution is the square of t-distribution, assuming that the assumptions of parametric tests hold true, any value of F-statistics more than 3.96 is sufficient to reject the null hypothesis with 5 per cent level of significance.

NOTES

### Combined Test

ANOVA is a combined test. It indicates that the rejection of the null hypothesis implies that all group means are not the same. But, it may be possible that some group means are the same and some are not. For example, if there are three groups, rejection of the null hypothesis means that all group means are not equal. This is a confusing statement because of the following possibilities:

**Null hypothesis**: All group means are the same, that is, $\overline{x}_1 = \overline{x}_2 = \overline{x}_3$

**Alternate hypothesis**: All group means are not equal and have the following possibilities:

$$\overline{x}_1 \neq \overline{x}_2 = \overline{x}_3$$
$$\overline{x}_1 = \overline{x}_2 \neq \overline{x}_3$$
$$\overline{x}_1 \neq \overline{x}_2 \neq \overline{x}_3$$
$$\overline{x}_1 \neq \overline{x}_3 = \overline{x}_2$$

In order to go in much detail, it is required to apply post-hoc tests along with ANOVA.

**SELF ASSESSMENT QUESTIONS**

12. A hypothesis statement is usually associated with the population parameters. (True/False)
13. The probability distributions are the bases of hypothesis testing. (True/False)
14. Independent-Sample t-test is used to test the difference between sample mean and population mean. (True/False)
15. Paired-Sample t-test is used to test the difference between pre- and post-event outcomes. (True/False)
16. The Analysis of Variance (ANOVA) helps in explaining the amount of variation in the dataset. (True/False)

**ACTIVITY**

Search and enlist the differences between one-way ANOVA and two-way ANOVA in a tabular format.

## 4.6 CORRELATION

Correlation is a statistical technique to show the association between a pair of variables. Let us try to understand this concept by determining the correlation between closely related values, such as the height and weight of human beings. It is generally observed that taller people are heavier as compared to people who are shorter in height.

People may have different weights even if their heights are same. It is also possible that a shorter person is heavier than a taller person. Sometimes, the mean weight of

a group of persons of height 5'3'' is less than the mean weight of a group in which each person is of height 5'4".

Here, you see that correlation can work appropriately with specific kinds of data. It is mainly suitable for quantifiable data in which exact numbers or values are provided. This technique cannot be used for categorical data, such as favourite colour, brands of products purchased, gender, etc.

The result obtained using correlation is known as correlation coefficient and is represented by 'r'. It is also known as Pearson's correlation coefficient. The range of coefficient lies between –1.0 and +1.0. The closeness of r with +1 and –1 shows that two variables are closely related. If the value of r is close to 0, then it signifies that no relationship exists between two variables.

If the value of r obtained is positive, then it signifies that if one variable gets larger, then the other variable also gets larger. The negative value of r signifies that if one variable gets larger, then the other variable gets smaller, which is generally known as an inverse correlation. Figure 19 shows the graphical representation of the correlation coefficient:

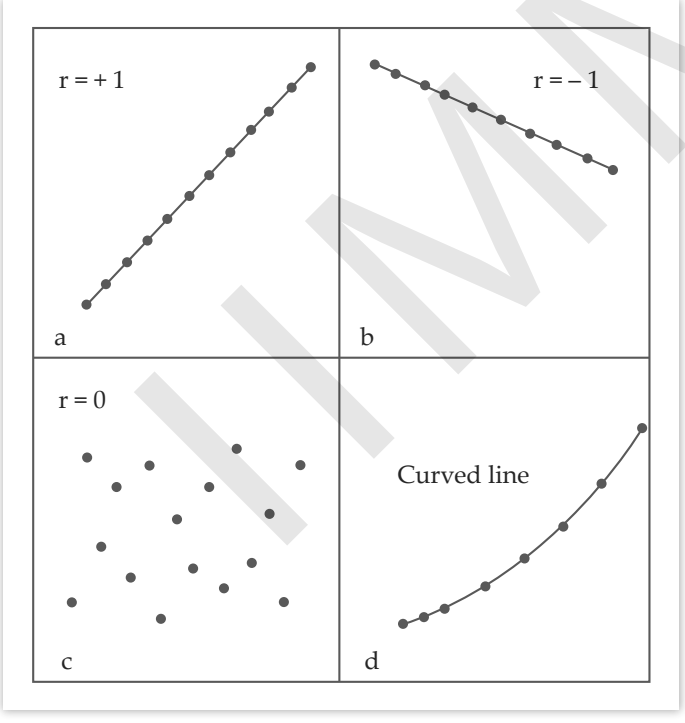



Figure 19: Displaying Graphical Representation of Correlation Coefficient

**Source:** *https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-andregression*

However, the correlation coefficient does not work well in case of the curvilinear relationship in which the relationship between variables cannot be represented using a straight line on a graph. For example, consider two variables, age and healthcare. These variables are related to each other; however, they do not follow a straight line. It is because older people and kids require more healthcare and attention as compared to teenagers or adults.

NOTES

**SELF ASSESSMENT QUESTIONS**

17. The result obtained using correlation is known as correlation coefficient and is represented by _____.
18. The range of coefficient lies between –1.0 and +1.0. (True/False)

## 4.7 REGRESSION ANALYSIS

Regression analysis is usually used to model a relationship between a response variable (dependent variable) and one or more predictor (independent) variables. There are various types of regression. However, the basic function of these regression models is to examine the influence of one or more independent variables on a dependent variable.

Regression analysis helps in identifying which variables have an impact on a variable of interest; for example, the impact of demand on supply, impact of money supply on inflation, etc. By performing regression analysis, we can determine the factors that matter the most, which factors have negligible impact (hence, can be ignored), and how these factors influence each other. To understand regression analysis, it is important to know the following:

- **Dependent variable:** Regression analysis is carried out to understand or predict the dependent variable.
- **Independent variables:** Regression analysis involves use of hypothesised factors or the independent variables that have an impact on the dependent variable.

### 4.7.1 SIMPLE REGRESSION ANALYSIS

The simple linear regression helps in summarising and studying relationships between two continuous quantitative variables. In linear regression, X = predictor/explanatory/independent variable, and Y = response/outcome/dependent variable.

The simple linear regression is depicted by the regression line:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where

$\beta_0$ = intercept of the line

$\beta_1$ = slope of the line

If $\beta_1$ is close to 0, it indicates little or no relationship, but if $\beta_1$= large positive or negative, it indicates large positive or large negative relationship.

$\varepsilon$ = Error term or disturbance

There exists a broadly linear relation between x and y when all the (x, y) pairs do not lie on the straight line. Therefore, it is required that we fit the linear regression line. Variance associated with Y is written as:

$$\text{Var}(Y) = \sigma^2$$

At times, X is a random variable and, in such cases, we consider the conditional mean of Y, given that X = x.

$$E(y|x) = \beta_0 + \beta_1 x$$

Conditional variance of Y, given that X = x.

$$Var\ (y|x) = \sigma^2$$

In case $\beta_0$, $\beta_1$ and $\sigma^2$ are known, the model is complete. However, the parameters $\beta_0$, $\beta_1$ and $\sigma^2$ are usually unknown and ε is not observed.

The value of the model $Y = \beta_0 + \beta_1 X + \varepsilon$ is dependent on the estimation of values of $\beta_0$, $\beta_1$ and $\sigma^2$. To determine the values of $\beta_0$, $\beta_1$ and $\sigma^2$, n pairs of observations of $(x_i, y_i)$ for i = 1, 2, ……, n are collected.

$$\beta_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$\beta_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

| i | x(Consumer Spending in ₹ 000) | y(GDP in ₹ 000) | xy (in ₹ 000) | $x^2$ (in ₹ 000) | $y^2$ (in ₹ 000) |
|---|---|---|---|---|---|
| 1 | 19 | 120 | 2280 | 361 | 14400 |
| 2 | 19.2 | 120.06 | 2305.152 | 368.64 | 14414.4 |
| 3 | 19.4 | 120.07 | 2329.358 | 376.36 | 14416.8 |
| 4 | 19.6 | 120 | 2353.764 | 384.16 | 14421.61 |
| 5 | 19.8 | 121 | 2395.8 | 392.04 | 14641 |
| Σ | 97 | 601.22 | 11664.07 | 1882.2 | 72293.82 |

From the above table $\Sigma x = 97$, $\Sigma y = 601.22$, $\Sigma xy = 11664.07$, $\Sigma x^2 = 1882.2$, $\Sigma y^2 = 1882.2$ and n = 5.

**Step 2:** Find $\beta_0$ and $\beta_1$

$$\beta_0 = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$= \frac{(601.22)(1882.2) - (97)(11664.07)}{5(1882.2) - (97)^2}$$

$$= \frac{1131616.284 - 1131414.79}{9411 - 9409}$$

$$= \frac{201.494}{2}$$

$$= 100.747$$

$$\beta_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$= \frac{5(11664.07) - (97)(601.22)}{5(1882.2) - (97)^2}$$

$$= \frac{58320.35 - 58318.34}{9411 - 9409}$$

$$= \frac{2.01}{2} = 1.005$$

**Step 3:** Insert these values in the equation,

$$Y = 100.747 + 1.005X$$

### 4.7.2 MULTIPLE REGRESSION ANALYSIS

In linear regression, you studied that an independent variable may affect the dependent variable differently. You studied the linear regression equation and the associated parameters. Here it is relevant to mention that, in practice, a dependent variable may be dependent upon more than one independent variable.

Multiple regression is depicted by the regression line:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots + \beta_k X_k$$

Note that the given multiple regression equation has k independent variables, where,

$\beta_0$ = Intercept of the line

$\beta_1$ = Slope for $X_1$; $\beta_1$ = First independent variable explaining variance in Y

$\beta_2$ = Slope for $X_2$; $\beta_2$ = Second independent variable explaining variance in Y

$\beta_k$ = Slope for $X_k$; $\beta_k$ = $k^{th}$ independent variable explaining variance in Y

The simplest form of multiple regression is created with two independent variables as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

To determine how well an equation fits the data is expressed by $R^2$. When $R^2$ value is 0, it means that there is no relationship between the Y and the X variables. When $R^2$ value is 1, it represents a perfect fit, which means that there is no difference between the observed and expected values of Y. It must be noted that the p-value is a function of $R^2$, the number of observations and the number of X variables.

## 4.7.3 BUILD REGRESSION MODEL IN EXCEL

You already studied about regression analysis and its types, such as simple (linear) regression and multiple regression. In linear regression, you studied that an independent variable may affect the dependent variable differently. You studied the linear regression equation and the associated parameters. Also, you already know that when we find the value of a dependent variable using the value of two or more independent variables, it is called multiple regression. You can perform regression analysis in excel. There are various ways to build a regression model in excel. Here, you will learn to build a linear regression model in excel with analysis ToolPak.

Let us take an example of sales number for smart phones for the last 12 months and find out the average monthly smart phones manufactured for the same period. The list of average monthly smart phones manufactured (independent variable) for the last 12 months entered in column B and the number of smart phones sold (dependent variable) is entered in column C. There are so many factors that can affect the sales of smart phones, but here we are focussing only on these two variables, as shown in Figure 20:

| | A | B | C |
|---|---|---|---|
| 1 | Month | Smartphones Manufactured | Smartphone Sold |
| 2 | January | 40 | 30 |
| 3 | February | 45 | 40 |
| 4 | March | 38 | 36 |
| 5 | April | 42.5 | 40.5 |
| 6 | May | 33 | 30 |
| 7 | June | 35.5 | 35 |
| 8 | July | 48 | 46 |
| 9 | August | 28 | 26 |
| 10 | September | 34 | 33 |
| 11 | October | 47 | 45 |
| 12 | November | 52 | 50 |
| 13 | December | 51 | 49.5 |

**Figure 20:** Regression Data

Perform the following steps to build the regression model in excel:

1. *Open* an MS Excel workbook.
2. *Click* the Data Analysis button in the Data tab, as shown in Figure 21:

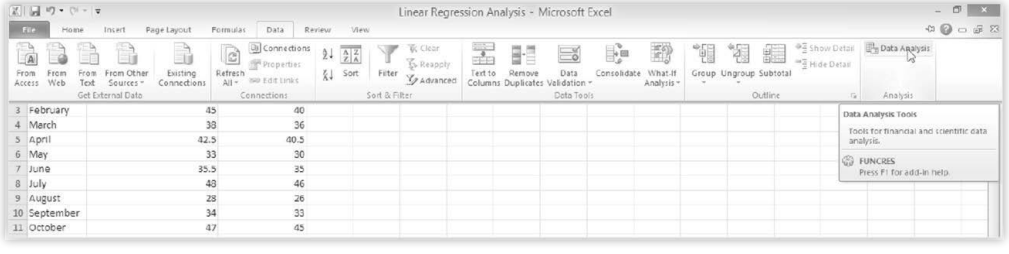



**Figure 21:** Selecting Data Analysis

The Data Analysis dialog box appears.

3. *Select* the Regression option from the Analysis Tools list box and click the OK button, as shown in Figure 22:

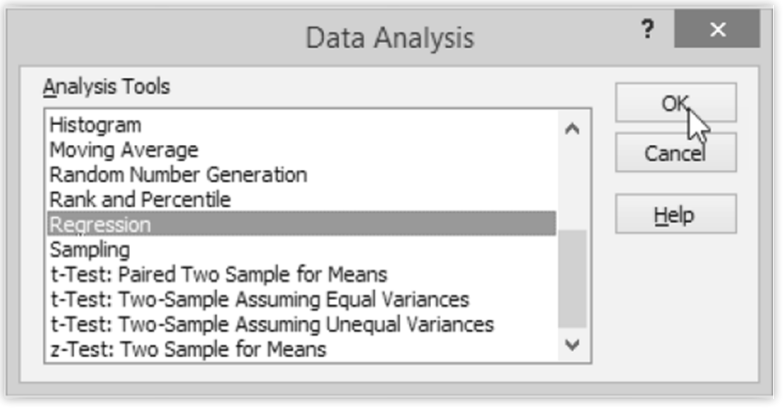



Figure 22: Selecting Regression from Data Analysis Dialog Box

The Regression dialog box appears, as shown in Figure 23:

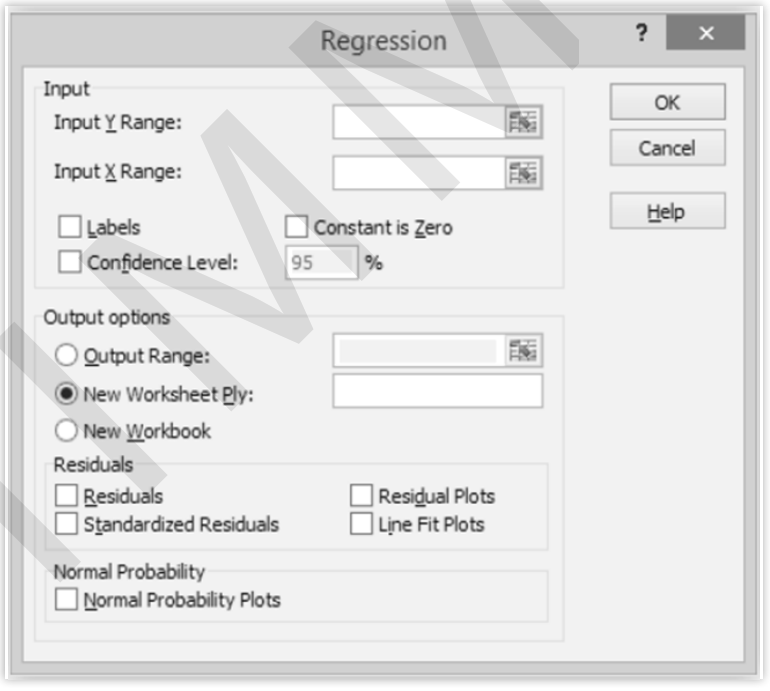



Figure 23: Displaying the Regression Dialog Box

4. *Configure* the following settings in the Regression dialog box:

   a. Select the Input Y Range. It is your dependent variable. In our case, it is smart phones' sales (C1:C13) (Figure 24).

   b. Select the Input X Range. It is your independent variable. In this example, it's the average monthly smart phones' manufacture (B1:B13) (Figure 24).

      If you want to build a multiple regression model in excel, then select two or more adjacent columns with different independent variables.

   c. Check the Labels box headers at the top of your X and Y ranges (Figure 24).

d. Choose your preferred Output option. In our case, we chose a new worksheet (Figure 24).

e. Select the Residuals check box to get the difference between the predicted and actual values as shown in Figure 24:

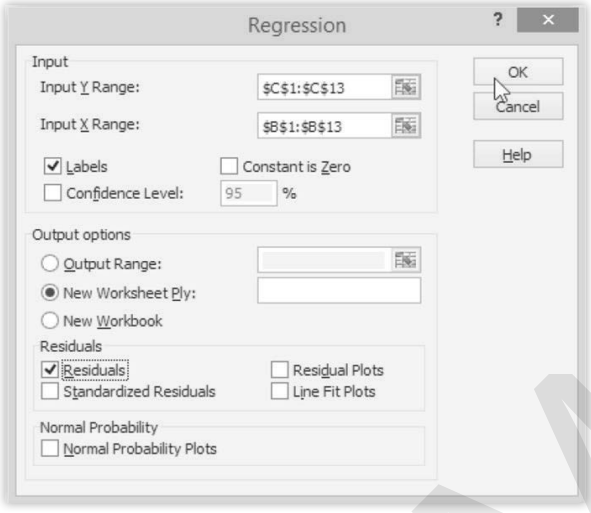



FIGURE 24: SELECTING DATA AND CONFIGURING FOR REGRESSION ANALYSIS OUTPUT

5. *Click* the OK button. It will generate the output of regression analysis based on the selected data. Now, you can observe the regression analysis output created by Excel, as shown in Figure 25:

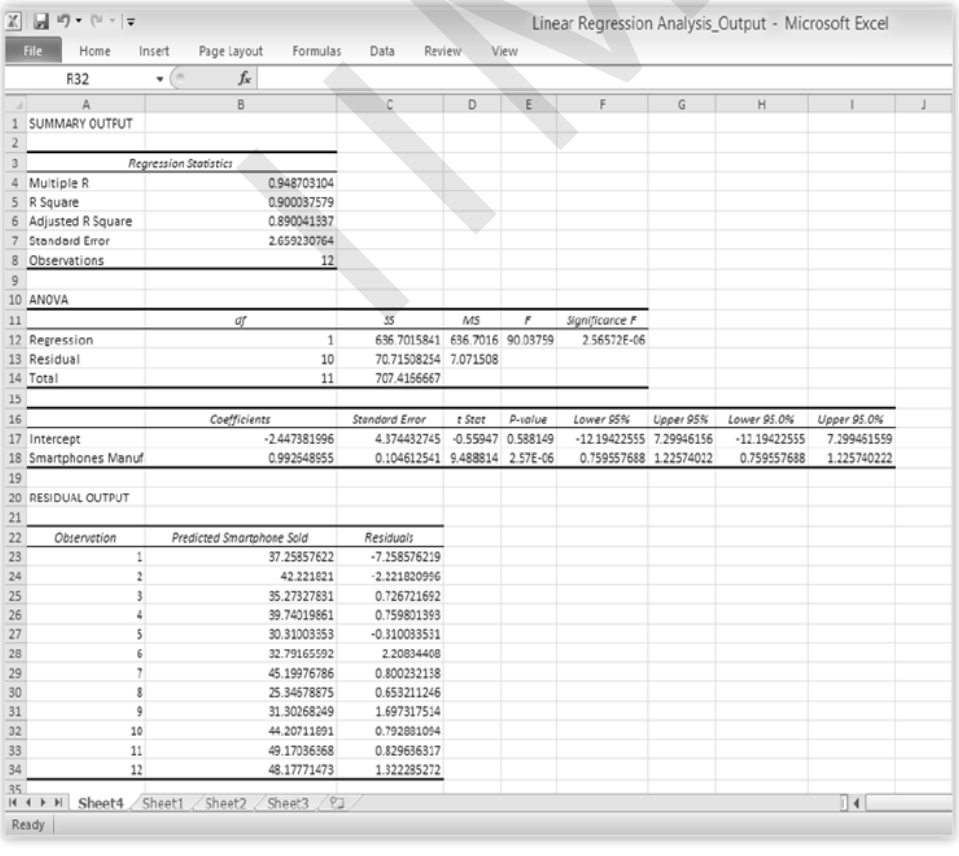



FIGURE 25: OUTPUT OF REGRESSION

Notes

As you have seen, building regression model in Excel is simple because Excel does all the calculations automatically. It generates 4 types of regression analysis output (Figure 24). These 4 types of regression analysis output are as follows:

- **Summary output:** This shows the fitness of your source data to calculate the linear regression equation. The information provided in the summary output is as follows:
  - **Multiple R:** It signifies the Correlation Coefficient that is used to measure the strength of a linear relationship between two variables.
  - **R-square:** It signifies the Coefficient of Determination, which indicates the goodness of fit.
  - **Adjusted R-square:** It denotes the R square that is adjusted for the number of independent variables existing in the model.
  - **Standard error:** It displays the accuracy of the regression analysis.
  - **Observations:** It represents the number of observations in your model.
- **Analysis of Variance (ANOVA):** It splits the sum of squares into individual components, and gives the information about the levels of variability within the regression model. ANOVA is not commonly used for a simple linear regression analysis in Excel. The components of ANOVA are as follows:
  - **df:** It denotes the number of the degrees of freedom related to the sources of variance.
  - **SS:** It denotes the sum of squares.
  - **MS:** It denotes the mean square.
  - **F:** It denotes F statistic, or F-test for the null hypothesis and is used for testing the overall importance of the model.
  - **Significance F:** It represents the P-value of F.
- **Coefficients:** It helps you build a linear regression equation in Excel. Consider the following equation:

  $$y = bx + a$$

  In this case, y represents the smart phones sold, x represents the average number of smart phones manufactured, and b represents the intercept. On putting values of a and b in the equation, it becomes

  $$y = 0.99x – 2.35$$

  With the average manufacturing of smart phones equal to 40, the sales would be

  $$y = 0.99*40 – 2.35$$

  $$y = 37.25$$
- **Residuals:** It helps you determine the variation between actual values and predicted values. In this case, you can see that the actual value 30 is different from the predicted value 37.25. So, the residual = 30 – 37.25 = – 7.25, which can be seen in the output obtained on the Excel sheet.

SELF ASSESSMENT QUESTIONS

19. Which one of the following is a statistical method that is used to model a relationship between two or more variables of interest?

    a. Business analysis

    b. Regression analysis

    c. Probability analysis

    d. Hypothesis analysis

20. When we find the value of a dependent variable using the value of two or more independent variables, it is called _____.

    a. Linear Regression

    b. Logistic Regression

    c. Multiple Regression

    d. Ordinal Regression

## 4.8 SUMMARY

- Data science is a multi-disciplinary subject that has developed as a combination of mathematical expertise (data inference and statistics) and algorithm development, business acumen and technology in order to solve complex problems.
- Data warehouse can be used to discover data and development of a data product that helps in generating value.
- Field of data science involves use of techniques, such as machine learning, statistical skills, cluster analysis, data mining, algorithms and coding and visualisation.
- Probability theory is a branch of mathematics that is concerned with chance or probability.
- Probability theory deals with concepts by expressing them in the form of axioms which formalise in terms of probability space.
- A Bernoulli distribution has only one trial and only two possible outcomes, namely 1 (success) and 0 (failure).
- In a Bernoulli distribution, a random variable X can take value 1 (success) with probability p or can take value 0 with probability q (= 1 – p).
- In a uniform distribution, there may be any number of outcomes and the probability of getting any outcome is equally likely.
- Normal distribution results in a bell-shaped symmetrical curve. This distribution occurs naturally in many situations.
- Statistical inference is the process of drawing conclusions or making predictions about a population based on a sample taken from that population.
- Statistical inference operates within distinct paradigms, such as classical, bayesian, likelihoods, and those based on the Akaike Information Criterion (AIC).

Notes

- Alternate hypothesis ($H_A$) is the hypothesis that must be accepted if the sample data leads to rejection of $H_0$.
- Hypothesis testing, also called significance testing, is a method which is used to test the hypothesis regarding the population parameters using the data collected from a sample.
- The Analysis of Variance (ANOVA) has an advantage over t-test when the researcher wants to compare the means of a larger number of population (i.e., three or more).
- Regression analysis is a statistical method that is used to model a relationship between two or more variables of interest.
- Regression analysis is usually used to model a relationship between a response variable (dependent variable) and one or more predictor (independent) variables.

## 4.9 KEY WORDS

- **Arithmetic Mean:** Arithmetic mean is the mean of a variable representing its average value.
- **Continuous Distribution:** A continuous distribution refers to the set of probabilities of the possible values of a continuous random variable.
- **Point Estimation:** Point estimation involves using sample data to estimate a single value of an unknown population parameter.
- **Interval Estimation:** Interval estimation provides a range of values within which the population parameter is likely to fall.
- **Frequentist Inference:** Frequentist inference assesses propositions by envisioning repeated samplings from a population distribution similar to the observed data.
- **Likelihood-Based Inference:** Likelihood-based inference revolves around estimating model parameters from observed data using the likelihood function.
- **Event:** If a subset of the sample space is taken, it is called an event.
- **Hypothesis Testing:** Hypothesis testing, also called significance testing, is a method which is used to test the hypothesis regarding the population parameters using the data collected from a sample.
- **Median:** It is known as the 'positional average' of a variable.
- **Mode:** The mode of a variable is the observation with the highest frequency or highest concentration of frequencies.
- **Probability Density Function (PDF):** The probability distribution can be described using an equation called Probability Density Function (PDF).
- **Probability Theory:** It deals with concepts by expressing them in the form of axioms which formalise in terms of probability space.
- **Regression Analysis:** It is a statistical method that is used to model a relationship between two or more variables of interest.
- **Sample Space:** The probability space assigns a value between 0 and 1 to a set of outcomes which are called sample space.

## 4.10 CASE STUDY: LINEAR REGRESSION FORECAST

In stock market analysis, Linear Regression technique is used to forecast future values (prices) of stocks/shares using the recent trend of prices. The forecasted values help in determining whether the stocks or the market (as a whole) would show an upward or downward trend.

Linear Regression technique is used to calculate the value of one (dependent) variable when the value of one or more independent variables is known. The linear regression results in a straight line which best fits the available data points (prices). In the study of financial and stock markets, technical analysis techniques are used. One of the technical analysis techniques is linear regression which results in a linear regression line that shows the market trend with respect to time. It helps in discovering when the markets deviate from the trend. The linear regression forecasts are calculated for a particular regression period (say, X). In linear regression, the regression line at each time period is drawn using the previous (X – 1) time periods. It can also be used to predict the trend for Y future time periods.

When linear regression is carried out for a stock market, any deviation from the usual trend may indicate temporary market disturbance or it may point towards a major trend reversal. However, it must be remembered that the linear regression technique can be used for stock market trends only when the markets are not highly volatile and the prices do not fluctuate a lot. The volatility of market decreases the fit of the linear regression equation to the data.

The trend line is derived by using the least squares fit method at each bar. The distance between the resulting linear regression line and the data points is minimised to plot a straight line which reveals a trend. Forecasting market trends is a highly tedious job because it involves calculating the regression lines at each bar (time period). In addition, the prices used for calculating linear regression may require smoothing using moving average. The value of each bar is taken as the end point of the line. The linear regression equation is represented as:

Regression = Reg (Price, X) + Slope * Y

Here, X = Regression period; Y = Forecast period; and Slope = Slope of the regression line

The slope of the linear regression line represents the strength of the trend. The linear regression line is overlaid by drawing the upper and lower bands by calculating the number of standard deviations above or below the regression line.

The upper and lower bands are calculated as follows:

- Upper Regression Band = Regression + SD (Price, X) * N
- Lower Regression Band = Regression – SD (Price, X) * N

Here, N = Number of standard deviations and X = Regression period

When the data points are smoothened and the linear regression forecasting is done using moving averages, the linear regression line does not show a lot of delay because

Notes

moving average fits a line to data points rather than averaging them. Price data may be smoothened using moving average for a specified number of time periods. Similarly, forecasting can be done for a specified number of time periods.

Usually, the Linear Regression Forecast is presented as a channel with regression value in the upper, middle and lower bands. The upper and lower bands are drawn using a particular number of standard deviations from the regression value. Stock market analysts use linear regression (technical analysis) technique to identify a short-term trend as well as for identifying trend reversals. Linear regression can be carried out for identifying intra-day, daily or weekly trends. However, the linear regression technique is not suitable for making forecasts for long periods.

In Linear Regression, crossover or trend identification is used to show the overall trend line. The upper and lower bands may also be used to represent the support and resistance levels. When one standard deviation is used, the upper and lower channel lines contain 68% of all the historical prices. Similarly, when two standard deviations are used, the upper and lower channel lines contain 95% of all the historical prices.

In case the prices break outside any of the channels, there are two possibilities:

- Presence of buy/sell opportunity (buy below the lower trend and sell above the upper trend)
- Ending of a major trend

Any movement in the stock prices beyond two standard deviations indicates that the concerned stock may be overbought or oversold with a 5% probability. In such a case, the security may bounce back. The traders may look for instances when the prices close back inside the linear regression channel and they may take buy/sell positions accordingly. When the prices remain outside the Linear Regression Channel for longer periods of time, it indicates a trend reversal or change.

**Source:** *http://www.blastchart.com/Community/IndicatorGuide/Overlays/LinearRegressionForecast.aspx*

## QUESTIONS

1. What is the significance of standard deviation overlay?

   (**Hint:** Traders use overlay as an indication to identify the overbought and oversold positions.)

2. After thoroughly studying the above case, identify one drawback of the Linear Regression technique.

   (**Hint:** Linear Regression technique should not be used for longer periods of time because the fit of the trend to the data would not be reliable.)

3. Which technique is used by stock market analysts to identify a short-term trend as well as for identifying trend reversals?

   (**Hint:** Stock market analysts use linear regression (technical analysis) technique to identify a short-term trend as well as for identifying trend reversals.)

4. Which technique is used to identify intra-day, daily or weekly trends?

   (**Hint:** Linear regression can be carried out for identifying intra-day, daily or weekly trends.)

5. What does slope of the linear regression line represent?

   (**Hint:** The slope of the linear regression line represents the strength of the trend.)

## 4.11 EXERCISE

1. 1,000 students of class 9 of ABC School score a mean IQ of 100 in an IQ test with a standard deviation of 15. Assume that you are a researcher and want to determine what percent of the students would score between mean – 1 SD and mean + 1 SD.
2. Enlist and describe at least four different types of distributions. Also specify whether each of the distributions is a continuous or a discrete distribution.
3. Explain in detail the types of error and t-test.
4. Write a short note on the process of determining sample size for population.
5. Describe simple and multiple regression in detail.

## 4.12 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Measures of Central Tendency | 1. | mean |
| | 2. | Median |
| | 3. | mode |
| | 4. | Mean |
| Probability Theory | 5. | sample space |
| | 6. | d. Continuous distribution |
| | 7. | $p^x(1–p)^{1–x}$ |
| Statistical Inference | 8. | b. Bayesian Inference |
| | 9. | c. A measure of uncertainty associated with the estimate |
| | 10. | False |
| | 11. | complexity |
| Hypothesis Testing | 12. | True |
| | 13. | False |
| | 14. | False |
| | 15. | True |
| | 16. | True |
| Correlation | 17. | r |
| | 18. | True |
| Regression Analysis | 19. | b. Regression analysis |
| | 20. | c. Multiple Regression |

Notes

## 4.13 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Linoff, G. (N.D.). *Data Analysis using SQL and Excel®.*
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Hoboken, NJ:Wiley-Interscience.


### E-REFERENCES


- ScienceDaily. (2018). Probability theory. [online] Available at: https://www.sciencedaily.com/terms/probability_theory.htm [Accessed 30 Nov. 2018].
- Faculty.math.illinois.edu. (2018). Basic Probability Theory. [online] Available at: https://faculty.math.illinois.edu/~r-ash/BPT.html [Accessed 30 Nov. 2018].
- Statistics Solutions. (2018). Hypothesis Testing - Statistics Solutions. [online] Available at: https://www.statisticssolutions.com/hypothesis-testing/[Accessed 30 Nov. 2018].
- Stattrek.com. (2018). Hypothesis Tests. [online] Available at: https://stattrek.com/hypothesis-test/hypothesis-testing.aspx [Accessed 30 Nov. 2018].

CHAPTER 5

# Decision Making and Support

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the concepts of decision making
- Explain the need of decision making system
- Elucidate the techniques of decision making
- Define the concept of data science for decision making in key areas
- Discuss data mining with decision tree

## 5.1 INTRODUCTION

In the previous chapter, you have studied the various techniques of statistics that are used frequently in data science. You have also studied the important concepts such as probability theory, statistical inference, sampling theory, hypothesis testing and regression analysis for analysing data and making decisions.

Decisions are inevitable and help an organisation to pave its way towards better performance. These decisions could be related to strategies, business activities or HR and each one of them is made in the best interest of the organisation. The process of decision making is an integral part of the managerial process in any organisation. However, this process is highly complex and involves experts from different domains. In case of large organisations, there is often a team of experts specially trained to make all sorts of decisions, while usually in small organisations, all significant decisions are taken by the managerial board.

The decision-making process is collective and consultative, and has a considerable impact on the overall growth and prospects of the organisation. However, there are some advantages and disadvantages in the process that reflect the consequences on the overall performance of the organisation.

As you know, decisions are taken to support organisational growth. Therefore, it requires the manager to be able to take critical decisions at any level – top-, middle-, or entry-level. The foundation of management in an organisation is built on managerial decisions, which are reflected through its day-to-day operations.

Many big corporations use effective communication tools in addition to the normal consultation process to make decisions that would have large-scale implications.

Discussions and consultations along with standard procedures and techniques are the two main tools that maintain and eventually facilitate decision making.

For example, when the strategic management team suggests a decision on initiating a new business activity, it must follow a series of deliberate discussions and consultations. Decisions that are taken by strategic managers often reflect new and innovative business initiatives. Thus, an extensive debate and research is required before finalising a decision.

Moreover, the final decision to roll out a product or service is accomplished through collective interim decisions taken by various internal and external units.

This decision proves to be reflective with the research done and consultations within various levels in the organisation. The overall process is a sequence of steps where one decision is taken at one point, and where each level has far-forcing implications on the overall decision-making strategy of the organisation.

In this chapter, you will first learn about the concept of decision making. Next, you will learn about the techniques of decision making. Then, the chapter discusses the development of Decision Support System (DSS). Also, the chapter explains the application of DSS.

## 5.2 CONCEPT OF DECISION MAKING

Decision making is concerned with the future and involves the act of selecting one the best course of action from the various courses of action. It is one of the major functions of management, which is difficult but very important. The most important responsibility of management in any organisation is to set up organisational goals and allocate the available resources effectively and efficiently. Resources are always limited and need to be used judiciously to achieve maximum profits.

Accounting information can improve the understanding of the management with respect to the alternative resource allocation. This information is provided by the cost management information system in terms of supply cost and revenue data that are useful to make strategic decisions.

### 5.2.1 TYPES OF DECISIONS

As you have studied, the decision-making process refers to selecting the best choice from the available options. It requires a hard thinking and intellectual weighing of different options to arrive at a certain choice depending on the requirement of the situation. Decisions are needed when different options are available, when a problem occurs and a solution is required, and when an opportunity comes along and there is a need to make a choice.

As each individual has different perspectives and different intellectual skills,the ability to make a decision also varies. Decision makers are categorised into various types. Similarly, the approach of decision making is also characterised into different kinds. The type of decision generally depends on an individual as well as on the situation at hand. The most common types of decisions that an organisation usually makes are given as follows:

- **Programmed decisions:** These are standard decisions that typically follow a repetitive practice. For example, most organisations have standard procedures to address customer complaints. Similarly, an inventory manager can order the product that is dipping beyond the cut-off mark to maintain stock. The procedure to take programmed decisions is fixed. Therefore, it can be written down in a sequence of steps which everyone can follow as astandard. They could also be written in the form of a computer program.
- **Non-programmed decisions:** These are non-standard and non-routine decisions, where every decision is different from the previous one. There is no need to set guidelines or rules for such decisions as each situation is either uncertain or unplanned, for example, a decision on whether the firm should go for a merger/ acquisition or not. Similarly, selecting a college for further studies is a non-programmed decision. These decisions are taken when the situation is unique

and information is unstructured. A decision maker needs to collect information, establish a link between the pieces of information available, consider the different alternatives and delve into a continuous thinking process.

- **Strategic decisions:** These are the long-term decisions that could set the trend of business, for example, a decision on whether the company should launch a new service or product, stop offering a product, or acquire a company for some specific service or product. Another example could be to train employees for enhancing performance and sustaining over a long term.
- **Tactical decisions:** These are medium-term decisions that are taken in regard to implementing strategic decisions, for example, market analysis for a new product or the staff required. The grocery stores sell data received through the bar code scanner to organisations such as Information Resources, Inc.(IRI) which are responsible for collation and further selling of it to grocery vendors and wholesalers. These people can study selling pattern of their competitors and respond according to the prevailing situation. At a tactical level, the forecast defined here is applied as a policy. At Continental Airlines, the tactical decisions are taken based on the query generated by the staff that accesses the Flight Management Dashboard application to check a particular flight status at a specific time.
- **Operational decisions:** These are short-term decisions that guide us about how to perform the regular operations, for example, the decision to hire a particular logistic company to make deliveries. Another example from the financial sector can be EMC Insurance Companies, which find it difficult to determine the amount of money required to be held in reserve against any potential case payouts. As a solution, EMC opted for PolyVista, a data analytics software, to reveal the hidden patterns, relationships, and anomalies within the firm's warehouse of claim data.

## 5.2.2 DECISION-MAKING PROCESS

This process is lengthy as well as time-consuming, but is carried out in a scientific manner. The decision-making process suggests a number of general guidelines and methods that need to be followed regarding how you should take a decision. It involves many steps that are arranged logically. **Peter Drucker** published the book 'The Practice of Management' in 1955. In this book, he suggests the scientific method of decision making. This method is considered as a base model and organisations can make changes depending on the nature of business. This basic model, given by Peter Drucker, consists of seven steps as shown in Figure 1:


Figure 1: The Decision-Making Process

Notes

Let us elaborate each of the steps as mentioned in Figure 1.

1. **Identify the problem:** This is the first step in a decision-making process. A well-defined problem can be solved easily. A manager should also identify the underlying 'critical factors' related to the problem. In addition, the manager requires considering the reasons and check if he/she can control them or not. For example, suppose an organisation is suffering from losses. After investigation, it finds out that its inferior product quality is the reason behind the loss incurred. This investigation is an example of identifying the problem. Another example is of the world-leading retailer, Tesco, which defined the strategic priorities in Balance Scorecard called the corporate steering wheel. Any action could be mapped to the wheel. If it does not hit the outlined strategic objective, it is identified as a problem.
2. **Analyse the problem:** The next step that decision-making process involves is the deep analysis of the problem. In order to know who is responsible for decision making and who must be reported to for the decision, it is important to classify the problem. The following four factors are generally evaluated:

   I. Future prospects of the decision

   II. Impact of the decision on other aspects of the organisation

   III. Number of qualitative considerations involved

   IV. Uniqueness of the decision

   For example, an organisation can analyse the quality issues in detail to collect the relevant information.
3. **Collect the relevant data:** The next step is to obtain facts and the relevant data. The flow of information due to strong communication media and modern development in the field of IT has enhanced considerably. Therefore, in such a scenario, identifying and picking out the appropriate and useful information is a big task. The filtered information should be used carefully for problem analysis. For example, after a detailed analysis of the problems pertaining to quality, an organisation can collect the relevant data such as the frequency of the occurrence of the problem. Tesco collects real-time data using its loyalty card – Clubcard, to observe consumer trends.
4. **Develop the alternative solutions:** After the problem has been identified, analysed and the relevant data is collected, the manager needs to determine the available alternative choices and the corresponding courses of action that are available to solve the problem at hand. Time and cost limitations along with psychological obstacles should also be considered to confine the number of alternatives. For example, various alternative solutions for quality problems in an organisation can be the implementation of new techniques, installing sophisticated machinery and installing a quality monitoring system. Tesco uses the real-time data to gain a detailed insight unlike any of its competitors. On the basis of the buying patterns of the consumers, Tesco can introduce many new schemes to increase sales and customer loyalty.
5. **Select the best solution:** In the decision-making process, when the alternative solutions have been developed, the subsequent step is that of selecting the best alternative to obtain the best result. The selected alternative should be conveyed

to those who are liable to be influenced by it. This decision can be implemented effectively if it is accepted by all group members.

6. **Convert decision into action:** Once the best decision has been selected, the next step is to translate the selected decision into an effectual action. In the absence of such appropriate action, the decision is just considered to be an assertion of good intentions. The role of the manager is to convert 'his/her decision' into 'their decision' with the help of his/her leadership skills. The manager should take his/her subordinates into confidence and convince them about the appropriateness of the decision. After this, the manager should follow up in order to ensure proper execution of the decision.
7. **Ensure feedback:** This is the last step of the decision-making process. The manager needs to formulate provisions to ensure that the feedback is taken continuously and that the actual developments are in tune with the expectations. It is the process of checking the effectiveness of the decision taken. Feedback helps in establishing whether the decision that has been taken needs to be continued or modified taking the changed conditions into account.

There are many companies that enjoy the advantage of business intelligence in effective decision making. For example, the fast food chain, McDonald's uses Business Intelligence (BI) to make strategic decisions such as what to add to the menu or which under-performing stores need to be closed or what new schemes should be implemented to drag them out from a low-profit zone.

Yahoo Inc. also uses BI to bring changes to its websites. Millions of users hit the organisation home page each hour. To bring changes to the home page, the organisation randomly selects a few thousand users as an experimental group and checks their behaviour. It can obtain the result of analysis in just a few minutes. This fast access to the results helps the organisation in optimising the offering to increase the number of hits and profits. At any given time, Yahoo generally runs about 20 such experiments.

## Self Assessment Questions

1. Decision at the operational level tends to be more:
   a. Programmed decision
   b. Tactical decision
   c. Semi-structured decision
   d. Unstructured decision
2. Rearrange the list of decision-making process in right sequence:
   I. Select the best solution
   II. Analyse the problem
   III. Develop alternative solutions
   IV. Convert decisions into action
   V. Follow up the action
   VI. Identify the problem
   VII. Collect relevant data

   Select the correct option:
   a. VI, II, VII, III, I, IV, V
   b. V, II, VII, III, I, IV, VI
   c. VI, II, VII, III, I, V, IV
   d. VI, II, VII, I, III, IV, V

**Activity**

Search the importance and uses of Group Decision Support System (GDSS) in organisations.

## 5.3 UNDERSTANDING DECISION SUPPORT SYSTEM

The definition of DSS has been evolving since 1970s and the definition in its present form has been described by two people, named **Ralph Sprague and Eric Carlson.** Their definition of DSS poses it in the form of a system that is based on computers and helps decision-making authorities to confront the ill-structured problems by interaction directly with data and analysis models.

DSS is used to analyse business data and present an interactive information support to all decision makers. Its role starts right from the stage of problem identification and continues till the decision is implemented. DSS uses analytical models, dedicated databases, insights and the judgement of the decision maker and an interactive, computer-based modelling approach to sustain unstructured decisions.

In an age of ever-evolving competition in the global business environment, you do not enjoy the liberty of being able to spend too much time on taking decisions. You are always expected to make more and faster decisions than ever before. This is the reason why businesses need leaders who can take quick decisions. The entire decision-making process has become extremely accelerated. In such a situation, it is impracticable to depend solely on human response. Therefore, companies need a DSS to react and adapt to the persistently changing business environment.

Therefore, for being successful in the business environment created today, your company requires information systems by which diverse information and decision making needs can be supported. In addition, the system also needs to support you in taking prompt decisions. DSSs assist in assessing and resolving the questions posed everyday in businesses. To do so, DSSs analyse raw data, documents, personal knowledge and business models from where useful information is gathered.

To attain maximum performance in the existing business environment, you need to achieve a competitive advantage. Otherwise, proper functioning of your company is in doubt, it may ultimately come to a close.

Your decisions can be considered better when on implementation, they are able to reduce the costs effectively, use assets more economically, enhance revenue, cut risks down and provide improvement in customer service.

### 5.3.1 MODEL-DRIVEN DSS

The model-driven DSS helps decision makers in analysing decisions and making choices from different alternatives. It manipulates data to generate statistical and financial reports as well as simulation models. The model-driven DSS follows 'what-if' analysis as an analytical tool. This type of DSS is helpful in analysing the effect of a change in certain variables towards the efficiency of the business. It can be used on a standalone PC, client/server, or the Web. The data and parameters provided by decision makers are used by model-driven DSS to analyse a situation and make

decisions. In general, model-driven DSS uses complex financial, optimisation, simulation, or multi-criteria models for supporting decision-making process.

### 5.3.2 DATA-DRIVEN DSS

The data-driven DSS focuses mainly on internal as well as external data of an organisation which is obtained from the data warehouse for decision making. Managers and other executives mainly depend on data-driven DSS as they need to consider the database to make various types of decisions. This type of DSS can be implemented by using a mainframe or client-server technology. This system utilises online analytical processing tools for data analysis. The examples of data-driven DSS are Geographic Information System (GIS), which represents geographical data through maps and Executive Information System (EIS) used by senior executives for decision making.

### 5.3.3 DSS USER INTERFACE

A user interface allows a user to interact with a system. It is what a user can see and use. A user interface generally includes different elements such as buttons, menus, icons, etc. In the case of DSS, a user interface decides how information will be entered and displayed in a system. In other words, it gives a visual way to develop communication between the user and DSS. The quality of a DSS depends on its user interface for good decision making. The following are some styles used for creating a user interface:

- **Command-line interface:** In this interface, a user has to enter commands in order to instruct a machine to do what they want to do.
- **Menu-driven interface:** It is difficult to learn commands, hence the solution is the menu interface in which a user can get a list of options. The menu makes it easier for a user to choose the required option.
- **Graphical user interface:** Visible objects are available in this interface to give instructions to a system. GUI is more focused on graphical objects rather than textual commands.
- **Voice user interface:** In this interface, a user can interact with a system through voice or speech. User's voice is required to instruct and control the system. These interfaces become common nowadays.
- **Touch user interface:** In this interface, a user can interact with a system by touching the interface of the system. Moreover, there is no need to learn textual commands for interaction.

**Self Assessment Questions**

3. Which types of data Decision Support System (DSS) analyse?
   a. Raw data
   b. Documents
   c. Personal knowledge and business models
   d. All of the above

Notes

4. DSS is used to analyse business data and present an interactive information support to all decision makers. (True/False)

**Activity**

Enlist and explain the components of a Decision Support System (DSS).

## 5.4 TECHNIQUES OF DECISION MAKING

A decision, when taken scientifically, leaves a little scope for confusion and generally meets the required goals. The accuracy and rationale for such decisions can be justified even if they do not accomplish the required purpose. Though there is no standard to ensure that the decision has been taken correctly, yet there are some important techniques that can assist a manager. Some of the widely used techniques are given as follows:

- **Operations research:** This is the study in which scientific methods are applied to the problems that are complicated and cropping up in the direction and administration of a sizeable system that comprises men, machines, raw materials and capital in various areas such as business, industry, defense and government. As **Robert J. Thierauf** has stated, *Operations research utilises the planned approach and an interdisciplinary team in order to represent functional relationships as mathematical models for the purpose of providing a quantitative basis for decision making and uncovering new problems for quantitative analysis.*

  Operations research facilitates the authority that is making the decision to formulate decisions objectively through a fact-based direction and support to the decision, and easing the responsibility of the manager with respect to efforts and time taken. Certain managerial problems that are generally subjected to operations research analysis comprise inventory control, production scheduling, expansion of plant, sales policies, etc. DANOPT is a company that uses operations research for business analytics to outperform in the market.

- **Models:** These are based on mathematical relationships and are also known as mathematical models. These models facilitate the concept of optimisation while making decisions. Models play a very important role in the calculation and selection of the best possible alternative solution for a particular problem. For example, linear programming is a mathematical model which helps in solving optimisation problems. Also, Rapid Decision is a BI solution that uses data models and its strength lies in the involvement of a wide variety of metadata fields to enhance usefulness and understanding.

- **Simulation:** The simulation techniques help in testing various alternatives in order to verify whether they are feasible or not and to check their possible outcomes. This quantitative technique is used for assessing various courses of action. The technique is developed on the basis of facts and assumptions and involves mathematical models that are computerised to show the actual decision making when influenced by uncertain conditions, identify bottlenecks, test process designs and test changes. For example, Denim manufacturing Turkey-based textile company, ISKO invested

in WITNESS Predictive Simulation Technology, from Lanner, to optimise uptime of weaving machines.

- **Game theory:** It helps find out an optimum solution for developing an effective strategy for a given situation, irrespective of the condition to maximise profits or minimise losses in a competitive age. It includes mathematical study of an appropriate approach during uncertainty. For example, big players such as Microsoft, Chevron and BAE Systems have appreciated the significance of game theory in defining high-risk and complicated decisions.
- **Decision trees:** This technique uses a graphical representation to visualise all the possible outcomes based on the decisions. This algorithm uses Tree representation which comprises a root and children nodes along with Leafnode. The root classifies the main decision or condition, followed by alternate solutions which are its branches (child nodes). This structure helps identify all the alternatives and accordingly the decision-making process becomes easier and effective. Decision trees help in classifying different paths related to a problem, making it faster to make the best decision. Decision trees can be used when the analyst wants to make sure all paths related to a condition are well-checked and analysed based on their reward depending on the problem.
- **Analytical hierarchy process (AHP):** It introduced in 1980 by Thomas Saaty and is a tool which is used to deal with complex decision making in an effective and efficient way. It helps a decision maker to take better decisions by setting desired goals and priorities. AHP, due to its specific design, checks the consistency in the evaluations of the decision maker, avoiding any possibility of biasing in the process of decision making. AHP reduces the complexity of decisions by reducing complex decisions into sets of pairwise comparisons and then synthesises the result. AHP contains a set of evaluation constraints along with a set of alternatives from which the decision is to be made. This does not guarantee that the optimal option for every criterion will be chosen. Rather, the option which strikes the balance of optimisation among all criteria is decided to be most suitable. A weight is assigned to each evaluation criteria, which is influenced by the decision maker's pairwise criteria comparison. This weight is directly proportional to the importance of the corresponding criterion. In the next step, AHP assigns a score to each option based on the option's pairwise comparison done by the decision maker, based on that criterion. This process results in the accumulation of criteria weight and corresponding option score which eventually determines the global score for each option, followed by a ranking. The magnitude of the score reflects the performance of the option for the given criteria and overall ranking. The global score for an option is equal to the weighted sum of all the scores obtained by it in accordance with all criteria.
- **Influence diagrams:** An influence diagram is used in the decision-making process to graphically represent a decision-making problem. A decision tree is also used for the same reason, but it is very complex. You can easily understand the relationships and dependencies between variables through an influence diagram, which is more compact than a decision tree. The decision, chance variable, objective, and general variable are the four nodes that are the building blocks of an influence diagram. There is one more important thing in the influence diagram that is an arrow and

Notes

the head of an arrow indicates the flow of influence. Each node in the influence diagram is connected with another node through the arrow.

**Self Assessment Questions**

5. _____ play a very important role in calculation and selection of the best possible alternative solution for a particular problem.
6. The stimulation technique is developed on the basis of _____ and assumptions, and involves _____ models that are computerized to show the actual decision making.

## 5.5 DATA MINING WITH DECISION TREES

Decision trees represent a potent and widely embraced method for data mining and machine learning, functioning as a supervised learning model that learns from labeled data. Labeled data encompasses both input features and a target variable, which the model endeavors to predict.

The process of constructing decision trees involves recursively partitioning the data into smaller subsets based on feature values. This partitioning persists until a predetermined stopping criterion is met, such as when all data in a subset belongs to the same class or when the subset's data points fall below a specified threshold.

An illustration of a decision tree predicting car purchases based on age, income, and credit score exemplifies this process. The tree poses the initial query, "Age < 30?" at the root node. Depending on the response, it branches into two paths—one for those with incomes below ₹50,000 and another for those earning ₹50,000 or more. This branching continues until a leaf node is reached, signifying a prediction regarding car purchase likelihood.

**Advantages of Decision Tree**

- **Interpretability:** They are easy to interpret, as the decision rules are transparent and straightforward.
- **Categorical data handling:** They can process categorical data without requiring preprocessing.
- **Robustness:** They are robust to outliers and missing data.
- **Versatility:** They can be applied to classification and regression tasks. However, certain drawbacks exist.
- **Overfitting:** Decision trees may overfit to training data, potentially impacting performance on unseen data.
- **Sensitivity to noise:** They can be sensitive to data noise.
- **Computational cost:** Training can be computationally expensive, particularly for sizable datasets. Despite these drawbacks, decision trees find application in diverse data mining tasks, including fraud detection, customer segmentation, medical diagnosis, credit risk assessment, image recognition, and text classification.

### 5.5.1 FOUR-LAYER MODEL

A four-layer model in data mining with decision trees is a specific architecture for building decision trees that utilizes four distinct layers:

- **Presentation layer:** The presentation layer in business analytics refers to the interface or visualization component of a system where data insights and analytical results are presented to end-users. It involves the design and creation of dashboards, reports, and other graphical representations that make complex data understandable and actionable for decision-makers.
- **Query layer:** The query layer in business analytics is responsible for retrieving and processing data from various sources to answer specific questions or fulfill analytical requests. This layer facilitates the extraction of relevant information from databases, data warehouses, or other data repositories. Users interact with the query layer to formulate queries, request specific data sets, or perform analyses.
- **Classification layer:** The classification layer in business analytics refers to a component that is focused on categorizing or grouping data based on certain attributes or characteristics. Classification algorithms are commonly used in machine learning and predictive analytics to assign predefined labels or categories to data points.
- **Storage layer:** The storage layer in business analytics is where data is stored and managed. It encompasses databases, data warehouses, data lakes, and other storage systems that house the raw and processed data used in analytical processes. The storage layer is fundamental for maintaining data integrity, ensuring data accessibility, and supporting efficient data retrieval.

**Benefits of Four-Layer Model**

- **Improved user interaction:** The GUI and query layer enable users to directly interact with the data and the decision tree model, enhancing their understanding and facilitating exploration.
- **Enhanced model explainability:** The decision tree›s structure and splitting criteria provide clear explanations of the model›s predictions, making it easier to interpret and debug.
- **Flexible data management:** The storage layer allows for efficient data handling and supports various data formats and sizes.
- **Scalability:** The modular architecture enables easy scaling of individual layers to accommodate larger datasets and more complex models.

**Applications of Four-Layer Model**

- **Biological data analysis:** Identifying patterns and relationships in biological data for research and drug discovery.
- **Customer segmentation:** Clustering customers based on their characteristics and behavior for targeted marketing campaigns.
- **Fraud detection:** Identifying fraudulent transactions in financial data.

- **Risk assessment:** Assessing the risk associated with loans, investments, or insurance policies.
- **Medical diagnosis:** Assisting medical professionals in diagnosis and treatment decisions.

**Here are some examples of how the four-layer model can be applied:**

- **Predicting protein function:** Researchers can use the four-layer model to analyze biological data about proteins, such as their sequence, structure, and interactions. The decision tree model can then be used to predict the function of unknown proteins.
- **Identifying fraudulent transactions:** Financial institutions can use the four-layer model to analyze financial data and identify fraudulent transactions. The decision tree model can be used to flag transactions that are likely to be fraudulent, allowing the bank to investigate them further.
- **Predicting customer churn:** Companies can use the four-layer model to analyze customer data and predict which customers are likely to churn. The decision tree model can be used to identify at-risk customers and take steps to prevent them from leaving.

### 5.5.2 CLASSIFICATION TREES

Classification trees are a fundamental component of data mining with decision trees. In the context of machine learning, particularly supervised learning, classification trees are designed to predict the class or category of a target variable based on input features. Here's an overview of how classification trees work in the context of data mining. The primary goal of a classification tree is to predict the class or category of a target variable. This is achieved by recursively partitioning the data based on input features.

**Tree Construction**

- **Splitting criteria:** The tree-building process involves selecting optimal splitting criteria for each node. The goal is to create splits that maximize the homogeneity of data within each resulting subset.
- **Gini index or entropy:** Common criteria for measuring homogeneity include the Gini index or entropy. These metrics quantify the impurity or disorder within a set of data, and the tree aims to minimize them.

**Nodes and Leaves**

- **Root node:** The initial node at the top of the tree represents the entire dataset.
- **Internal nodes:** Nodes generated through the splitting process, posing questions based on feature values.
- **Leaf nodes:** Terminal nodes where predictions are made. Each leaf node corresponds to a specific class or category.

**Splitting Process**

- **Recursive splitting:** The data is recursively split into subsets based on the values of input features. This process continues until stopping criteria are met, such as a predefined tree depth or a minimum number of data points in a node.

**Decision Rules**

- **Interpretability:** One of the key advantages of classification trees is their interpretability. Decision rules at each node are clear and easy to understand, making the model accessible to non-experts.

**Handling Categorical Data**

- **Categorical variables:** Classification trees naturally handle categorical variables without the need for extensive preprocessing.

**Predictions**

- **Majority voting:** When a new data point traverses the tree, it follows the decision rules at each node until it reaches a leaf. The majority class in the leaf node becomes the predicted class for the data point.

### 5.5.3 CHARACTERISTICS OF TREES

Decision trees, a widely employed machine learning algorithm, excel in performing both classification and regression tasks. The following are key characteristics that define decision trees:

- **Hierarchy and structure:** Decision trees exhibit a hierarchical structure, with nodes symbolizing decisions or test conditions, and branches extending based on the outcomes of these decisions.
- **Root node:** The topmost node in a decision tree is termed the root node, signifying the initial decision or feature used to split the data.
- **Internal nodes:** Internal nodes denote decision points where data is partitioned into subsets based on specific conditions or features.
- **Leaves (Terminal Nodes):** Leaves, positioned at the terminus of the tree, represent the ultimate predicted outcomes or values. Each leaf corresponds to a distinct class (for classification) or a predicted value (for regression).
- **Decision rules:** Decision trees rely on decision rules, formulated based on feature values in the dataset, to navigate through the tree.
- **Splitting criteria:** Splitting criteria, such as Gini impurity (for classification) or mean squared error (for regression), guide decision trees in determining how to divide the data at each node.
- **Feature importance:** Decision trees offer a metric for feature importance. Features closer to the root node, contributing to more substantial splits, are generally considered more influential in making predictions.
- **Prone to overfitting:** Decision trees are susceptible to overfitting, particularly when deep and capturing noise in the training data. Pruning techniques can be employed to alleviate this concern.

Notes

- **Non-parametric model:** Classified as non-parametric models, decision trees refrain from making assumptions about the underlying data distribution, allowing them to capture intricate relationships without predefined assumptions.
- **Interpretability:** Decision trees boast high interpretability, facilitating an understanding of the rationale behind the model's predictions. The path from the root to a leaf node represents a set of rules guiding predictions.
- **Handling categorical and numerical data:** Decision trees adeptly handle both categorical and numerical features, employing distinct strategies for data splitting based on feature type.
- **Ensemble methods:** Decision trees often serve as foundational components for ensemble methods like Random Forests and Gradient Boosting. These methods amalgamate multiple trees to enhance overall predictive performance.

### 5.5.4 TREE SIZE

The size of a decision tree, often referred to as its "tree size," plays a crucial role in balancing model complexity and generalization performance. It is determined by the number of nodes, including internal nodes that represent decision points based on features, and leaf nodes that represent the final outcomes or predictions. The size of a decision tree is crucial as it directly impacts the model's ability to generalize to new, unseen data.

A larger tree with more nodes has the potential to capture intricate patterns in the training data. However, there's a risk of overfitting, where the model learns noise or specific details of the training data that don't generalize well to new data. On the other hand, a smaller tree may oversimplify the underlying patterns and fail to capture important relationships in the data.

Finding the right balance in tree size is crucial. It involves considering factors like the complexity of patterns in the data, the risk of overfitting, and the need for a model that generalizes well to new, unseen data. Some common techniques to control the size of decision trees are:

- **Pruning:** Pruning selectively removes branches or nodes from a fully grown decision tree to prevent overfitting, promote generalization, and improve the model's performance on new, unseen data.
- **Minimum Sample Split:** This technique sets a threshold on the minimum number of samples required to split a node, ensuring that nodes with insufficient data won't split and preventing the creation of small branches.
- **Maximum Depth:** This technique controls the overall structure of the tree by limiting its maximum depth, preventing it from becoming excessively deep and complex.
- **Maximum Features:** This technique is valuable in scenarios with numerous features, as it allows control over the number of features considered for each split, preventing the model from becoming overly specialized and promoting better generalization to new data.

- **Minimum Leaf Size:** This technique involves setting a minimum number of samples required in a leaf node, preventing the creation of small, potentially noisy leaves and contributing to a more generalized model.

## 5.5.5 THE HIERARCHICAL NATURE OF DECISION TREES

Decision trees are structured hierarchically, resembling a family tree. The "root" of the tree corresponds to the initial decision or question. This root then branches into different possibilities based on certain conditions or features. Each branch represents a decision or outcome. As you move down the tree, you encounter more branches, each indicative of further decisions or distinctions.

The process continues until you reach the "leaves" of the tree. These leaves represent the final outcomes or predictions. Each path from the root to a leaf signifies a unique sequence of decisions based on the input features. It's akin to a series of if-else questions, where each decision point narrows down the possibilities until a specific outcome is determined.

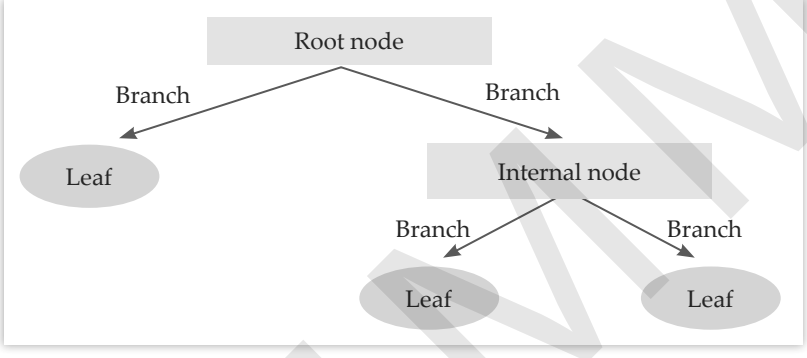



Figure 2: Hierarchical structure of decision

This hierarchical structure allows decision trees to systematically break down complex decision-making processes into simpler, more manageable steps. It facilitates the understanding of how a decision is reached and enables the model to adapt to various scenarios by learning from data patterns at different levels of granularity.

## 5.5.6 TRAINING THE DECISION TREE

Training a decision tree involves using a dataset to teach the model how to make decisions. The algorithm identifies patterns and relationships in the data, creating a hierarchical structure that allows the tree to predict outcomes or classifications for new, unseen instances based on learned rules. Steps involved in training the decision tree are:

1. **Initiating the Process:** The training of a decision tree begins with a dataset containing examples of inputs and corresponding outcomes.
2. **Selecting Features:** The algorithm identifies the most relevant features within the dataset to make decisions at each node of the tree.
3. **Create Decision Nodes:** It entails dividing the dataset based on the chosen feature; this action establishes decision nodes representing different paths within the tree.

4. **Repeat for Subsets:** This process is recursively applied to subsets; for each decision node, repeat steps 2 and 3 until the data is well-segmented into distinct categories.
5. **Assign Outcomes to Leaves:** This final step involves assigning outcomes to the leaves. Once the tree is built, determine the predicted outcomes for each set of conditions, completing the training process.

**Self Assessment Questions**

7. What is a decision tree in data mining?
   a. A tree used for landscaping
   b. A visual representation of data
   c. A hierarchical model for making decisions
   d. A database structure
8. Which of the following is not a component of a decision tree?
   a. Root node
   b. Decision node
   c. Leaf node
   d. Trunk node
9. What is the purpose of pruning in decision trees?
   a. To encourage overfitting
   b. To reduce the complexity of the tree and prevent overfitting
   c. To increase the depth of the tree
   d. To speed up the training process
10. Which algorithm is commonly used for constructing decision trees?
    a. K-means
    b. Support Vector Machines (SVM)
    c. 4.5
    d. Apriori

## 5.6 APPLICATION OF DATA SCIENCE FOR DECISION MAKING IN KEY AREA

Chief economist **Hal Varian** at Google announced that data scientist would be the'highly significant job in the 21st century'. Hence, now the question is: Why does the data science profession or job become important for all the professionals? And the straightforward answer is: In the past decade, there has been an explosion of information/data and computing power. We have created predictions benefiting the full human race by creating sense out of that knowledge. Knowledge Science helps organisations, governments, establishments, and people leverage knowledge to form strategic choices. Corporations such as Netflix, LinkedIn, Facebook and

Google have knowledge at the centre of their planning and competing benefits. As a lot of corporations adopt knowledge as their core quality, professionals have to be compelled to get adapted to the changing scenario of knowledge adoption. Let us discuss the importance of data science in some key areas or organisations which are as follows:

- Economics
- Telecommunications
- Bioinformatics
- Software engineering
- Healthcare
- ICT
- Logistics
- Process industry

Let us discuss the application of each field in detail.

### 5.6.1 ECONOMICS

After many weeks of interviewing, you have got job offers from three different companies. The offers differ greatly, which creates some confusion. You have created a small list of the offers:

- Giant national firm, $12 per hour beginning wage, insurance and dental advantages paid by the corporate, a two-week holiday each year, and potential for fast advancement.
- Little native firm, $20 per hour beginning wage, insurance and dental advantages offered. However, you need to pay the premiums, a 2-week without-pay vacation every year, share choices and program benefits, and potential for advancement.
- Regional firm, $15 per hour beginning wage, full insurance and dental advantages, one-week holiday, smart programme, and moderate advancement potential.

Regardless of the shape of the organisation or enterprise, success within the world of business sometimes depends on economic choices. As a result of economic decision making depending heavily on analysing data, it is crucial for that data to be helpful to economic decision makers. All economic choices of any consequence need the utilisation of some form of accounting data, typically within the style of money reports.

### 5.6.2 TELECOMMUNICATION

The evolution comes in the telecommunication industries due to the prompt expansion of mobile devices and smart phones and because of this, telecommunication industries need to gather huge amount of data from call records, data usage and server logs, etc., to get valuable information from this large amount of data in order to improve customer experience and growth of business.

Notes

Data science can help handle and improve accountability of data by boosting services related to network, customers and security. It is one of the best ways to better understand about the prospective customers in order to predict their action and behaviour and to deliver the right assistance and solutions in the telecommunication industries.

Data science in telecommunication helps in the following ways:

- Analysing the company's traditional databases
- Gathering, mapping and analyzing the data from different data sources
- Identifying the duration of the heaviest data uses over the network, and taking steps to solve it
- Analysing the customers who are facing problems with paying bills, and assisting them and taking appropriate steps to make recovery of payment easy
- Analysing the call and data statistics

### 5.6.3 BIOINFORMATICS

Being an interdisciplinary field, bioinformatics uses various computational methods which are used to analyse a huge amount of data such as cell count, genetic sequences and protein structures to predict new solutions and to expand biological understanding. Some of the technologies such as next-generation sequencing generate enormous amount of data which should be properly organised and clustered to make sense out of it. Big data like this, if utilised optimally, can be extremely useful in drug discovery or preventive medicine design. With the introduction of data science in this field, management of big data and data visualisation has been made very easy and scalable. Although bioinformatics and data science are two distinct disciplines, they share a common objective of cleaning, understanding and processing data. Presently, a large number of tools, based on machine learning are present which are used in bioinformatics. Recently, TensorFlow, a deep learning library from Google, has demonstrated how it can be used in biological computations. Application of data science in the field of bioinformatics is relatively new, but in short time, it has established itself as a great contributor for disease diagnosis and prediction.

### 5.6.4 SOFTWARE ENGINEERING

Software engineering can be defined as a detailed study of designing, developing and maintaining a software product. Since most of these software are directly used by a huge population or a corporation which is processing a large amount of data, these products are source of valuable information which can be used for various purposes. This is the reason why data science should matter to a software engineer very much. Over the course of time, many software engineers have realised that they can leverage this data flow to find solutions to their own questions, some of them being crucial, for example:

- Which feature of the software is most popular among consumers and which is not?
- What problem is being created by a bug and who should fix it?
- Before shipping, what should be the scalability of the software?

Since its advent, data science has been continuously shaping the decisions of software engineers by providing an insight into the scenario using the available data.

Let us take some cases where data science has affected the perspective to look at a problem. Research professionals, while examining the failures on Hadoop MapReduce and similar systems, found the following facts:

- Enough data was collected in error logs to reproduce data.
- Only a few nodes were required to debug the whole cluster.
- Simple testing and error handling could prevent a majority of failures.

This analysis was able to deduce a simple but vital conclusion that if engineers would have also tested that how their code behaved if things go wrong instead of right, most of the catastrophic failures could be averted. To improve software, a software engineer can implement statistical tools of data science to examine important information such as server metrics and logs. Data science helps them ask vital questions regarding a product and then find the answer using the relevant data.

### 5.6.5 HEALTHCARE

In the past decade, healthcare has also been impacted greatly by the advancement of technology. For instance, in US, approximately 1.2 billion documents related to healthcare of patients are generated annually. The collection, structuring and processing of such high amount of data are done to understand the health issues deeply. Thousands of data scientists and machine learning experts are providing their contribution in the healthcare industry in different ways and helping doctors to diagnose a patient with great accuracy and efficiency. Data science is also playing a key role in the advancement of healthcare industry.

Approximately, 2 terabytes of data is generated by the human body and due to advancement in the technology, we are capable of collecting most of it. This data can be related to a patient's heartbeat, sleeping patterns, levels of blood glucose and stress, brain activity, etc. Companies such as IBM, Apple and Qualcomm are providing advance equipment and framework to collect patients' data with high accuracy. Machine learning algorithms are also playing a vital role in detecting and tracking common health ailments related to heart or respiration. As regards the collection and analysis of the acquired patterns of heartbeat and breathing, the technology is advanced enough to detect disorders in the patient's health on the basis of the collected data. These days, obesity has emerged as a common ailment globally. To tackle this health issue, Omada Health, a medicinal company, has launched a data science-based preventive medicinal programme, called first digital therapeutic, to help obese patients in changing their daily routine to lose weight or keep the body weight under control. This program also helps patients in avoiding the risks which may occur on their body due to obesity.

Omada uses smart devices such as pedometer and scales to collect patient's behavioural data to customise its programme for helping every patient. This customisation acts as personal health coach for each patient, which helps in gaining deeper knowledge about the patient's health and modifies the programme accordingly.

Another deep learning organisation, Enlitic, uses data science for increasing the accuracy and efficiency of diagnosis. Enlitic has created a deep learning algorithm for reading imaging data generated using X-rays, CT scans and analysing it. It them compares its analysis with the results of clinical and laboratory reports. It has been found that the algorithm has delivered results with 70 per cent accuracy and fifty thousand times faster. Therefore, you can easily conclude that data science is playing a significant role in the field of healthcare and medicine.

### 5.6.6 INFORMATION AND COMMUNICATION TECHNOLOGY

Like healthcare, telecommunications and other sectors, Information and Communication Technology (ICT) has shown remarkable advancement in the past decade. The ICT companies collect data that are generated by the digital footprints of people surfing the Internet and the data are increasing exponentially. The storage capacity of the devices used for storing the data has increased ten times globally in a very short period of time. Various ICT technologies, such as Cloud Computing, the Internet of Things (IoT), are getting adopted to handle such large amount of data.

This also led to the evolution of data science for transforming data into business insights. The ICT technologies achieved high optimization in handling data by utilising distributed storage mechanism and computational capabilities due to innovation in data science. Both these technologies, the data science and the ICT, are now interdependent, which is immensely helpful for both these areas in aiding development. Data science plays a significant role in understanding the external and the internal reasons which are impacting the business. The reasons are determined using the data produced from social media platforms, search engines, organisational portals and are getting used extensively in widespread business applications. There is sharp rise in the demand of workforce who can work in interdisciplinary domain of data science and ICT. The data scientists are required who must have required skills and enough knowledge about the emerging information technologies and also must be capable enough to implement business solutions efficiently. Data science has somehow made the adoption of ICT technologies easier for the people. Moreover, one does not need to acquire high skills for adopting data science. Due to the availability of libraries and user-friendly applications, business users can create and implement data science easily and achieve swift business insights and results.

### 5.6.7 LOGISTICS

Global logistics sector is one of the fastest growing industries. While generating revenue at such a fast pace, it has yet to unleash the monetary and analytical potential of huge amount of data it generates. Several systems such as business systems, social media and lot of other devices generate huge data which can be helpful for both the consumer and the producer using business analytics.

Now, this huge amount of data needs data science to make sense out of it.

Let us take an example to elaborate the situation. Few years ago, if there was a problem in a vehicle, there were two ways to address this. First would be the owner taking his/her vehicle to the service garage or second would be a repair service van

getting to the location of the vehicle. Nowadays, smart vehicles can detect, analyse and report any kind of anomaly to the owner prior to any unexpected problem.

This saves huge cost and time for both the business as well as the consumers. Using data science, one can predict the future points of failure up to a significant precision using analytics and data-driven algorithms. Better prediction of demands has enabled logistics companies to successfully cut out their inventory by 20 per cent to 30 per cent along with increasing their fill rate by up to 7 per cent. These are only few of the applications from the domain of applications which can be implemented using data science and analytics.

Predictive analysis is the key to use data science successfully. Following are some examples in logistics industry:

- Since logistics vehicles need frequent maintenance and long lifetime on road, data analytics can be used to predict the mechanical parts which are most likely to fail and replace them to save time and money.
- Increase or decrease in demand can be predicted in a particular demography, which can make the company ready to move resources in a particular direction.

Data science techniques can help in improving various aspects such as automation, tracking vehicles more accurately and freight uses, including readiness for natural constraints such as weather. Multiple factors such as multiple data sources collated in a meaningful way, and the art of asking the right question when finely tuned with data science, packs the potential to push the logistic industry on a very profitable path from both the consumer's and the supplier's points of view.

### 5.6.8 PROCESS INDUSTRY

The process industries are those industries in which the production of items is done in large amount. The process industry requires reformation due to advancement of new information technologies. The mechanism of improving processes and gathering effective knowledge plays a significant role in all the facets of process industry, which include system integration, sustainability design, quality control, process control, decision support, etc. In order to automate the process industry from machine automation to information automation, and finally, to knowledge automation, data science and analytics are required. In the past few years, a large amount of data has been collected in the process industry because of the uses of scattered control systems. This large amount of data has been hardly utilised for detailed analyses. Nowadays, the significance of extracting information from the collected data has emerged and acquired the main role in the process industry.

The useful information from the collected data can be extracted by analysing the patterns of data. Based on the information, the following can be done:

- Processes can be monitored easily
- Faults can be diagnosed easily
- Quality can be measured easily
- Decision-making process can be made easy

NOTES

**SELF ASSESSMENT QUESTIONS**

11. Data science plays a vital role in different areas of decision making. (True/ False)
12. Data science is a booming industry and it helps decision maker to take right decisions in short period. (True/False)
13. Sentiment analysis (customer behaviour) is not possible through data science. (True / False)
14. _____ a deep learning library from Google has demonstrated how that can be used in biological computations.

## 5.7 SUMMARY

- The process of decision making is an integral part of the managerial process in any organisation.
- The decision-making process is collective and consultative and has a considerable impact on the overall growth and prospects of the organisation.
- Decision is taken to support organisational growth.
- Discussions and consultations along with standard procedures and techniques are the two main tools that maintain and eventually facilitate decision making.
- An extensive debate and research is required before finalising a decision.
- The most important responsibility of management in any organisation is to set up organisational goals and allocate the available resources effectively and efficiently.
- The decision-making process involves a series of steps that we need to take logically.
- Decision Support Systems (DSS) belong to particular categories of information systems that support the decision-making activities.
- DSS uses analytical models, dedicated databases, insight and judgement of the decision maker and an interactive, computer-based modelling approach to sustain unstructured decisions.
- DSSs assist in assessing and resolving the questions posed every day in businesses.
- Economic decision making depending heavily on analysing the data, it is crucial for that data to be helpful to economic decision makers.
- Data science can help handle and improve accountability of data by boosting services related to network, customers and security.
- Thousands of data scientists and machine learning experts are providing their contribution in the healthcare industry in different ways and helping doctors to diagnose a patient with great accuracy and efficiency.
- Machine learning algorithms are also playing a vital role in detecting and tracking common health ailments related to heart or respiration.

- The ICT companies collect data that are generated by the digital footprints of people surfing the Internet and the data are increasing exponentially.
- Various ICT technologies, such as Cloud Computing and the Internet of Things, are getting adopted to handle such large amount of data.

## 5.8 KEY WORDS

- **Data science:** It is a perfect combination of various tools, algorithms and machine learning principles to achieve hidden patterns from huge amount of data.
- **Big data:** The name suggested the data which is big or large. Nowadays, the variety of data available in market coming from different sources and our traditional data storage and data processing application software are unable to deal with.
- **Machine learning:** Human training process is become outdated, now technologies provide systems the ability to automatically learn and improve from experience without any explicit programming.
- **Deep learning:** This technology is a part of machine learning and inspired by the structure and function of the brain called artificial neural networks. Nowadays, deep learning algorithms are used for reading imaging data generated using X-rays, CT scans and analysing it.
- **DSS:** DSS stands for Decision Support System and it is used to analyse business data and present an interactive information support to all decision makers.

## 5.9 CASE STUDY: ROLE OF ANALYTICS IN FINANCIAL DECISION MAKING

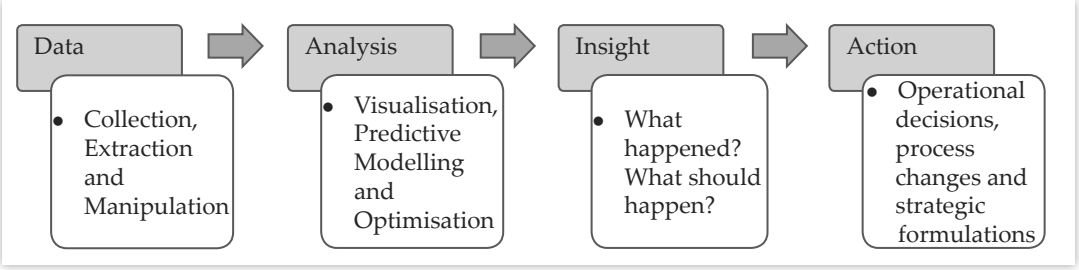
Organisations today collect large amounts of data from internal and external sources. The current technology leads to collection of a large amount of data in an efficient manner. However, leveraging this data and extracting the maximum value from it to improve its market positioning is a challenge. Information derived by analysing the data is used to make certain decisions that directly or indirectly have a financial impact on the organisation in terms of market share, profitability and customer satisfaction. Here, the role of Business Analytics (BA) becomes relevant. BA includes a lot of advanced statistical techniques and tools which can be used to extract the maximum value from the available information or data. In a research conducted by Lesser, Lavalle, Shockley, Hopkins and Kruschwitz in 2011, it was found that high-performing organisations may make use of BA more often than their low-performing peers.

BA helps an organisation in various types of decision making. However, the most prominent decisions for an organisation are the financial decisions. Therefore, the use of BA in making financial decisions must be evaluated thoroughly.

While using BA, it must be remembered that it involves three elements, namely technological, human and business processes. BA is the process of turning data into actions by using various tools and techniques.

NOTES

The process of BA involves steps as shown in the following figure:




A wireless provider **Ipro** is a large private organisation which provides wireless services in the United States. IPro is the organisation that introduced innovative voice and data services. Ipro's major contribution lies in providing advanced wireless services in rural America, which were earlier available only in the urban areas.

Until recently, Ipro was using Excel spreadsheets and an old general ledger system. The data from these sources was fed into a financial reporting tool which did not have the required functional and infrastructure support to prepare and generate comprehensive budget. Different information or data sets required to prepare a budget were on disparate spreadsheets which were not designed to handle budgeting.

To overcome this problem and to be able to use the available data, Ipro's technology team decided to implement the IBM Cognos TM1 solution. Cognos TM1 is a Business Analytics software tool developed by IBM to aid an organisation in financial performance management and reporting. The Cognos software tool helped **Ipro** in the following ways:

- All the relevant departments of Ipro were able to input detailed information in a central place.
- Finance department did not need to look up for individual invoices.
- Individual transactions that added up to totals in the financial reports could be referenced individually.
- Transparency.
- Up-to-date information was available to all the users.
- Fast and flexible reporting.
- Ad-hoc analysis.
- Cognos could interface with Excel.
- Planning, budgeting, forecasting and analysis processes were automated.
- Financial results and analysis were integrated with operational plans forfaster execution.
- Cognos provided visualisations and could make predictions.
- Cognos could perform what-if analysis and test alternative assumptions.

- Timely and reliable plans could be made and these helped in turning insights into actions.

**Source:** *https://www.ibm.com/developerworks/community/blogs/9ca03a38-fc0f-4a62-8f5f-f163b1d03769/entry/Impact_of_Analytics_in_Financial_Decision_Making_Evidence_from_a_Case_Study_Approach?lang=en*

## QUESTIONS

1. Why the role of Business analytics(BA) is relevant in organisations?

   (**Hint:** BA includes a lot of advanced statistical techniques and tools which can be used to extract the maximum value from the available information or data.)

2. How many elements are involved while using BA?

   (**Hint:** While using BA, it must be remembered that it involves three elements, namely technological, human and business processes.)

3. What are the four steps involved in Business Analytics? Explain in the context of Ipro.

   (**Hint:** Four steps involved in Business Analytics are: Data, Analysis, Insights and Actions. Cognos provided **IPro**timely and reliable plans which helped in turning insights into actions.)

4. Why did the organisation Ipro feel the need to deploy a Financial Business Analytics software?

   (**Hint:** Ipro was using Excel spreadsheets and an old general ledger system. The data from these sources was fed into a financial reporting tool which did not have the required functional and infrastructure support to prepare and generate comprehensive budget.)

5. How Cognos software helped orgnisation Ipro in overcoming the problems?

   (**Hint:** The Cognos software tool helped IPro in the following ways:

   - All the relevant departments of Ipro were able to input detailed information in a central place.
   - Finance department did not need to look up for individual invoices.)

## 5.10 EXERCISE

1. What do you understand by decision making?
2. What are the differences between programmed decisions and non-programmed decisions?
3. Explain the significance of data science in ICT.
4. List the steps involved in the decision-making process.
5. Write short note on the following topics:
   - Strategic decisions
   - Tactical decisions

## 5.11 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Concept of Decision Making | 1. | a. Programmed decision |
| | 2. | a. VI,II,VII,III,I,IV,V |
| Understanding Decision Support System | 3. | d. All of the above |
| | 4. | True |
| Techniques of Decision Making | 5. | Models |
| | 6. | facts, mathematical |
| Data Mining with Decision Trees | 7. | c. A hierarchical model for making decisions |
| | 8. | d. Trunk node |
| | 9. | b. To reduce the complexity of the tree and prevent overfitting |
| | 10. | c. 4.5 |
| Application of Data Science for Decision Making in Key Area | 11. | True |
| | 12. | True |
| | 13. | False |
| | 14. | Tensor flow |

## 5.12 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Sauter, V. (2014). Decision Support Systems for Business Intelligence.Somerset: Wiley.
- Lawrence, K. and Kleinman, G. (2010). Applications in multicriteria decisionmaking, data envelopment analysis, and finance. Bingley, UK: NorthAmerica.


### E-REFERENCES


- Towards Data Science. (2018). How Data Science Is Enabling Better Decisionmaking. [online] Available at: https://towardsdatascience.com/how-datascience- is-enabling-better-decision-making-1699defd6899 [Accessed 20 Nov.2018].
- Decision-making-solutions.com. (2018). Decision Making Techniques.[online] Available at: https://www.decision-making-solutions.com/decision_making_techniques.html [Accessed 20 Nov. 2018].
- Essays, Research Papers and Articles on Business Management. (2018). Top 7Steps involved Decision-Making Process. [online] Available at: http://www.businessmanagementideas.com/decision-making/top-7-steps-involveddecision-making-process/3363 [Accessed 20 Nov. 2018].

CHAPTER 6

# Introduction to OLTP & OLAP

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Explain the need for online analytical processing
- Describe the difference between OLAP & OLTP
- Discuss the different types of OLAP architectures
- Examine the comparison between OLTP & OLAP
- Define the different types of OLAP Operations

## 6.1 INTRODUCTION

In the previous chapter, you studied the basic concept of decision making. The chapter explained the various types of decisions, processes, decision support system and different techniques used in decision making. You also studied the different applications of data science for decision making in different areas such as economics, telecommunication, engineering, healthcare, and information and communication technologies.

Both the OLTP (Online Transaction Processing) and the OLAP (Online Analytical Processing) are the data processing systems but process the data differently. OLTP is customer-oriented software which has a large number of users who perform short transactions such as order entry, retail sales and financial transactions. It is used by IT professionals to manage the current data or transactions but does not use historical data. Historical data is that old data of an organisation which is used for decision making and this software is not for that purpose. OLTP follows an access pattern which consists of short, atomic transaction, recovery mechanisms and concurrency controls.

On the other hand, OLAP is market-oriented software which is used by analysts and managers of an organisation to perform data analysis in multiple dimensions. Multiple dimensions of a business may be planning, budgeting, financial reporting, analysis, forecasting, etc. It helps decision makers to take better decisions for the business. In OLAP, different architectures or data models are available. Some of the OLAP architectures are Relational Online Analytical Processing (ROLAP), Multidimensional Online Analytical Processing (MOLAP), Hybrid Online Analytical Processing (HOLAP) and Desktop Online Analytical Processing (DOLAP).

All architectures involve in the creation of a multidimensional data structure where dimensions denote business entities. Multidimensional data structure typically runs at off-peak uses times to build and update a persistent data structure. OLAP provides a user-friendly environment for interactive analysis and querying of data. The OLAP operations are slicing and dicing, drill down, roll-up and pivot, which will be discussed later in the chapter.

This chapter begins by explaining the concept of OLTP and OLAP in detail. The chapter covers the needs of OLAP and different OLAP architectures. It describes the difference between OLAP and OLTP. Moreover, this chapter also covers OLAP operations such as roll up, drill down, slice, dice and pivot.

Notes

## 6.2 ONLINE TRANSACTION PROCESSING

As discussed earlier, OLTP processes a large number of client transactions. Daily transactional data is stored in OLTP systems and some basic commands such as insert, update and delete can be performed on the data. It is carried in a client-server system. Most of the organisations use Database Management System (DBMS) to support Online Transaction Processing (OLTP). Some examples of OLTP systems are financial transaction systems (banks), retailers, order entries, airlines, etc.

OLTP transactions are very particular in performing their tasks and they generally involve single record or small group of records. For example, a person may transfer money from his/her account to his/her friend's account. In this case, the transactions will involve only two accounts, i.e., the account of the person and his/her friend's account. Some of the characteristics of an OLTP application are as follows:

- It involves a small amount of data
- It allows indexed accessing of data
- It can involve a large number of users
- It allows frequent queries and updates
- It allows fast responses

In OLTP, more than one user can access the same data from the database system, which is possible because of concurrency. In a concurrent environment, if a user wants to make any changes in the data, then he or she has to wait until other users finish the processing.

In OLTP, atomicity ensures that all the steps of a transaction must be completed altogether. If any step in a transaction fails, then all subsequent steps must fail. For example, for booking a seat in a train, a passenger needs to first reserve a seat and then make a payment for it. Atomicity consolidates the two actions, i.e., first reserve a seat and then makes a payment to book a seat. If the payment fails, then the booking will not be complete subsequently.

### Self Assessment Questions

1. _____ is mainly used in industries for efficient processing of a large number of client transactions.
2. OLTP is not carried in a client-server system. (True/False)
3. Online transaction process concerns about _____ and _____.
4. _____ controls allow two users to access the same data in the database system.
5. _____ controls allow that all the steps in a transaction are completed successfully as a group.
6. OLTP brokering programs can distribute transaction processing among multiple computers on a _____.

**Activity**

Search the role of ACID properties in OLTP and make a report.

## 6.3 ONLINE ANALYTICAL PROCESSING

To attain informed business decisions and take the required actions, business analysts and executives often need to analyse business data from different aspects. Consider the case of a business organisation that sells different household products across the different regions of the country. To arrive at an appropriate decision about increasing the sale of a particular product, business executives need to analyse data according to the market trends and popularity of the product in various regions. The business executives may also need to compare the marketing statistics of a product during the different seasons of a year or a particular period in different years. To perform the comparison, the business data needs to be stored in multidimensional databases. The selective extraction and analysis of data are then accomplished by using the Online Analytical Processing (OLAP) method.

OLAP facilitates the performance of complex calculations, trend analysis and complicated data modelling. OLAP is different from OLTP, which is another processing approach characterised by various online transactions, including INSERT, UPDATE and DELETE.

The following points depict the need for OLAP in organisations:

- Tool for analysis of huge data, i.e., billions of rows or petabytes of data, was in demand. The analysis must be fast and should be interactive.
- There was a need for multidimensional analysis as data was coming from various parts of an enterprise or from data marts. For example, data was getting generated from purchase department, sales department, stores, etc. Every department has its way of storing data and some important attributes related to it. Thus, you need a combined analysis with respect to all departments or some departments with their important attributes.
- In the multidimensional analysis, a user expects fast processing, complex calculation, data integrity, concurrency, atomicity, etc. Traditional tools such as query products, spreadsheets, language interfaces and report writers are not capable enough to satisfy a user's high expectations.
- Standard tools can be used for analysis, but when data grows in size, performance gets hampered. It requires more time for complex analysis over a set of varied attributes from different dimensions.
- Complex analysis using OLAP is useful to senior management of various companies to forecast future trends.

**Self Assessment Questions**

7. OLAP is a processing approach that performs multidimensional analysis of business data. It also facilitates the performance of

   a. Complex calculations
   b. Trend analysis
   c. Complicated data modeling
   d. All of these

Notes

8. OLAP and OLTP are same. (True/False)
9. In complex analysis, OLAP is not useful for senior management to forecast future trends. (True/False)
10. OLAP architectures are
    a. ROLAP
    b. KOLAP
    c. TOLAP
    d. All of these
11. _____ data structure is constructed when some OLAP technologies require an ETL (extract, transform and load) process, which typically runs at off-peak usage times to build and update a persistent multidimensional data structure.

**Activity**

Create a PowerPoint presentation on OLAP and OLTP and show it in your class.

## 6.4 DIFFERENT OLAP ARCHITECTURES

In the market, different OLAP architectures are available, but the mainly used OLAP architectures are ROLAP, MOLAP, HOLAP and DOLAP. In the OLAP architectures, a multidimensional data structure is created, where dimensions refer to business entities (i.e., sales regions, products, time, geography, etc.). A multidimensional data structure is created whenever some OLAP technologies need an extract, transform, and load (ETL) process. An ETL process normally runs at a time when demand is less to construct and update a persistent multidimensional data structure. Let us discuss different OLAP architectures in detail.

### 6.4.1 ROLAP

In ROLAP, data will not be in a summarised form because the size of the database is greater than 100 GB. Its response time is slow or you can say poor, but it depends on the query type. Figure 1 shows the architecture of ROLAP.

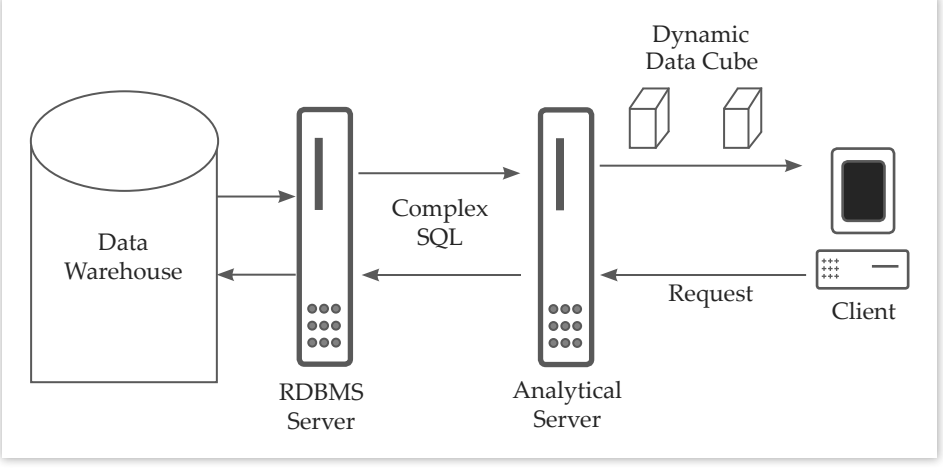



Figure 1: Architecture of ROLAP

In Figure 1, the ROLAP clients and a database server both are based on RDBMS. The OLAP server sends request to the database server and obtains the response.

The multidimensional cubes are generated dynamically as per the requirement sent by the user. To hide the structure from the user, a layer of metadata is created.

This layer supports mapping between a relational model and business dimensions. In this architecture, the user interacts with the database frequently. The query is processed by the OLAP server and/or the database server.

The advantages of the ROLAP architecture are as follows:

- It provides flexibility as it can address ad-hoc queries also. Moreover, there is no limit on the number of dimensions.
- Its tools are less expensive.
- RDBMS requires less storage.
- It requires low network bandwidth because only the results are sent to the client over the network.
- Its tools are simple and easy to use.

The disadvantages of the ROLAP architecture are as follows:

- Its response time is poor.
- Iterative analysis and drill-down are very slow.

### 6.4.2 MOLAP

As ROLAP works on RDBMS, executing complex queries hampers the overall performance. In order to overcome this drawback, the MOLAP architecture was proposed with the idea that multidimensional cubes must be pre-computed. Figure 2 shows the architecture of MOLAP:

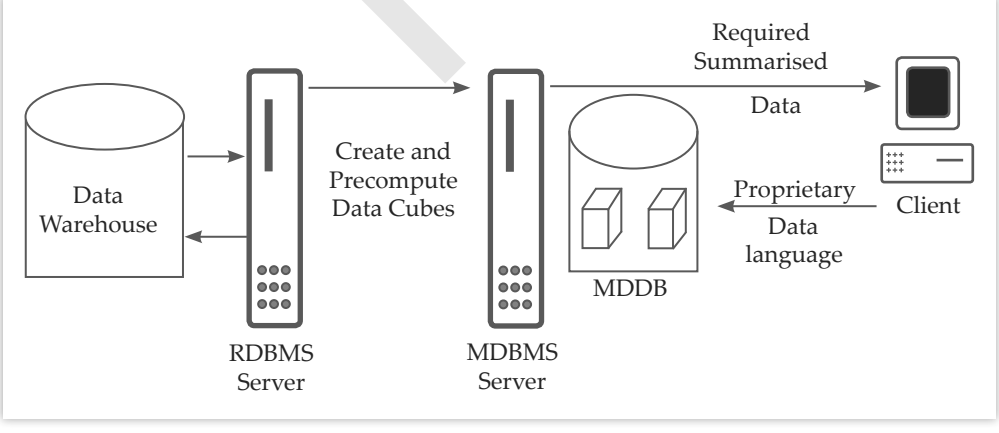



FIGURE 2: Architecture of MOLAP

The architecture of the MOLAP system is shown in Figure 2. The system consists of the following:

- The OLAP client that provides the front-end GUI for giving the queries and obtaining the reports.

- The OLAP server which is also known as Multidimensional Database Management System (MDBMS) server. This is a proprietary database management system which stores the multidimensional data in 'multidimensional cubes' and contains the data in the summarised form, based on the type of reports required.
- A machine that carries out data staging, which converts the data from RDBMS format to MDBMS and sends the 'multidimensional cube' data to the OLAP server.

The multidimensional database uses large multidimensional arrays as storage structures. Multidimensional database management systems are not open source software, but they are proprietary software systems. Consolidation and fabrication of summarised cubes are provided by these systems.

The main advantages of the MOLAP are as follows:

- It provides excellent performance in terms of response time to queries as the data is already stored in the summarised form.
- Only the results are sent to the client machine, and hence the network bandwidth requirements are low.
- Drill-down is very fast.
- It provides an interface that is effective and easy to use.
- Complex analysis can be done using the MOLAP servers.

The disadvantages of the MOLAP are:

- MDBMS are proprietary to the vendor, who sells the data warehouse development tools.
- The data cubes require high storage.
- The data cubes have to be updated periodically as the summaries keep changing.
- MOLAP tools are costlier as compared to ROLAP tools.
- MOLAP lacks flexibility because the data cubes in the server have to be predefined keeping in view the requirements analysis. Hence, drill-up is very difficult.

Hence, multidimensional OLAP is used when the database size is small, say less than 100 GB, and the data is available in the summarised form.

### 6.4.3 HOLAP

This system tries to accommodate the advantages of the ROLAP and the MOLAP models. The ROLAP has a good database structure and simple queries can be handled efficiently. On the other hand, the MOLAP can handle complex aggregate queries faster. However, MOLAP is computationally costlier.

Therefore, you can have a midway. In HOLAP, the relational database structure is preserved to handle simple and user-required queries. Instead of computing all the data cubes and storing them in the MOLAP server, HOLAP server stores only some important and partially computed cubes or aggregates so that when we require

higher scalability and faster computation, the required aggregates can be computed efficiently. Thus, HOLAP possesses advantages of both the ROLAP and the HOLAP.

Figure 3 shows the architecture of HOLAP:

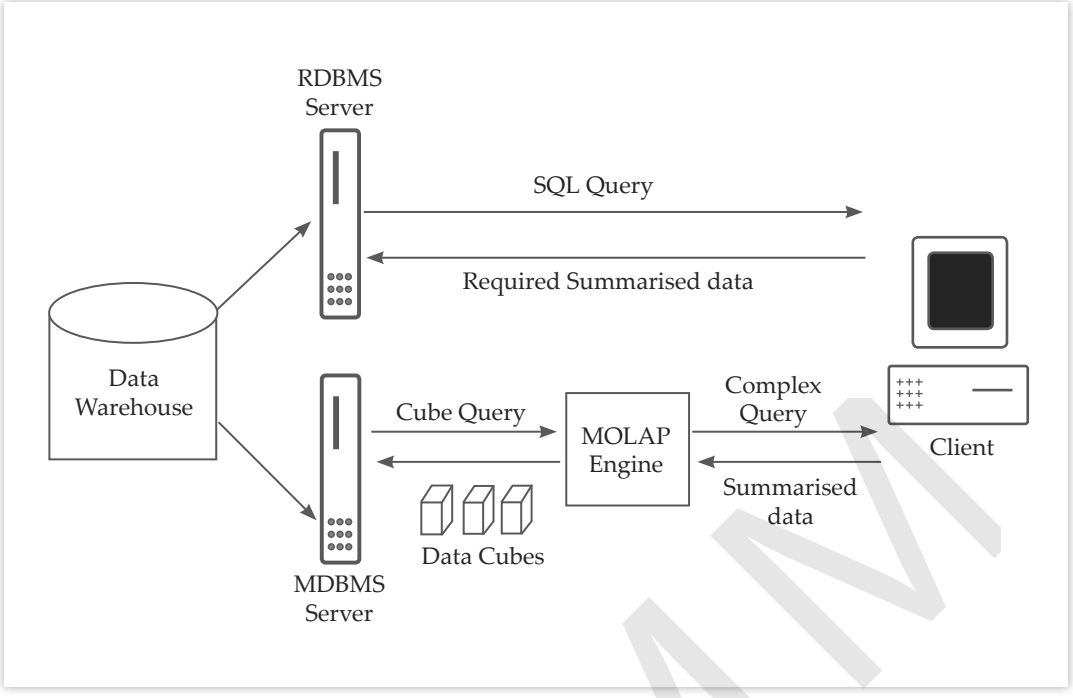



FIGURE 3: HOLAP Architecture

The advantages of the HOLAP are as follows:

- Better performance compared to ROLAP and MOLAP models.
- Users can write simple as well as complex queries to get the required data.
- In HOLAP, the processing time is less than MOLAP.

The disadvantages of the HOLAP are as follows:

- Chances of data redundancy in the HOLAP server may create problems in a network of low bandwidth.
- The HOLAP server can maintain only a limited amount of data.
- Whenever a new record is inserted, it needs to be processed. HOLAP does not process any new record.

### 6.4.4 DOLAP

It is a technology by which the user is able to download a part of OLAP locally and then may perform the required processing locally. DOLAP provides data after performing a multidimensional analysis of the client machine, which has already obtained the data from multidimensional database servers.

Notes

DOLAP is a single-tier technology that enables user to download a small hypercube on their desktop machine from a central point or server to perform multidimensional analyses. This central point can be a data mart or data warehouse.

Note that in the case of DOLAP, during the process of multidimensional analyses, the client machine remains disconnected from the server.

The advantages of the DOLAP are as follows:

- It facilitates user to modify or calculate data locally from the result set obtained from the server.
- It provides good query performance in collecting, aggregating and calculating the data during multidimensional analysis.
- It is useful for mobile users as they cannot always stay connected with the data warehouse.

The only disadvantage of the DOLAP is that it provides limited functionality and data capacity.

**Self Assessment Questions**

12. ROLAP is the preferred technology when the database size is large. (True/False)
13. ROLAP response time is fast, minutes to hours, depending on the query type. (True/False)
14. ROLAP systems are based on _____ model.
15. The _____ architecture was proposed with the idea that multidimensional cubes must be pre-computed.

## 6.5 COMPARISON BETWEEN OLTP & OLAP

The traditional RDBMS technology and the data warehouse technology differ in terms of the type of processing. The database applications are tuned for OLTP, whereas the data warehouses are tuned for OLAP.

OLTP systems are used to store transactions on a daily basis. They are used majorly to carry out update, delete and insert operations. On the other hand, OLAP is a software technology majorly used by Business Analysts, Project Managers and CEOs to gain insight into data using the various views in a fast, interactive and consistent way. They may view data in a summarised fashion with the dimensions of their choice. Table 1 summarises the differences between OLTP and OLAP:

**Table 1:** Differences Between OLTP and OLAP

| Differential Point | OLTP | OLAP |
|---|---|---|
| Data source | Operational and current business data pertaining to day to day business activities | Consolidated data obtained by complex aggregated functions |

| Differential Point | OLTP | OLAP |
|---|---|---|
| Purpose of data | Management of fundamental business activities | Support planning, problem solving and strategic decision making |
| What the data represents | A snapshot of ongoing business activities | A multidimensional view of different business activities |
| Inserts and updates | Short, fast and frequent inserts and updates by end user | Periodic updates by batch jobs |
| Queries | Standardised, simple SQL queries returning a few records | Complex queries involving aggregation |
| Query processing speed | Very fast | Relatively dependent on the amount of data involved |
| Space requirement | Relatively small | Relatively large due to the existence of historical aggregated data |
| Database design | Normalised tables | A few de-normalised tables using a star or snowflake schema |
| Backup and recovery | Frequently made | Rarely made, instead, data is reloaded |
| Application | Operational, Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), System Control Module (SCM), call centres, Point-of-Sale applications, etc. | Management Information Systems (MIS), Decision Support Systems (DSS), etc. |
| Users | Employees | Managers and executives |

**Self Assessment Questions**

16. The database applications are tuned for _____ whereas the data warehouses are tuned for _____.
17. OLTP is used to store transactions on a daily basis. (True/False)

## 6.6 OLAP OPERATIONS

OLAP provides an environment which is user-friendly for interactive data analysis. The OLAP operations allow viewing, analysis and querying of data.

The popular OLAP operations are as mentioned below:

- Slicing
- Dicing
- Drill-down
- Roll-up
- Pivot

Let us discuss each operation in detail.

Notes

### 6.6.1 SLICING

In the slice operation, we consider one particular dimension from any cube and provide a new sub-cube. Here, you get the details about one dimension by keeping other dimensions at a particular level.

Figure 4 shows the slice operation:

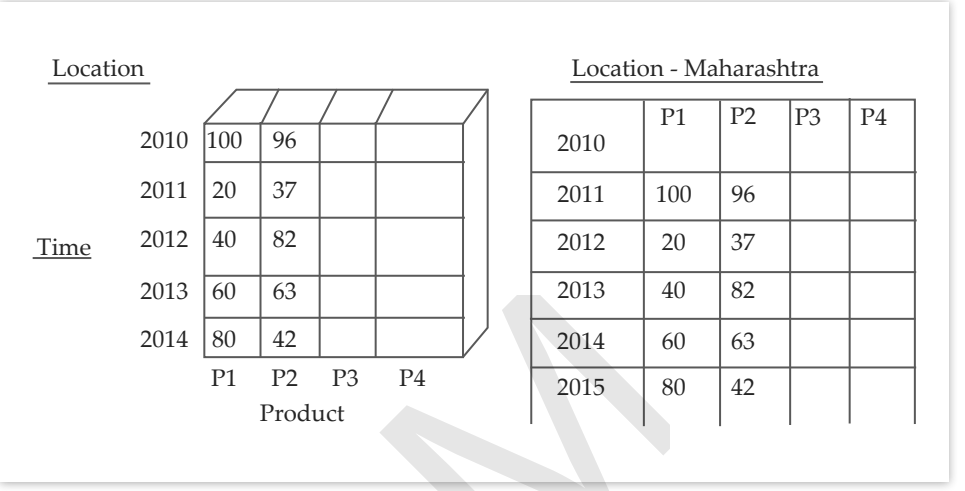



**Figure 4: Slice Operation**

In Figure 4, you consider only the location dimension with values of Maharashtra.

### 6.6.2 DICING

The logical partitioning of a cube considering more than one dimension is called dice operation.

Here, we get a sub-cube by selecting two or more dimensions. Figure 5 shows dice operation:

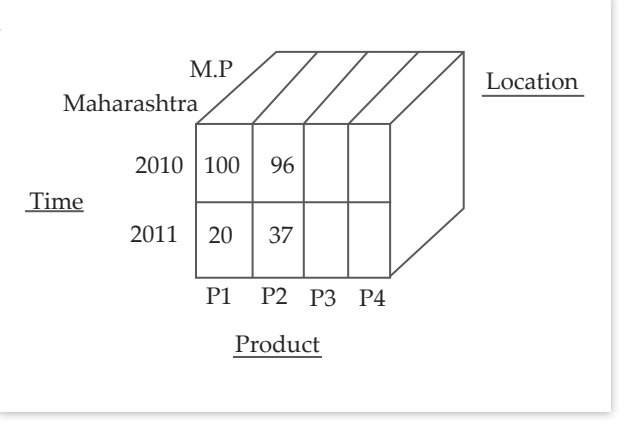



**Figure 5: Dice Operation**

Figure 5 shows a dice operation considering time in 2010 and 2011 and locations as Maharashtra and M.P.

### 6.6.3 DRILL-DOWN

Drill-down operations help in going to lower levels in the hierarchy. If you consider the time dimension, then the drill-down operation will help you in getting aggregates from month to week to day by exploring details at the lower levels. Figure 6 shows the diagram of a drill-down operation:

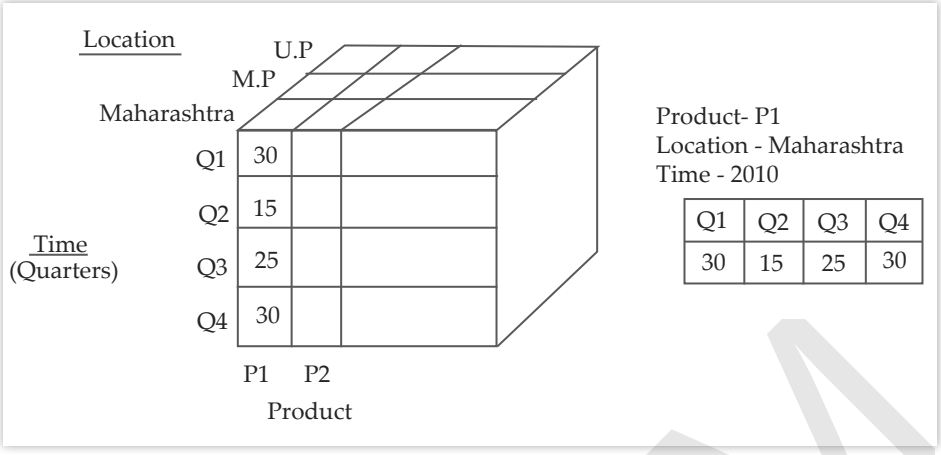



Figure 6: Drill-down Operation

In Figure 6, the drill-down operation is performed in which we are drilling down a year to find out the lower level details from Quarter 1 to Quarter 4.

### 6.6.4 ROLL-UP

Roll-up operations are used to climb up at a higher level in the hierarchy. If you consider the time dimension, then the roll-up operation will help you in getting aggregates for a day, a week or a month by summing up at lower levels. Figure 7 shows the diagram of a roll-up operation:

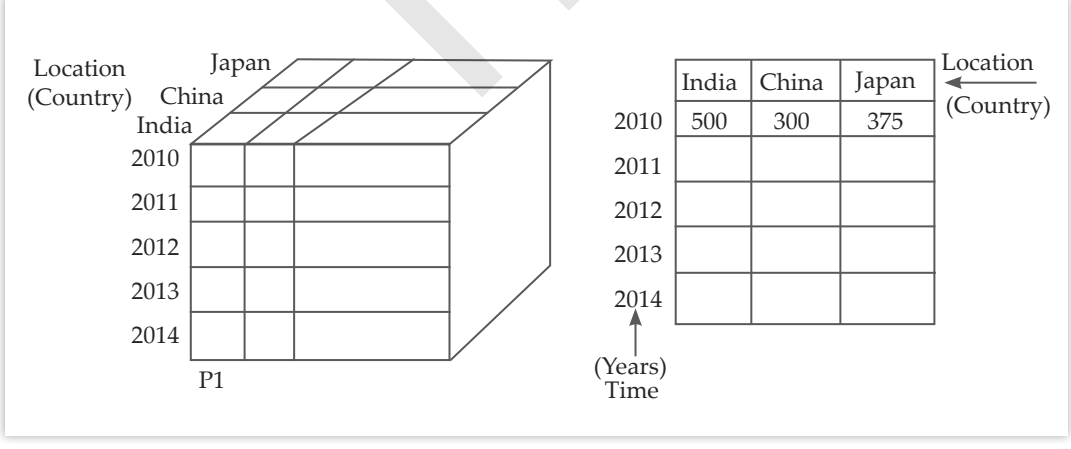



Figure 7: Roll-up Operation

In Figure 7, you are performing the roll-up operation for hierarchies of location dimension (from states to countries).

Notes

### 6.6.5 PIVOT

In pivot, an analyst can rotate a cube to see its different faces. It is a visualisation operation. Figure 8 shows the pivot operation:

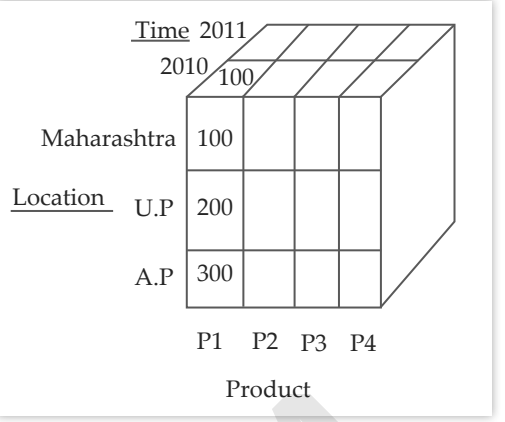



**Figure 8: Pivot Operation**

**Self Assessment Questions**

18. OLAP operations are
    a. Slicing and Dicing
    b. Drill-down and Roll-up
    c. Pivot
    d. All of these
19. In the _____ operation, you consider one particular dimension from any cube and provide a new sub-cube.
20. The logical partitioning of a cube considering more than one dimension is called _____ operation.

## 6.7 SUMMARY

- OLTP is used to process a large amount of client transactions.
- OLTP is carried in a client-server system.
- Concurrency and atomicity are the two basic concepts of OLTP.
- In OLTP, more than one user can access the same data from database system. This is possible because of concurrency.
- Atomicity consolidates more than one transaction together and ensures that all transactions must be completed successfully as a group. If any transaction gets failed, then all other steps get failed automatically.
- OLAP is a processing approach that performs multidimensional analysis of business data. It also facilitates performance of complex calculations, trend analysis and complicated data modelling.
- In complex analysis, OLAP is useful for senior management to forecast future trends.
- Different types of OLAP architectures are ROLAP, MOLAP, HOLAP and DOLAP.

- Multidimensional data structure constructed when some OLAP technologies require an ETL process, which typically runs at off-peak usage times to build and update a persistent multidimensional data structure.
- Relational OLAP is the preferred technology when the database size is large.
- ROLAP response time is poor, minutes to hours, depending on the query type.
- ROLAP systems are based on the relational data model.
- In ROLAP, the multidimensional cubes are generated dynamically as per the requirement sent by the user. To hide the structure from the user, a layer of metadata is created. This layer supports mapping between the relational model and business dimensions.

## 6.8 KEY WORDS

- **Online Transaction Processing (OLTP):** It refers to customer-oriented software which has a large number of users who perform short transactions like order entry, retail sales, financial transactions, etc.
- **Online Analytical Processing (OLAP):** It refers to market-oriented software which is used by analysts and managers of an organisation to perform data analysis in multiple dimensions.
- **Concurrency**: It allows two users to access the same data in the database system.
- **Atomicity**: It allows all the steps in a transaction to be completed successfully as a group.
- **Web service:** A service offered by an electronic device to another electronic device, communicating with each other via the World Wide Web.
- **Data marts:** It is used to store data which is specific to a particular group.
- **Relational Online Analytical Processing (ROLAP):** It is based on the relational data model.
- **Multidimensional Database Management System (MDBMS) server:** It is aproprietary database management system which stores the multidimensional data in 'multidimensional cubes' and contains data in the summarised form, based on the type of reports required.

## 6.9 CASE STUDY: DEPLOYMENT OF AN OLAP SOLUTION

Agecol, Inc. is a leading outsource collection agency for government debts in the USA. Agecol approached Win metrics as it wanted to implement an OLAP system to meet its various needs. All the financial service-related organisations need to prepare and provide regular performance reports to their clients and other stakeholders. Agecol prepares its performance report and names it as CARE report. Agecol's clients depend majorly on the recovery percentages as reflected in the CARE report. Initially, Agecol deployed its CARE report using a 40-page C program. However, the CARE report for all clients was taking over 24 hours to process. In addition, it generated a lengthy and inflexible report. The report was inflexible as the level of detail was fixed and the report could not focus on a single client. The results generated in the report could not be analysed, further, for example, distinguishing clients by loan type. The overall results were displayed, but there was no way to

Notes

validate the results using the detailed records used to prepare it. It was felt that Agecol required the CARE information delivered to them in the form of an Excel pivot table.

Merrill Eastman, the then CEO of Agecol, decided to implement an OLAP system. During the implementation, the old CARE report had to be transformed into an OLAP cube. The OLAP system's implementation was being favoured because it pre-computes numeric aggregations for cross-product of all the relevant dimensions so that summary information for any combination of dimensions can be displayed on demand. In simple terms, the OLAP system transforms a relational database into a pivot table.

Agecol knew that there are a number of OLAP software systems available in the market. However, they decided to adopt SQL Server Analysis Services because of the following reasons:

- Agecol owned MS SQL Server licenses
- MS SQL Server is easy to use and administer
- MS SQL Server provides bundled Analysis Services
- MS SQL Server is tightly integrated with MS Excel which is used extensively for financial reporting and analysis.

The pace of implementing the OLAP CARE report was slow because the CARE report specifications were not fixed. In addition, the people who were implementing OLAP were not able to decide as to how the content of the old CARE report could be extracted out of the OLAP cube. Other important challenges that required to be overcome were as follows:

- Exporting 80M facts and dimension rows from Informix to SQL Server in less than 4 hours
- Transforming the exported information into an SQL Server data mart with no referential integrity errors
- Computing the distinct counts within the cube that had different granularity than the basic revenue facts
- Mapping the same facts to multiple members in the same dimension
- Deciding what ragged hierarchies should be used as dimensions of the cube
- Deciding how cube aggregates can be validated and understood
- Tying CARE cube aggregates to General Ledger to validate data integrity

The first CARE cube was delivered in 8 weeks and soon the second cube was also delivered. By that time, Agecol recognised the importance of SQL Server OLAP system. Three software engineers of Agecol were trained to build cubes and they developed General Ledger, General Ledger Budget, Payroll, Collector Performance and Revenue Forecasting cubes. The key benefits realised by Agecol after implementing the OLAP system were as follows:

- Agecol could develop the performance report for its clients.
- Agecol saved $200K in accounting software expenses.

- OLAP system implementation helped the organisation in selling $167,000,000 of equity to Venture Capitalists who were impressed by Agecol's performance in just 9 months. This meant increased investment for the organisation.
- Balance Sheets and Profit and Loss statements are implemented in an account roll-up dimension.
- Agecol can drill down to any level of detail, especially in the General Ledger Budget cube.
- The time required to close accounting books was reduced by 5 days.

**Source:** *http://www.winmetrics.com/olap_casestudies.html*

## QUESTIONS

1. What are the limitations with Agecol's 40-C program?

   (**Hint:** The CARE report for all clients was taking over 24 hours to process. In addition, it generated a lengthy and inflexible report.)

2. Assume that you are a client of Agecol, Inc. While analysing Agecol's financial performance, what metrics would you look out for?

   (**Hint:** Agecol's clients depend majorly on the recovery percentages.)

3. Why was OLAP system's implementation being favoured?

   (**Hint:** It precomputes numeric aggregations for cross-product of all the relevant dimensions so that summary information for any combination of dimensions can be displayed on demand.)

4. Why Agecol decided to adopt SQL Server Analysis Services?

   (**Hint:** Agecol decided to adopt SQL Server Analysis Services because of the following reasons:

   - MS SQL Server is easy to use and administer
   - MS SQL Server provides bundled Analysis Services

5. List a few benefits accrued by Agecol which could be presented in front of other financial organisations to help them decide whether or not they should implement the OLAP systems in their respective organisations.

   (**Hint:** Agecol saved $200K in accounting software expenses. In addition, the time required to close accounting books was reduced by 5 days, etc.)

## 6.10 EXERCISE

1. Write a short note on the need for OLAP?
2. Why is Roll-up operation important?
3. What is the difference between OLAP and OLTP?
4. Explain the DOLAP architecture in detail.
5. What is the difference between slicing and dicing?
6. What is the difference between ROLAP AND MOLAP?

## 6.11 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Online Transaction Processing | 1. | OLTP |
| | 2. | False |
| | 3. | concurrency and atomicity |
| | 4. | Concurrency |
| | 5. | Atomicity |
| | 6. | network |
| Online Analytical Processing | 7. | d. All of these |
| | 8. | False |
| | 9. | False |
| | 10. | a. ROLAP |
| | 11. | Multidimensional |
| Different OLAP Architectures | 12. | True |
| | 13. | False |
| | 14. | relational data |
| | 15. | MOLAP |
| Comparison between OLTP & OLAP | 16. | OLTP, OLAP |
| | 17. | True |
| OLAP Operations | 18. | d. All of these |
| | 19. | slice operation |
| | 20. | dice operation |

## 6.12 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Anantapantula, S., Gomez, J., Kadur, S. and J, V. (2009). Oracle Essbase 9 implementation guide. Birmingham, U.K.: Packt Pub.
- Ho, C. (1997). Range queries in OLAP data cubes. Yorktown Heights, N.Y.: IBM Research Division.
- Jensen, C., Pedersen, T. and Thomsen, C. (2010). Multidimensional databases and data warehousing. [San Rafael, Calif.?]: Morgan & Claypool Publishers.


### E-REFERENCES


- SearchDataCenter. (2018). what is OLTP (online transaction processing)? - Definition from WhatIs.com. [Online] Available at: https://searchdatacenter.techtarget.com/definition/OLTP
- Datawarehouse4u.info. (2018). OLTP vs. OLAP. [online] Available at: https://www.datawarehouse4u.info/OLTP-vs-OLAP.html [Accessed 6 Dec. 2018].
- Database.guide. (2018). What is OLTP? | Database.Guide. [online] Available at: https://database.guide/what-is-oltp/

CHAPTER 7

# Data Warehousing

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the concept of data warehousing as an information environment
- Discuss the benefits of a data warehousing
- Explore the features of a data warehouse
- Identify the concept of increased demand for strategic information
- Discuss the importance of inability of past decision support system
- Compare the operational and decisional support system
- Describe the concept of information flow mechanism
- Explain the concept of data warehouse and data marts
- Define the significance of data warehouse design techniques
- Describe the concept of ETL process

## 7.1 INTRODUCTION

In the previous chapter, the concept of OLTP and OLAP were discussed in detail. The chapter also covered the needs of OLAP and different OLAP architectures. Thereafter, the chapter explained the differences between OLAP and OLTP. Finally, the chapter covered various OLAP operations such as roll-up, drill-down, slice, dice and pivot.

A data warehouse is maintained by organisations as a central storehouse of data that can be equally accessed by all the business experts and the end-users. The term 'data warehouse' was introduced by W. H. Inmon, a computer scientist also known as the Father of Data Warehouse. Data warehouses are used to store huge amounts of data, which help organisations in decision making, defining business conditions and formulating future strategies.

Both the data warehouse and the database store data but a data warehouse is more efficient than a database. A data warehouse is more effective in dealing with the information requirement of an organisation because it helps in fulfilling the information needs for the management.

Extraction, Transformation and Loading (ETL) is a process of extracting data from the source systems, validating it against certain quality standards, transforming it so that data from separate sources can be used together and delivered in a presentation-ready format and then loading it into the data warehouse. This organised form of data helps organisations as well as end-users to conduct analysis, create reports, formulate strategies and the decision-making process. Apart from the three processes of extraction, transformation and loading, ETL also involves transportation stage in which data is transported from various sources to the warehouse.

This chapter explains the significance of data warehousing and the need to implement it in organisations. The chapter also discusses the importance and the increased

Notes

demand for strategic information among business experts. Further, this chapter elaborates the benefits and features of data warehousing and the data warehouse architecture. Thereafter, you will learn about the information flow mechanism, data warehouse architecture, data marts and data warehouse design techniques. This chapter also describes the various steps involved in the ETL process. This chapter starts with an overview of the ETL process and proceeds to discuss its various stages in detail.

## 7.2 DATA WAREHOUSING: AN INFORMATIONAL ENVIRONMENT

A data warehousing refers to the process of building and using a data warehouse. The data warehouse helps organisations to fulfil their information-related requirements. Therefore, it is also called as an informational environment. Figure 1 shows how we can regard the data warehouse as an informational environment:

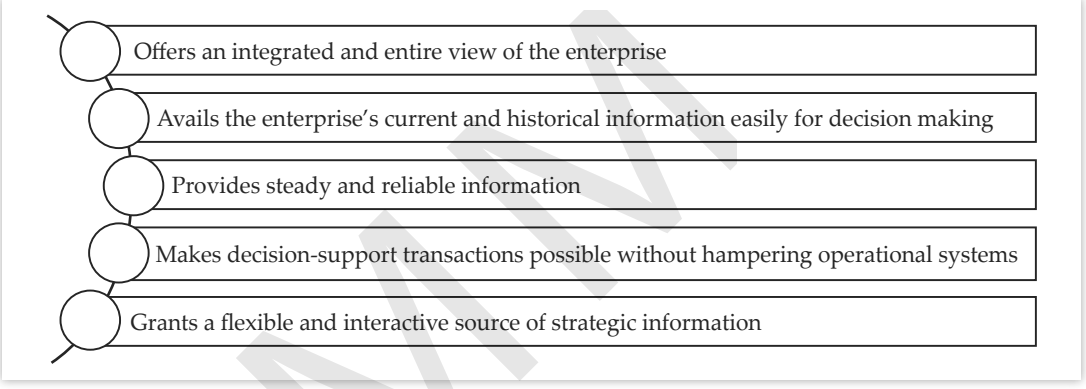



Figure 1: Data Warehouse as an Informational Environment

### 7.2.1 BENEFITS OF DATA WAREHOUSING

A data warehouse stores a replica of information from the source transaction systems. Let us discuss some benefits of implementing data warehousing:

- It collects data from numerous sources into a single database so that a single query engine can be executed to access the data.
- It diminishes the problematic situations of database isolation level lock disagreement in transaction processing systems caused by attempts to run large and complex analysis queries in transaction processing databases.
- It manages data history, even if the source transaction systems do not.
- It integrates data from several source systems, which deliver a central view across the enterprise. This is always a valuable benefit, but specifically when the organisation has been merged with other organisations and grown in size.
- It improves the quality of data, provides reliable codes and descriptions and flags. It even fixes bad data.
- It delivers the organisation's information constantly to business experts and managers.

- It provides a common data model for all data irrespective of the source of the data.
- It reorganises the data to make it easy to understand for business users.
- It restructures the data to make it deliver excellent query performance. It also helps in complex analytic queries, without affecting the operational systems.
- It enhances the value of operational business applications, especially Customer Relationship Management (CRM) systems.
- It makes decision-support queries easier to write.
- Data warehouses make it easy to develop and store metadata.
- Business experts or users become habitual to see many customised data on display screens fields such as rolled-up general ledger balances. These fields do not exist in the database.
- When we perform reporting and analysis functions on the hardware that handles transactions, the performance is often poor. Therefore, the data warehouse should be used for reporting and analysis.

### 7.2.2 FEATURES OF DATA WAREHOUSE

A data warehouse is a combination of data from enterprise-wide sources. A data warehouse consists of the following four main features:

- **Subject oriented:** A data warehouse helps experts or users to analyse data. Consider an example where you want to analyse the company's sales data. For this, you can create a warehouse that focuses on sales only. With the help of this warehouse, you can find the information such as the best customer for a particular item last year, best product and an increase in sales, and so on. This ability to define a data warehouse by a particular subject makes it subject oriented.
- **Integrated:** Integration is associated with subject orientation. Data warehouses must structure the data from various sources into a consistent format. These data warehouses must solve problems such as naming conflicts and inconsistencies among units of measure. Once these problems are solved by the data warehouses, they are regarded as integrated.
- **Non-volatile:** The meaning of non-volatile is that once the data is entered into a data warehouse, it should not be changed or altered. This is because the objective of a data warehouse is to analyse what has occurred.
- **Time Variant:** Business analysts require the large amounts of data to discover trends in business. This is very different from the Online Transaction Processing (OLTP) systems, where the performance requirements request that historical data should be moved to an archive. The term time variant signifies a data warehouse's focus on the change over time. Normally, data flows on a monthly, weekly or daily basis from one or many OLTPs to data warehouse.

### 7.2.3 INCREASED DEMAND FOR STRATEGIC INFORMATION

Strategic information helps business experts to make strategies to achieve business goals. It is very important for an organisation and provides critical information to

Notes

create business goals. Figure 2 shows some business objectives that can be achieved using strategic information:

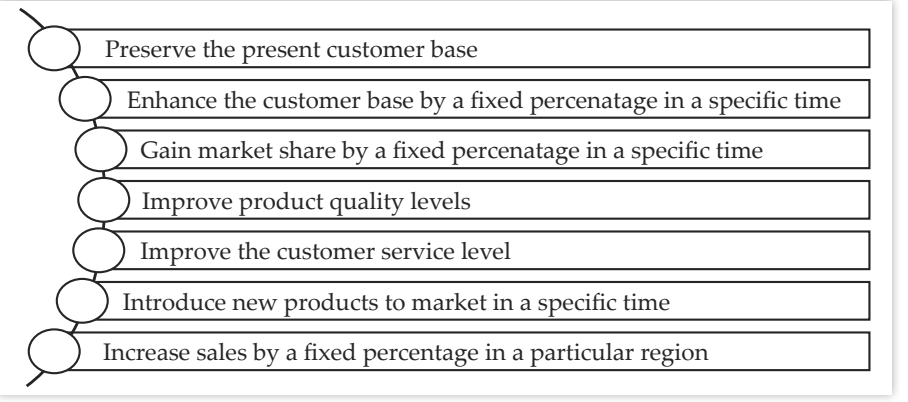



Figure 2: Business Objectives

Business experts and managers use strategic information to make decisions about these objectives for some important purposes. Some of these purposes are as follows:

- Gain a thorough knowledge of the company's operations
- Learn about important business factors and how these affect each other
- Monitor the change in business factors over time
- Compare the performance of the organisations to that of the competitors

Business experts and managers need to concentrate on the need and preferences of customers, new technologies, sales and marketing outcomes and quality of product, and services. There are so many types of information that is needed to make decisions regarding the creation and the execution of business strategies. You can group all these types of essential information and call it strategic information. Strategic information is not for executing daily operations such as generating invoices, making shipments and recording bank transactions. Figure 3 shows the features of strategic information:

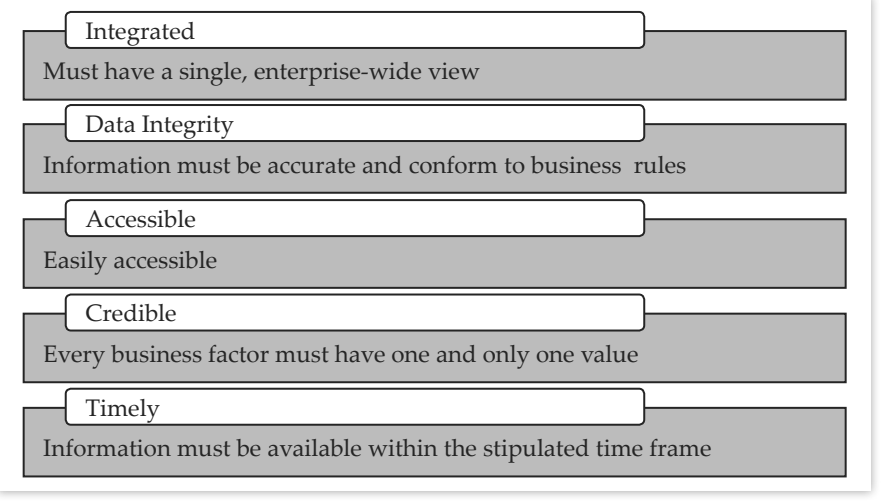



Figure 3: Features of Strategic Information

### 7.2.4 INABILITY OF PAST DECISION SUPPORT SYSTEM

Imagine you are working as an IT expert in a company. The marketing department in your company has been concerned about the performance of a particular region. The sales numbers from the monthly report of this region are considerably low. The marketing Vice President of the company wants to get some reports from the IT department to examine the performance of the organisation and tells you to deliver the same. However, there are no regular reports from any system according to the needs of the beginning.

You might face such situations many times during your career as an IT expert. Sometimes, you might get the information needed for such ad hoc reports from the databases and other sources and sometimes you may not get the required information. In the latter case, you may have to approach several applications, running on different platforms in your company environment to get the information. Sometimes, you may also be required to sort or present the information in different formats and all tasks can prove to be very cumbersome and time consuming in the absence of a data warehouse.

The fact is that for the last couple of decades or more, IT departments have been trying to deliver information to key personnel in their companies for making strategic decisions. Sometimes, an IT department could generate ad hoc reports from a single application but in most cases, the reports are created from multiple systems.

Most of these efforts by IT department in the past resulted in failure. Users often could not clearly define what they need in the first place. Once the first set of reports is delivered to them, they wanted more data in changed formats. This happened primarily because of the nature of the process of making strategic decisions. Information required for strategic decision making has to be available in a collaborative manner. The user must be capable to query online and get results. The information must be in an appropriate format for analysis.

### 7.2.5 OPERATIONAL VS DECISIONAL SUPPORT SYSTEMS

Let us now compare the operational and decision support systems. Table 1 shows the comparison between operational support systems and decisional support systems:

**Table1: Differences between Operational Support Systems and Decisional Support Systems**

| Operational Support Systems | Decisional Support Systems |
|---|---|
| Data represented by the operational support systems is transactions that happen in real-time. | Decisional support systems use the operational data at a particular point in time. |
| Operational data is regarded as update transactions. | Decisional support data is regarded as query transaction, which is read only. |
| The concurrent transaction volume in operational data is very high. | The concurrent transaction volume is found at low or medium levels. |
| Operational data usually consists of information about transactions and is stored in numerous tables. | Decisional support data is usually stored in less number of tables that store data derived from the operational data. |

Notes

Data in decisional support systems needs to be updated periodically to load new updated data that is derived from the operational data. The decision support data does not store the details of each operational transaction. Hence, you can say that decision support systems store data that is summarised, integrated and aggregated for decision support objectives.

### 7.2.6 INFORMATION FLOW MECHANISM

Now, you are going to study the overall flow of data and control through the data warehouse. These flows are metadata-driven and simplified by Java Metadata Interface (JMI) programmatic interfaces (both the metamodel-specific and the reflective) and the XML Metadata Interchange (XMI) bulk import/export mechanism maintained by JMI. Figure 4 shows the data warehouse information flow:

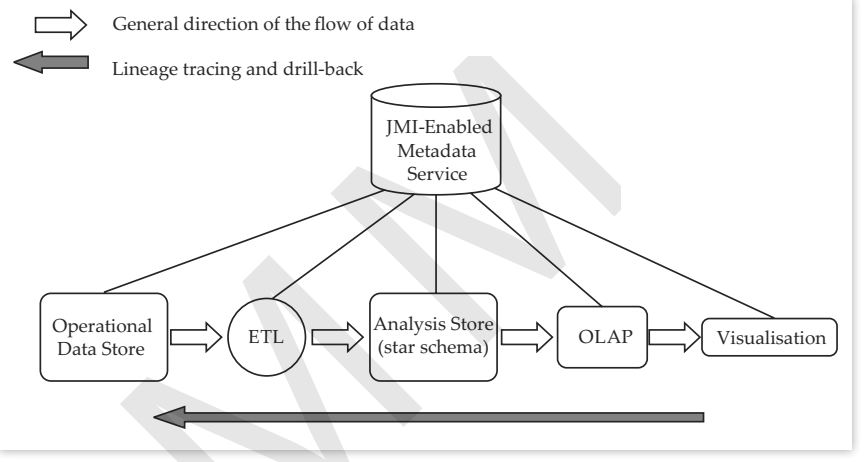



Figure 4: Data Warehouse Information Flow

Figure 4 shows the complete flow of information through the integrated data warehouse. The white arrows denote the general flow of data through the data warehouse, from the Operational Data Source (ODS) to the advanced visualisation/ reporting software. This data flow is characteristically metadata-driven.

Shared metadata is defined by a Meta Object Facility or MOF-compliant metamodel that is Common Warehouse Metamodel (CWM), but the JMI-enabled metadata service is not linked to any particular metamodel, and is capable of loading the CWM metamodel and dynamically generating an internal implementation of CWM Communication of shared metadata is accomplished through JMI interfaces.

Xmi Reader and Xmi Writer interfaces are used to transfer complete models or precise packages of models in a bulk format for loading into tools. On the other hand, metamodel-specific (CWM, in this case) JMI interfaces are used by the client tools for browsing and probably building or altering existing metadata structures. Finally, JMI reflection is used to enable metadata integration between tools whose metamodels vary, but are otherwise MOF-compliant.

**Self-Assessment Questions**

1. A data warehouse stores a replica of information from the source_____ systems.

2. A data warehouse collects data from numerous sources into a single database so that a single _____ engine can be executed to access the data.
3. A data warehouse improves the quality of data, provides reliable codes and descriptions, and flags. It even fixes bad data. (True/False)
4. Data warehouses make it difficult to develop and store metadata. (True/False)
5. Strategic information helps business experts to make strategies to achieve business goals. (True/False)
6. _____ information is not for executing daily operations such as generating invoices, making shipments and recording bank transactions.
7. Which of the following are the features of strategic information?

   a. Integrated
   b. Accessible
   c. Credible
   d. All of these

**Activity**

Assume that you are working as a data analyst in a company. The company has assigned you the responsibility to develop a data warehouse for it. Enlist the steps that you will follow to develop the data warehouse.

## 7.3 KEY COMPONENTS

As discussed earlier, the process of data warehousing allows the collection of data from various sources into a single database so that it can be accessed easily using the query engine. The key components of a data warehousing process are:

- Data Warehouse and Data Marts
- Fact and Dimension Tables
- Architecture of a data warehouse

Let us discuss each component in detail.

### 7.3.1 DATA WAREHOUSE AND DATA MARTS

While studying data warehouses, you must have come across the term data mart. Most people who have a little knowledge about data warehousing use the terms data warehouse and data mart synonymously. Even some authors and vendors use these terms synonymously. Now, let us discuss both the terms.According to many vendors, data warehouses are not easy to build and are also expensive. They make you believe that building data warehouses are just a waste of time. However, this is not accurate. These data mart vendors believe that data warehouses are obstacles which stop them from earning profits. Ordinarily, they would tell you about all the drawbacks of a data warehouse that you may encounter while implementing a data warehouse. Some vendors might suggest you to build a data warehouse by building a few data marts and let them grow.

Notes

However, using this method you might face many problems. Data mart companies try to advertise their products as being the data warehouses. This often confuses people. Many people purchased data marts and started using them without the data warehouses. But soon they realised that the architecture is defective. It should be understood clearly that the data warehouses and the data marts are two different things. There are some noteworthy differences between both of them.

A data mart is a collection of subjects that support departments in making specific decisions. For example, the marketing department will have its own data mart, while the sales department will have a data mart that is separate from it. Additionally, each department completely owns the software, hardware and other components that form their data marts.

Due to this, it is difficult to manage and organise data across various departments. Each department has its own control on data mart that how it looks. The data mart that they use will be explicit to them. Comparatively, a data warehouse is designed around the entire organisation and is not owned by any single department. The data contained in data warehouses is granular, but the information stored in data marts is not very granular.

Another thing that distinguishes data warehouses from data marts is that data warehouses store more information than data marts. The information stored in data marts is normally summarised.

The information that is stored in a data warehouses is mostly historical in nature. Data warehouses are designed to process this information. You have seen that there are many differences between data marts and data warehouses.

Table 2 shows the differences between a data warehouse and a data mart:

**Table 2: Differences between a Data Warehouse and a Data Mart**

| Data Warehouse | Data Mart |
|---|---|
| It has a corporate/ enterprise-wide scope. | Its scope is departmental that is specific to one department. |
| It is a union of all data marts. | It is a single business process. |
| Data is received from the staging area. | Data is received from Star-join (facts and dimensions). |
| It queries on the presentation resources. | It is technology optimal for data access and analysis. |
| It has a structure for corporate view of data. | It has a structure to suit the departmental view of data. |

## 7.3.2 FACT AND DIMENSION TABLES

Every data warehouse includes one or more fact tables. A fact table contains facts related to the business sale revenue such as details sales revenue, details of cash transactions and the transactions related to non-profit organisations. It normally can have large numbers of rows, which may range sometimes up to hundreds of millions of records for large organisations. These records contain one or more years of history of operations of the organisation. One important characteristic of a fact table is that

it contains numerical data (facts) that can be summarised and aggregated to provide information about the history of the operations of the organisation. Each fact table also contains an index that contains as foreign keys the primary keys of the related dimension tables. These dimension tables contain the attributes of the fact records. Fact tables should not contain any kind of descriptive data. It can contain only numerical fields and the index fields that relate the facts to corresponding entries in the dimension tables.

A dimension table, on the other hand, is a hierarchical structure that contains attributes to describe fact records in the fact table. Some of these attributes provide descriptive information, others are used to specify how fact table data should be aggregated or summarised to provide useful information to the analyst. Dimension tables consist of hierarchies of attributes that help in summarisation. For example, a dimension containing product information would often contain a hierarchy that separates products into categories such as food, drink and non-consumable items. Each of these categories is further subdivided into a number of times until an individual product is reached at the lowest level.

Dimension data is usually collected at the lowest level and then aggregated and transformed into higher-level totals, which prove to be more useful for business analysis. These natural progressions or aggregations within a dimension table are called hierarchies. Dimension tables are used to store the data and information that normally contain queries. Dimension tables contain information which is mostly textual and descriptive and can be used in result set as the row headers. Dimension tables are produced by dimensional modelling. Each table contains fact attributes that are independent of those in other dimensions. For example, a customer dimension table contains the data about customers, a product dimension table contains information about products, and a store dimension table contains information about stores. Figure 5 shows a fact table (Fact_Table_Sale) and the associated dimensions tables (Dim_Table_Date, Dim_Table_Store and Dim_Table_Product ):

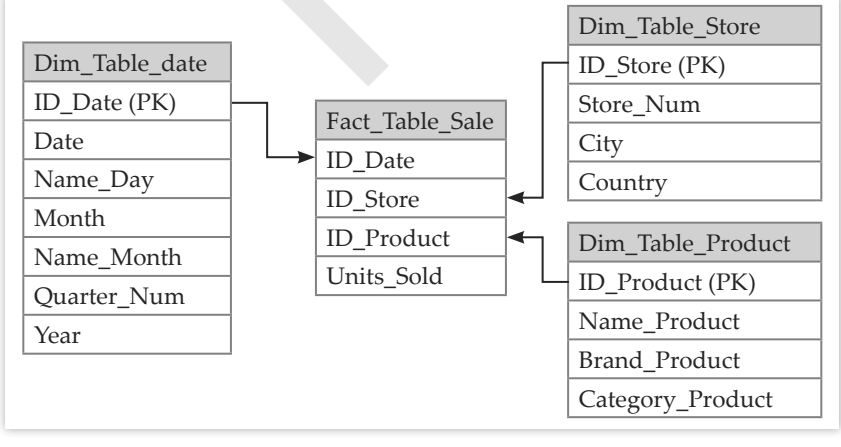



Figure 5: Fact and Dimension Tables

Queries make use of attributes in dimensions to specify a view into the fact information. For example, The Product, store and time dimensions might be used in the query to ask the question "What was the cost of non-consumable goods in the southwest in 1989?" Subsequent queries can go down along one or more dimensions

Notes

to study more detailed data such as "What was the cost of kitchen products in New York City in the third quarter of 1999?" In these examples, the dimension tables are used to specify how a numeric figure (cost) in the fact table is to be summarised.

### 7.3.3 DATA WAREHOUSE ARCHITECTURE

A typical data warehousing architecture is shown in Figure 6:

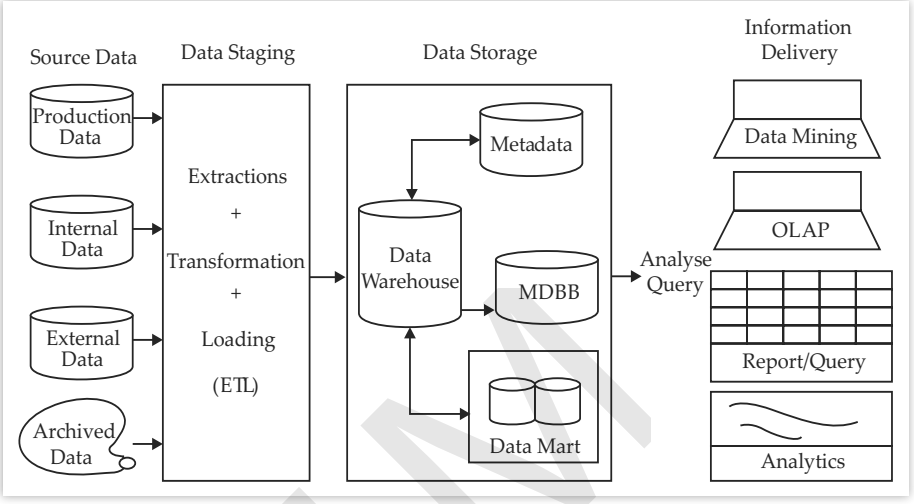



Figure 6: Architecture of Data Warehouse

The followings are the description of the layers of the data warehouse system:

- **Data source layer**: The data can be in any of these formats, i.e., plain text file, relational database, Excel file and other types of databases. The following are the types of data that can act as data sources:
  - **Production sources**: Represents sales data, HR data and product data
  - **Internal data**: Represents data of a department or an organisation such as employee data
  - **External data**: Represents data from outside the organisation or third party data such as census data or demographic or survey data
  - **Archived data**: Represents logs of the Web server along with the user's browsing data
- **Data staging layer:** Refers to the storage area for data processing where data comes before being transformed into the data that is entered in a data warehouse. The followings are the steps involved in transporting data from various sources to data warehouse:
  - **Extraction:** Refers to the process of extracting data from different source systems and validating it against certain quality
  - **Transformation:** Refers to the transformation of the data available indifferent source systems and validating it against certain quality
  - **Loading:** Refers to the process of loading the data either from data warehouse or data mart

- **Data storage layer:** Refers to the layer in which the transformed data and cleaned data is stored. On the basis of the scope and functionality, the following are the types of entities in this stage:
  - **Data warehouse:** It is maintained by organisations as central warehouse of data that can be equally accessed by all business experts and end users.
  - **Data mart**: When data warehouse is created at the departmental level, it is known as data mart.
  - **Metadata**: Details about the data is known as metadata. In other words, it is a catalogue of data warehouse.
  - **MDDB**: It is multidimensional database that allows data to be moulded and viewed in multiple dimensions.
- **Information delivery**: Provides the information that reached to end-users. The information can be in any form such as tables, chart, graphs, or histograms. The following are the tools used in this layer:
  - **Data mining:** Refers to the process of finding the relevant and useful information from a large amount of data
  - **OLAP:** Allows the navigation of data at different levels of abstraction such as drill-down, roll-up, slice, dice, and so on
  - **Query/reports:** Query and reporting tools are used for accessing and displaying data stored in the data warehouse to a user. A user enters the query and accordingly information is displayed mainly in form of reports
  - **Analytics:** Data is dynamically accessed from the data warehouse using various analytical tools efficiently. The data accessed can be used for analysis by end-users and decision making in an organisation

**Self-Assessment Questions**

8. A _____ is a collection of subjects that support departments in making specific decisions.
9. The information that is stored in a data warehouses is mostly historical innature. (True/False)
10. A _____ table is a special table that contains the data to measures the organisation's business operations.
11. _____ Layer refers to the layer representing various data sources that enter data into the data warehouse.
12. Detail about the data is known as _____.

**Activity**

Search and explore about the classification of metadata of a data warehouse.

## 7.4 DATA WAREHOUSE DESIGN TECHNIQUES

A data warehouse can be designed either by bottom-up or top-down approach. In other words, a global data warehouse contains data of the entire organisation and

segments it into different data marts or subset warehouse. Let us discuss both the bottom-up design and the top-down design approaches.

### 7.4.1 BOTTOM-UP DESIGN

Data marts in the bottom-up approach firstly help to create reports and then analytical capabilities for specific business process. Data marts store dimensions and facts. Facts comprise either atomic data and, if necessary, summarised data. A single data mart usually models a particular business area such as 'Sales' or 'Production.' These data marts can ultimately be integrated to create a complete data warehouse. The data warehouse bus architecture is mainly an implementation of 'the bus', a group of conformed dimensions and conformed facts, which are dimensions that are shared between facts in multiple data marts. Figure 7 shows the bottom-up design approach for designing a data warehouse:

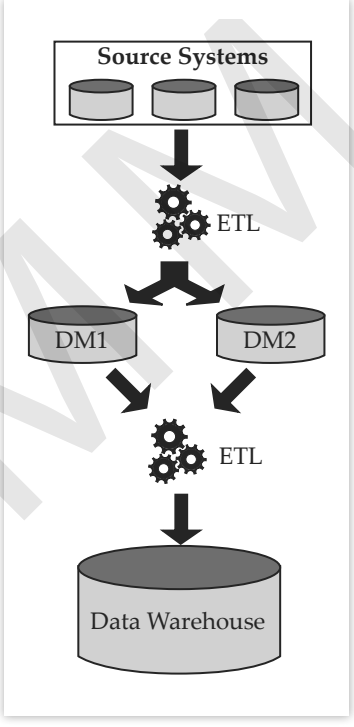



Figure 7: Bottom-up Design Approach

**Source:** *http://dwgeek.com/various-data-warehouse-design-approaches.html/*

The integration of data marts in a data warehouse is focused on the conformed dimensions that reside in the bus that describes the potential integration 'points' between data marts. The authentic integration of two or more data marts is then completed by a process recognised as 'drill across'. A drill across works by grouping or summarising data along the keys of the conformed dimensions of each fact contributing in the "drill across" followed by a join on the keys of these summarised facts.

Upholding strict management over the data warehouse bus architecture is essential to maintain the integrity of the data warehouse. The most significant management task is to ensure the dimensions among data marts that are reliable.

Business value can be returned as soon as the first data marts can be for example, the data warehousing might start in the 'Sales' department, by creating a Sales data mart. After the completion of the Sales-data mart, the business might choose to take the warehousing activities into the 'Production department' which results in a Production data mart.

The need for the Sales data mart and the Production data mart to be integrable is that they share the same 'Bus', means that the data warehousing team has done something to recognise and implement the conformed dimensions in the bus, and that separate data mart links that information from the bus. The Sales-datamart is already constructed and the Production-data mart can be constructed virtually independent of the Sales-data mart (but not independent of the Bus).

## 7.4.2 TOP-DOWN DESIGN

In the top-down approach, first the data warehouse is designed and then datamarts are created unlike bottom-up approach. The top-down approach is designed with the help of the normalised enterprise data model. 'Atomic' data, which is the lowest level of detail, is stored in the data warehouse.

Dimensional data marts that store data needed for specific business processes or specific departments are built from the data warehouse. In the **Inmon** vision, the data warehouse is in the middle of the 'Corporate Information Factory' (CIF), which provides a logical framework for providing Business Intelligence (BI) and business management skills.

Figure 8 shows the top-down design approach for designing a data warehouse:

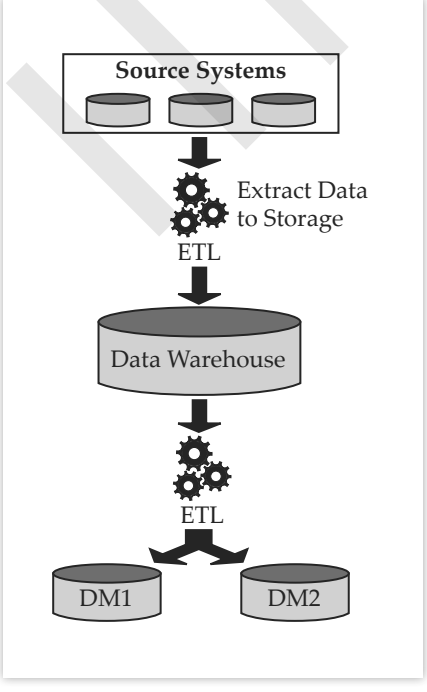



FIGURE 8: Top-down Approach

**Source:** *http://dwgeek.com/various-data-warehouse-design-approaches.html/*

Notes

The differences between the bottom-up approach and the top-down approach for designing a data warehouse are given in Table 3:

**Table 3: Differences between Bottom-Up Approach and Top-Down Approach**

| Points of Comparison | Bottom-Up Approach | Top-Down Approach |
|---|---|---|
| Building of data warehouse | This approach takes lesser time in building a data warehouse. | This approach is time consuming. |
| Cost | This approach requires low development cost initially but subsequent cost of development or changes is high. | This approach requires high development cost initially but subsequent cost of development or changes is low. |
| Time | This approach requires a lesser amount of time for an initial setup. | This approach requires a higher amount of time for an initial setup. |
| Maintenance | The maintenance in this approach is difficult. | The maintenance in this approach is easy. |
| Skills Required | This approach does not require highly-skilled people for designing. | This approach requires highly-skilled people for designing. |

**Self-Assessment Questions**

13. A data warehouse can be designed either by _____ or _____ approach.
14. The authentic integration of two or more data marts is then completedby a process recognised as 'drill across'. (True/False)
15. The _____ approach is designed with the help of the normalised enterprise data model.

## 7.5 ETL PROCESS

Data warehouse needs to be loaded regularly so that it can serve its purpose of providing relevant data to facilitate business analysis. To facilitate this, data from one or more operational systems needs to be extracted and copied into the data warehouse.

The integration, rearrangement and consolidation of a large amount of data is a challenge in the data warehouse environment over many systems in order to provide an organised information for business intelligence.

The three processes, extraction, transformation and loading, are responsible for the majority of operations taking place at the back end of data warehousing. Although, ETL is primarily a back-end process, it takes up almost 70% of the resources required for maintenance and implementation of a data warehouse.

First, the data extraction takes place from multiple sources which typically can range from relational databases, Online Transactional Processing (OLTP), Webpages, or various kinds of documents such as Word documents or spreadsheets.

After the extraction phase, comes the transformation phase wherein all the extracted data is accumulated at a special area called Data Staging Area (DSA).The homogenisation, transformation and cleaning of data take place at the DSA.

This transformation takes place with checks in place such as filters and integrity constraints to ensure that the data that reaches the warehouse conforms to the business rules and schema of the target data warehouse. In the final step, data isloaded in the data warehouse. ETL provides a well-defined process for extracting data from varied sources and loading it in the data warehouse in a consolidated format.

Figure 9 shows the major steps of the ETL process:

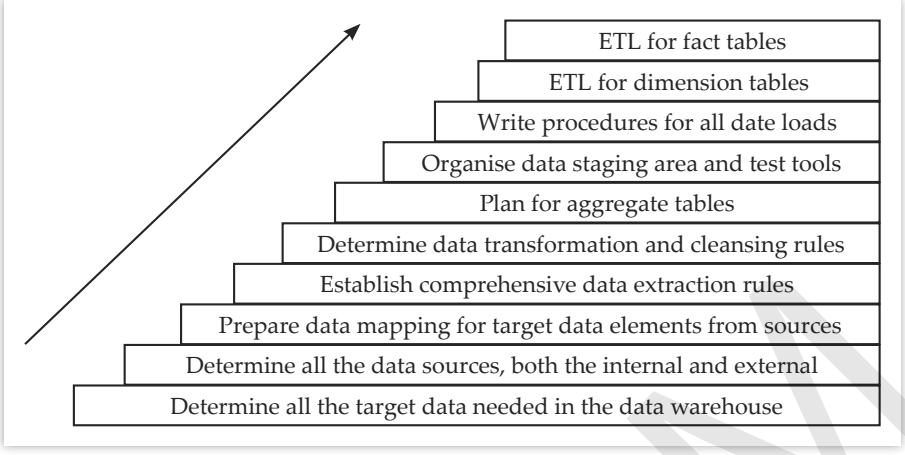



FIGURE 9: Major Steps in the ETL Process

EXHIBIT

**GOOGLE CLOUD USING ETL ARCHITECTURE**

Google Cloud supports ETL architecture for a cloud-native data warehouse on Google Cloud platform. ETL solutions automate the task of extracting data from operational databases, making initial transformations to data, loading data records into staging tables and initiating aggregation calculations. A lot of ETL tools are now capable in handling a very large amount of data that do not have to be necessarily stored in any data warehouses. With Hadoop-connectors to big data sources being provided by almost 40% of ETL tools, support for big data is growing continually at a fast pace.

### 7.5.1 DATA EXTRACTION

Data extraction is the first step in the ETL process. During this phase, first, the required data is identified and then extracted from varied sources source database systems and applications using as little resources as possible. The process of data extraction should not adversely affect the source in terms of its execution, reaction time or any kind of locking. Most of the times, it is not possible to identify the data of interest, thus during the data extraction stage, a lot of data gets extracted than is actually required. The size of the extracted data can range from hundreds of kilobytes up to gigabytes, depending on the source system and the business situation. Depending upon the capabilities of the source system, some transformation might take place during the extraction process itself.

To design and create an extraction process is the most time-consuming part of the entire ETL process. The source systems are diverse and varied in design. They are

complex and are usually poorly documented, thus extracting useful data from them can be a challenging part. Moreover, the data is extracted not once but periodically to update the data in a data warehouse. The extraction process has no control over the source system, nor on its performance or availability to suit the needs of data warehouse extraction method.

There are various techniques to extract data from different kinds of sources. The extraction process that we choose is influenced by various factors such as data source system, transportation process and the time needed to refresh the data warehouse.

## 7.5.2 IDENTIFICATION OF DATA SOURCE

The first and foremost task of the data extraction stage is identification of all the suitable data sources. This process is not only used to identify the data source but also enhances the importance of data warehouse in terms of the data extracted from data sources. Let us discuss the data identification process in detail.

Let us assume that an organisation designs a database to provide strategic information on the orders that it fulfilled. To do that, it needs the records of previous as well as current fulfilled and pending orders. Now, if the orders are fulfilled through multiple channels, then the organisation also needs reports about these channels. The order fact table contains data related to order such as date of delivery, item no, item codes, discounts and credit limit.

The dimension table contains the details about products, customers and channels. The organisation also needs to ensure that it has the correct data source needed for the database and this data source is able to supply correct data to each data element. This is done by going through the verification process to authenticate the data source. You need to go through the source identification and ensure that whatever bit of data is entered into the data warehouse must be authenticated. Figure 10 describes pictorially the example you have just explained:

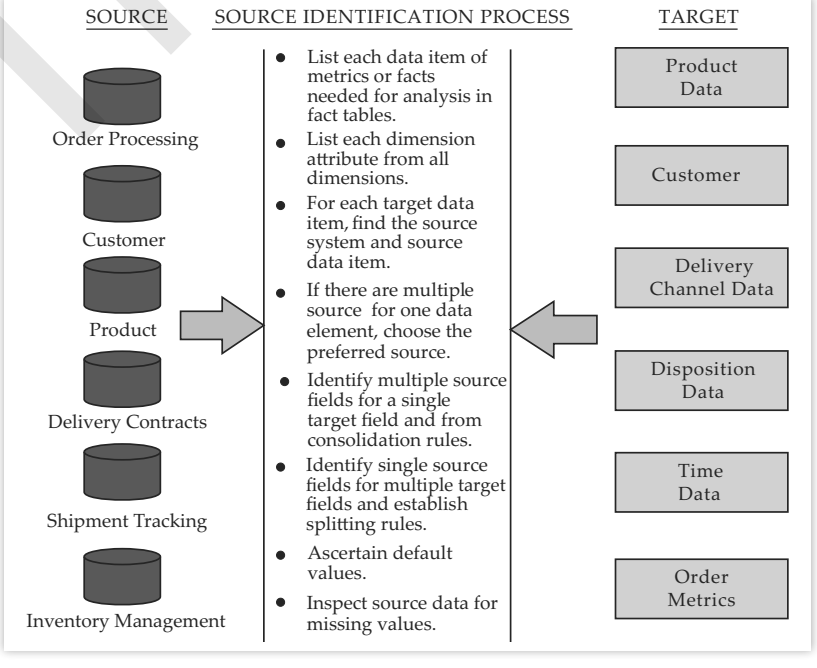



Figure 10: Data Source Identification Process

## 7.5.3 EXTRACTION METHODS IN DATA WAREHOUSE

Before starting to understand the various methods of data extraction, you must understand the nature and source of data that you are trying to capture. You also need to know how the captured data is going to be used. This is necessary because the data source is in a constant state of updating.

Whenever a new addition or modification takes place in the existing data, the data source changes. Thus, data in a system is said to be time dependent or temporal since the data in system changes with time. Let us discuss a few methods of data extraction:

- Logical Extraction Methods
- Physical Extraction Methods

### Logical Extraction Methods

There are two types of logical extraction:

- **Full extraction:** In this case, the data is extracted completely from the source system. Since this data extraction method reflects all the data that is available in the data source as it is, there is no need to keep additional information about the data, i.e., time stamp, that is, when was the last change made, and so on.
- **Incremental extraction:** In this method, only the data that has changed after a specific point of time is extracted. The extraction may be defined by an event that had occurred in the history. This event may define the last time of extraction, thus the data that has changed after this event has occurred is identified. This change in data is either provided by the source data itself such as a data column showing the last change or a separate table where any addition in the data information keeps getting recorded. Since the incremental method entails an additional logic of maintaining a separate table, many data warehouses do not want to use this extraction process. Instead the whole table from the source system is extracted to the data warehouse or data staging area and compared with the previous data from the source data to identify the changes. Although this approach might prove to be simpler for the source data but it clearly places a huge burden on the data warehouse processes especially in cases where the data volume is too high. Incremental data extraction can be done in two ways:
  - **Immediate data extraction:** In this technique, the data extraction is done in real time. Extraction occurs at the same time as the transactions are taking place at the source database and files, as shown in Figure 11:

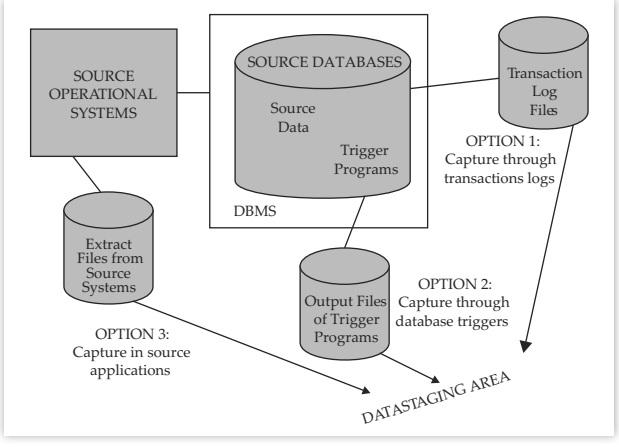



FIGURE 11: Immediate Data Extraction Process

Notes

Immediate data extraction is facilitated by the following three processes:

- **Capture through transaction Log:** This option takes the help of transaction logs of the DBMS, which are maintained for data recovery in case of failure. For every transaction such as addition, updation, or deletion of a row from a database table, the DBMS immediately makes an entry in the transaction log file. This transaction log is read by the data extraction technique and selects all the committed transactions.
- **Capture through database triggers:** This option is applicable to all the database applications. As you know, triggers are specially stored procedures that are stored on the database and are fired when certain predefined events occur. You can create triggers for all the possible events for which you need data to be captured. When a trigger is fired, it generates an output that is stored in a separate file. This file is used to extract data for the data warehouse. For example, if you need to capture all changes to the records in the employee table, write a trigger program to capture all updates and deletes in that table.
- **Capture through source applications:** This option is also referred to as application-assisted data capture. Here, the source application itself, assists in the data capture for the data warehouse. You have to accordingly modify the relevant application programs that write to the source files and databases. You rewrite the programs to include all additions, updates, and deletes to the source files as well as database tables. The changes to the source data can be contained in the separate files by other extract program.

❍ **Deferred data extraction:** In this technique, as compared to immediate data extraction, data extraction does not capture the changes in real time as shown in Figure 12:

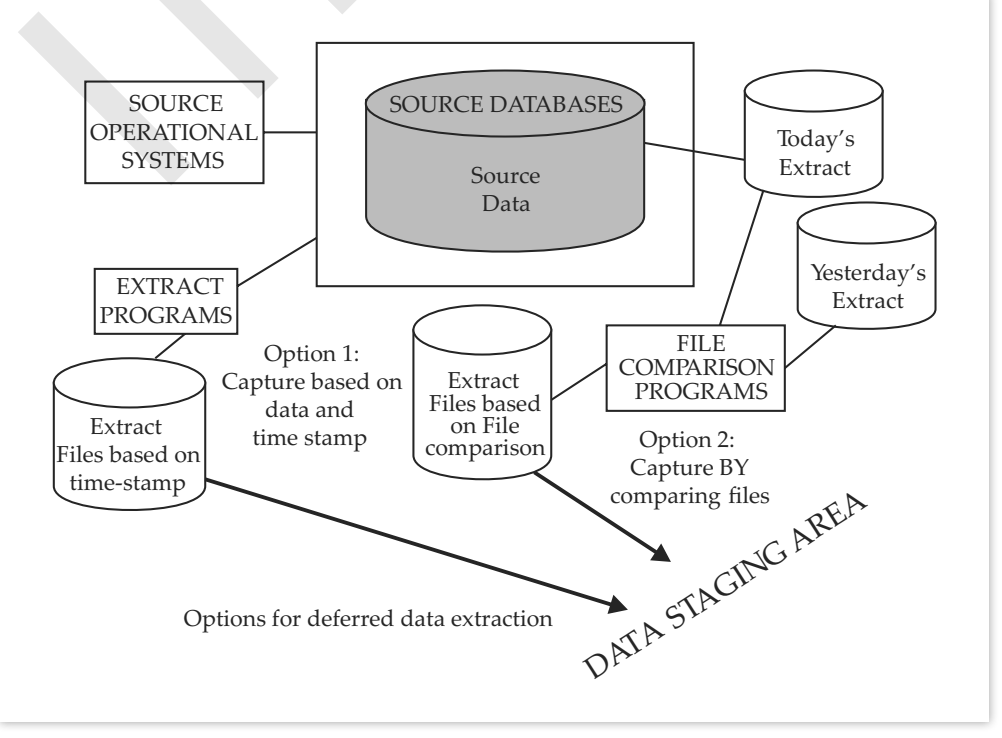



Figure 12: Deferred Data Extraction Process

Deferred data extraction takes place in the following two ways:

- **Capture based on date and time stamp:** In this case, every time a source record is created or updated, it must be marked with a stamp showing the date and time of creation or updation. This timestamp provides the basis for selecting records for data extraction. Here, the data capture occurs at a later time and not while each source record is created or updated. This technique works efficiently for the system where the number of revised records is small. However, deleting source records can cause problems. This problem is rectified by marking the source record for delete first, do the extraction run and then go ahead and physically delete the record.
- **Capture by comparing files:** This technique is also called the snap shot differential technique because it compares two snapshots of the source data. In this technique, it is necessary to keep prior copies of all the relevant source data as they may be required for comparing data in future. Though the technique is simple and straight forward, the actual comparison of full rows in a large file can be very time consuming and may prove in efficient in the long run. The technique may be only relevant way for the source records of some legacy data sources are used that do not have transaction logs and time stamps.

### Physical Extraction Methods

Depending on the logical method adopted for the extraction of data and the limitations and capabilities of the source data, two physical extraction methods are applied on the extracted data. The online source system or offline structure can be used to extract the data. This offline structure may already exist or it is generated by an extraction routine. Followings are the two mechanisms of physical extraction:

- **Online extraction:** In this scenario, the data is extracted from the source directly. The extraction process connects directly to data source and extracts data from the source itself. It can also connect to with an intermediate system that stores the data in a recon figured manner (i.e., snapshot logs or change tables). The point to notice here is that the intermediate system is not necessarily physically different from the source system. You need to consider the original source objects of prepared source objects for the distributed transactions in an online transaction.
- **Offline extraction:** In contrast to online extraction, data is extracted from outside the original system, not directly from the sources. The data is already in a pre-existing format or structure (i.e., redo logs, archive logs or transportable table spaces) or had been created by some extraction routine. A few of these external structures are:
  - **Flat files:** In flat files, data is in a defined, generic format. Some additional information is required about the source object for further processing.
  - **Dump files:** Dump files are created in the Oracle-specific format. Depending on the incorporated utility, the information about the containing objects may or may not be included.

- **Redo and archive logs:** These logs contain information in a specific additional dump file.
- **Transportable table spaces:** Transportable table spaces are powerful medium to extract and move large volumes of data between Oracle databases.

### 7.5.4 CHANGE DATA CAPTURE

An important factor to be considered while data extraction is incremental extraction also called Change Data Capture. If a data warehouse extracts the data from an operational system on a regular basis (i.e., within a scheduled cycle of 24 hours.) then it requires only that data which has been modified since the last extraction (i.e., the data that has been modified in the past 24 hours).

Due to efficient identification and extraction of only the most recently changed data, the extraction process (as well as all downstream operations in the ETL process) becomes much more efficient, because now it must extract a much smaller volume of data.

On the other side, for some source systems, identifying the recently modified data can be difficult or intrusive to the operations of the system. This proves to be a disadvantage for the efficiency and the speed of the system. In data extraction the change in the data capture is one of the most demanding issues. The following are some alternate techniques for implementing a self-developed change capture on Oracle Database source systems:

- **Timestamps:** Some operational systems have specific timestamp columns in their tables. This timestamp specifies the time and date when the specified row was last modified. This enables the identification of the latest data very easily by using the timestamp columns and reduces the overheads of extracting extra data. For example, the following query proves useful in extracting today's data from an orders table:

  SELECT * FROM orders WHERE TRUNC (CAST(order_date AS date),'dd') =TO_DATE(SYSDATE,'dd-mm- yyyy');

  If originally, the timestamp column is not present in an operational source system, it proves to be a difficult task to modify the system to include timestamps. Such type of modification would require changing the operational system table's design to include a new timestamp column and then updating the timestamp column with the help of a trigger which would be fired following every operation that modifies a given row.

- **Partitioning:** Some source systems use range partitioning, i.e., the source tables are partitioned along a date key. This helps in easy identification of new data. For instance, if data extraction is requires from an orders table which is partitioned by week then the data of current week is easily identifiable.

- **Triggers:** Triggers are created to keep track of the recently updated records in an operational system. Timestamp columns can also be used along with triggers to identify the actual time and date of the last modified given row. This can be done by creating a trigger on each source table where change data capture has been

implemented. Thus, for every DML statement that is executed on the source table, the trigger shall update the timestamp column with the current time. Hence, with the help of the timestamp column which provides the exact time and date when a given row was last modified, you can extract the required data.

This kind of technique is used for Oracle materialised view logs. These logs are used by materialised views to identify changed data. These logs are accessible to end users also. However, the format of the materialised view logs is not documented and might change over time.

These techniques are defined by the characteristics of the source systems. Some source systems might require some modifications to implement these techniques. Each one of such techniques should be assessed carefully by the owner of source system preceding the implementation.

All these techniques can be implemented along with the techniques of previously discussed data extraction. When the data contained in the file is being unloaded or data is used through a distributed query then the timestamps can be used. Materialised view logs are completely based on triggers, but this proves to be an advantage as the creation and maintenance of the change-data system is the database. Data extraction can be done in two ways:

- **Extraction using data files:** Most database systems provide methods for exporting or unloading data from the internal database format into flat files. While COBOL programs are used by mainframe systems, many databases as well as third-party software vendors provide export or unload utilities for data extraction. Data extraction does not inevitably involve the whole structure of the database being unloaded in flat files. Although in some cases, it may be suitable to unload entire database tables or objects but in other scenarios it may be more acceptable and beneficial to the given table such as any modifications in the source system since the last extraction or the results of joining multiple tables together. Different extraction techniques are implemented differently according to their capabilities to support these two scenarios. If the source system is an Oracle database, the following options are available for extracting data into files:
  - Extracting into Flat Files using SQL*Plus
  - Extracting into Flat Files using OCI or Pro*C Programs
  - Exporting into Export Files using the Export Utility
  - Extracting into Export Files using External Tables
- **Extraction through distributed operations:** In the distributed-query technology, one database can directly query tables located in various source systems. Specifically, a data warehouse or staging database can directly access tables and data located in a connected source system. Gateways are a form of distributed-query technology. They allow an Oracle database (such as a data warehouse) to access database tables stored in remote, non-Oracle databases. It is one of the most effortless approaches for transferring data among two oracle databases because it merges the transformation along with extraction into single step and requires minimum programming. However, this is not always feasible.

NOTES

## 7.5.5 TRANSFORMATION

Data transformation is the second step in the ETL process. It can include both the simple data conversions and the extremely complex data scrubbing (cleaning/error correction) techniques at the same time. Some data transformations can occur within the database, although most transformations are implemented outside the database, for example on flat files. The following headings will demonstrate the types of fundamental technology that can be applied to implement the transformations.

### Transformation Flow

From an architectural perspective, data can be transformed in two ways:

- **Multistage data transformation:** Data transformation logic consists of multiple steps for most data warehouses. For example, while inserting new records into a table, the transformation may take place in separate logical transformation steps to validate each dimension key. Figure 13 offers a graphical way of looking at the transformation logic:

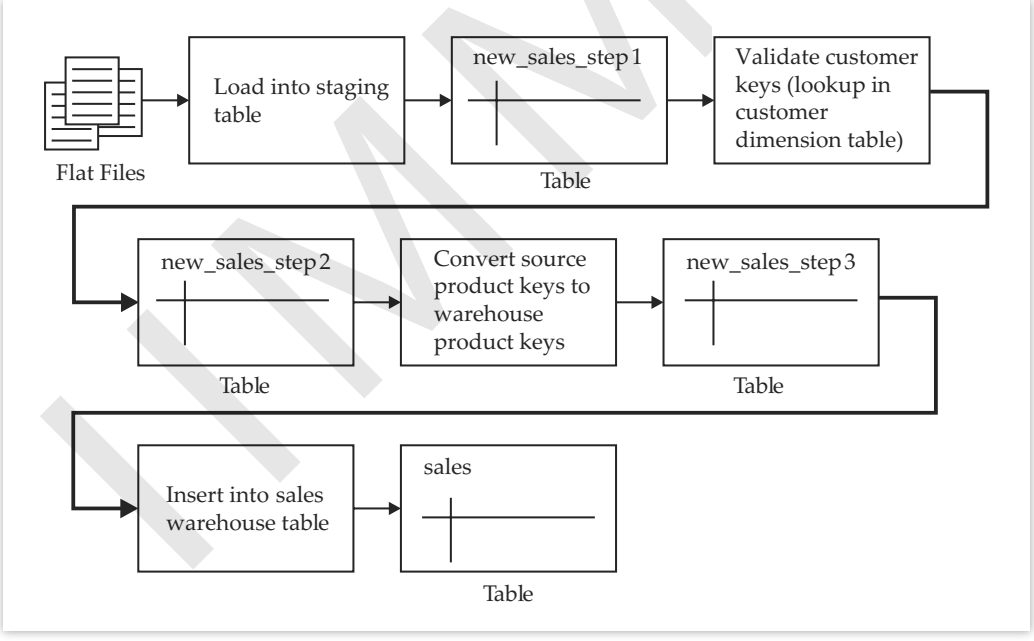



FIGURE 13: Multistage Data Transformation

If the Oracle database is used as a transformation engine, a common technique would be to implement each transformation as a separate SQL operation. After each such operation, a separate, temporary staging table is created (such as new_sales_step1 and new_sales_step2 as in Figure 13) to store the intermediate results for each step.

This load-then-transform technique provides a natural checkpoint scheme to the entire transformation process, thus enabling the process to be more easily monitored and restarted. However, due to this, there is an increase in time and space requirement that proves to be a major disadvantage of multi-staging data transformation.

To overcome this disadvantage, there is a possibility of combining many simple logical transformations into a single SQL statement or single PL/SQL procedure. Although combining steps may prove to optimise the performance, on the other hand, it may also introduce difficulties such as modification, addition, or dropping individual transformations or difficulty in recovering from failed transformations.

- **Pipelined data transformation:** Pipelined data transformation technique changes the ETL process flow dramatically. It renders some of the previous necessary process steps obsolete whereas a few others are remodelled to enhance the data flow. This enhances the whole process and creates a more scalable and non-interruptive data transformation procedure. The focus shifts from serial transform-then-load process (where most of the tasks are done outside the database) or load-then-transform process, to an enhanced transform-while-loading. Figure 14 shows the pipelined data transformation:

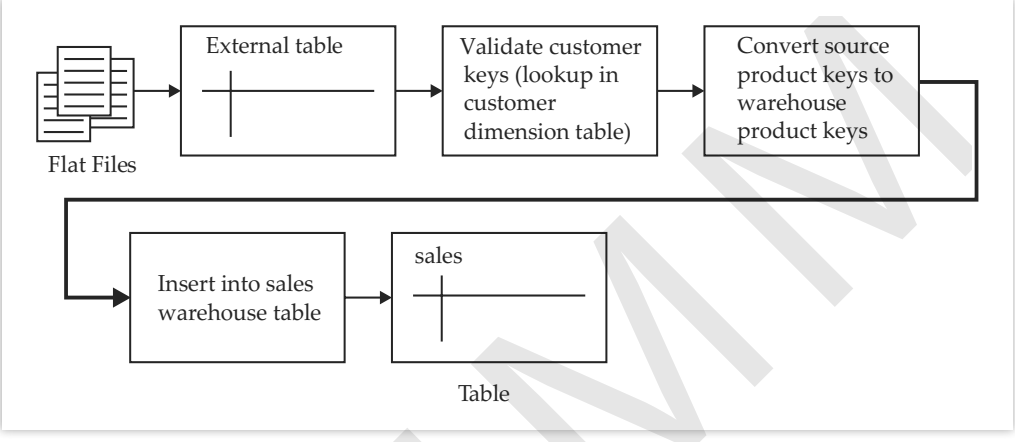



Figure 14: Pipelined Data Transformation

Figure 14 depicts the steps in a pipelined data transformation:

- Loading external table with flat files
- Validating customer keys (lookup in customer dimension table)
- Converting source product keys into warehouse product keys
- Inserting source product keys to warehouse product keys

### 7.5.6 STAGING

During the ETL process, it should be a possibility to restart, if need, some of the phases independently from the others. For example, if the transformation process fails, the process should not restart from the Extract step again. Only the failed step should be rectified and restarted. This can be ensured by implementing proper staging. Staging means that the data is dumped to a special location called the Data warehousing Staging Area (DSA) so that it is unaffected by any failed step. It can be read by the next processing phase individually. This area is also used frequently during the ETL process to store intermediate results of processing. However, the load ETL process can only access the staging area. It should not be made available

NOTES

to anyone outside the process, such as the end-users as it is not intended for data presentation to end-users, as shown in Figure 15:

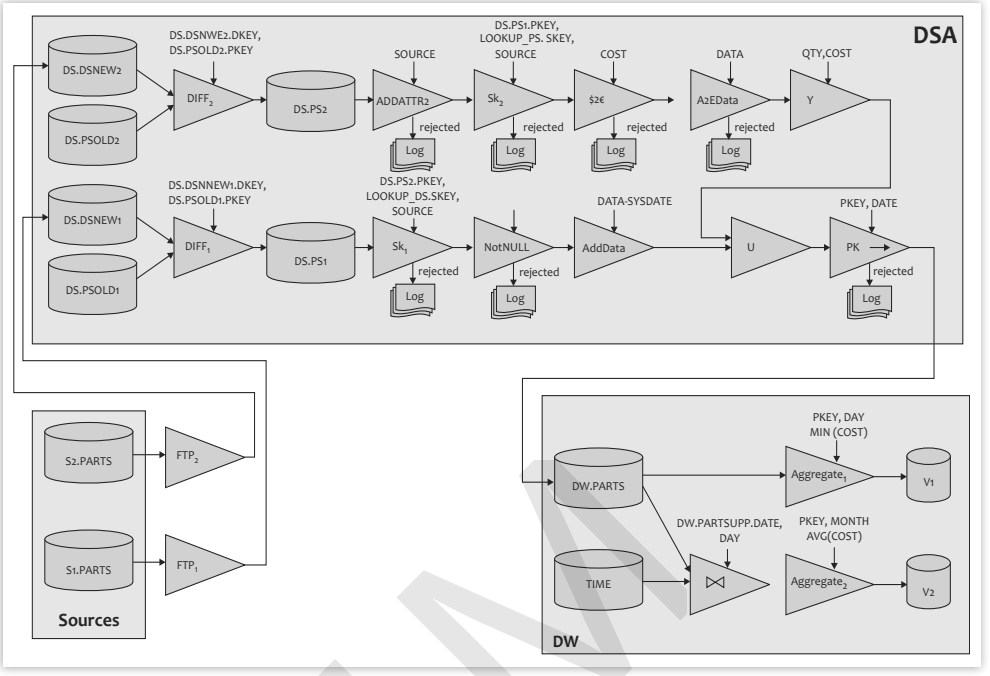



FIGURE 15: Staging Process

The Data Warehousing Architecture requires DSA for timing reasons. It is not possible to extract all the data from all operation databases due to couple of reasons – lack of hardware and network resources, geographical factors, varying business and data processing cycles. In short, all required data must be accumulated and made available to the process before data can be integrated into the data warehouse.

For example, on daily basis, the extraction of sales data might be reasonable, however, the daily extraction of financial data might not be suitable that requires a month-end reconciliation process. Similarly, it might be feasible to extract "customer" data from a database in Bangkok at noon eastern standard time, but this would not be feasible for "customer" data in a New York database.

Data in the data warehouse can be either persistent (i.e., remains in the DSA for a long period) or transient (i.e., only remains in DSA temporarily or for a very short period). All businesses do not require a DSA. Many businesses find it more feasible to use ETL to copy data directly from operational databases into the data warehouse rather than maintaining the Staging Area.

### 7.5.7 LOADING

Loading is the third step in ETL process, and is relatively simpler than the other two processes. During the loading process, it is ensured that the loading of data is completed correctly and with as little resources as possible. To increase the efficiency of the load process, it is desirable to disable any constraints and indexes before starting the load process and enable them back only after the load process completes. To ensure consistency, the referential integrity is maintained by the ETL tool.

The following mechanisms are used for loading a data warehouse:

- Loading a Data Warehouse with SQL*Loader
- Loading a Data Warehouse with External Tables
- Loading a Data Warehouse with OCI and Direct-Path APIs
- Loading a Data Warehouse with Export/Import

### 7.5.8 CLEANING

The cleaning step is one of the most important steps as it ensures the quality of the data in the data warehouse and helps in the integration of heterogeneous data sources. Data quality problems arise in single data collections as well as in case of multiple data sources. In single data source, such as files and databases, the problem of data cleaning arises due to misspellings during the time of entering the data, missing of useful data and other invalid data. When multiple data sources need to be integrated, i.e., in data warehouses, federated database systems or global Web-based information systems, the need for data cleaning increases manifold. In this case, data quality problem arises because the sources often contain redundant data in different forms. In order to provide access to clean, accurate and consistent data, consolidation of different data forms and elimination of duplicate data becomes necessary.

As you know, data is important for business to make critical business decisions. ETL testing plays a significant role in validating and ensuring that the business information is exact, consistent and reliable. It also helps in minimising the hazard of data loss in production. However, there are some challenges come during the ETL testing of data. Some of these challenges are shown in Figure 16:

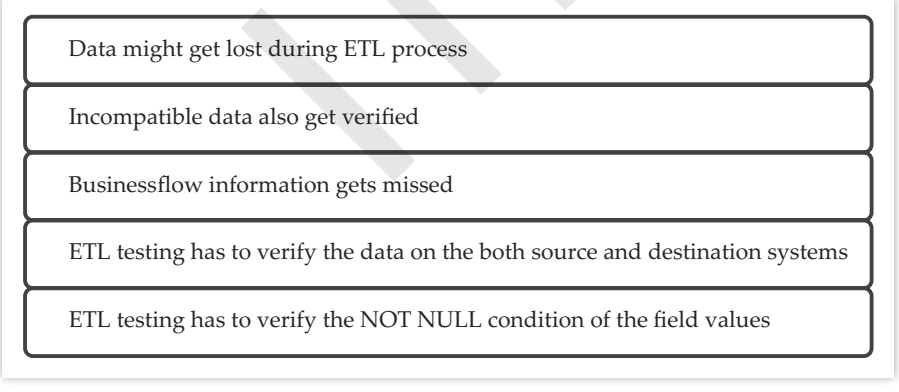



Figure 16: Challenges in ETL Testing

**Self-Assessment Questions**

16. _____ is primarily a back-end process, it takes up almost 70% of the resources required for maintenance and implementation of a data warehouse.
17. The three processes _____, _____, and _____ are responsible for the majority of operations taking place at the back-end of data warehousing.
18. Data extraction is the last step in the ETL process. (True/False)

Notes

19. In _____ extraction method, the data is extracted completely from the source system.
20. The _____ extraction method is time stamped.

## 7.6 SUMMARY

- A data warehouse can be defined as the centralised repository of the data integrating from various sources in order to support analytical techniques and decision-making processes of organisations.
- Data warehouse helps organisation to fulfil their information-related requirements. Therefore, it is also called as an information environment.
- Data warehouse helps you to analyse data.
- Strategic information is not for executing daily operations such as generating invoices, making shipments and recording bank transactions. Strategic information is much more important and helps organisations to take some of its most crucial decisions.
- Decision Support System (DSS) use data extracted from business events and transaction summaries, whereas Operational Support System (OSS) use data that is a detailed record of daily business operations of an organisation.
- Data in decisional support systems needs to be updated periodically to load new updated data that is derived from the operational data.
- A data warehouse has a structure which is distinct from a data mart.
- A data mart is a collection of subjects that support departments in making specific decisions.
- A data warehouse is designed around the entire organisation and is not owned by any single department.
- A fact table is a special table that contains the data to measures the organisation's business operations. Every data warehouse includes one or more fact tables.
- A dimension table, on the other hand, is a hierarchical structure that contains attributes to describe fact records in the fact table.
- Dimension tables are produced by dimensional modelling.
- Data mining refers to the process of finding relevant and useful information from a large amount of data.
- Data marts in the bottom-up approach firstly help to create reports and then analytical capabilities for specific business process.
- Upholding strict management over the data warehouse bus architecture is essential to maintain the integrity of the data warehouse.
- The top-down approach is designed with the help of the normalised enterprise data model.
- Dimensional data marts that store data needed for specific business processes or specific departments are built from the data warehouse.

- ETL is the process of extracting data from varied source systems and loading it into the data warehouse.
- After the extraction phase, comes the transformation phase wherein all the extracted data is accumulated at a special area called Data Staging Area (DSA).
- The first and foremost task of the data extraction stage is identification of all the suitable data sources.
- Identification of data source is a crucial step in the data extraction process.

## 7.7 KEY WORDS

- **Data warehousing:** It is the best way to integrate valuable data from different sources into the database of a particular application.
- **Strategic information:** It refers to information that helps business experts to make strategies to achieve business goals.
- **Data mart:** It is a collection of subjects that support departments in making specific decisions.
- **Fact table:** It is a special table that contains the data to measures the organisation's business operations.
- **Dimension tables:** It contains information which is mostly textual and descriptive and can be used in result set as the row headers.
- **Data Source Layer:** It refers to the layer representing various data sources that enter data into the data warehouse.
- **Data Staging Layer:** It refers to the storage area for data processing where data comes before being transformed into the data that is entered in a data warehouse.
- **Data Storage Layer:** It refers to the layer in which the transformed data and cleaned data is stored.
- **MDDB:** It refers to multidimensional database that allows data to be moulded and viewed in multiple dimensions.

## 7.8 CASE STUDY: DEVELOPING A MODERN DATA WAREHOUSE FOR FINANCIAL SERVICES ORGANISATION

Financial services organisations are constantly faced with the task of gaining new customers and providing accurate investment advice. However, all depends heavily on the use of various types of technologies for ensuring that all the fund transactions, client correspondence, market analytics and sales strategies remain reliable and responsive. In simple terms, we can say that the Information Technology department is a very critical department and it plays a major role in deciding the IT infrastructure of the organisation. In addition, the activities of the organisation are also planned using inputs from the IT department.

Ironside Group is an organisation that provides Business Analytics services and they help organisations in using their data to make better business decisions. Ironside has also helped FINSO (hypothetical name), a financial services organisation, to take care of its various IT needs. This was made possible by using a data warehouse, an information management solution, to process and prioritise various technologies.

Notes

FINSO employs more than 10,000 people and average revenues made by FINSO are more than USD 1 billion. FINSO hired Ironside to take care of areas, such as information management, business intelligence and financial performance management.

FINSO's IT department was having a hard time balancing its planning efforts and the pace of business. The IT department had taken steps to centralise all the data sources in a manner that accurate and actionable reporting could be done on such data. However, the infrastructure of FINSO had become obsolete and it could not handle the increased information needs of the FINSO staff. The IT team was also unable to integrate its data solution with the Business Intelligence reporting processes. FINSO wanted that Ironside helped their IT team in moving their existing data warehouse to a more sophisticated data warehouse that could address the following concerns:

- Old warehousing infrastructure was to be moved out of the existing obsolete environment.
- Historical data had to be stored so that the past events could be referenced accurately.
- All the pain points between the data warehouse and the Cognos BI reporting layer must be addressed.
- Increased query efficiency and timely analytics.

The Ironside information management team and FINSO's database engineers started working together to transition to a modern data warehouse. Ironside's executives conducted discovery conversations with the IT team of FINSO and they developed a full-scale migration and redesign plan to bring approximately 20 tables from various data sources, such as Excel, Oracle GL, A-Track, Remedy, TM1, etc. All these were moved to IBM PureData for Analytics (Netezza). Ironside also reengineered both the ETL processes. It was necessary to move and transform different information streams and the reporting layer so that the information could be made available for analysis.

Ironside successfully implemented several phases as follows:

- Collection of client requirements and use cases for developing the new data warehouse.
- Documentation of all the existing data warehouse logic including data extracts, transformations, schedules, etc.
- Specifications for the new ETL workflows were decided in consultation with FINSO's engineers.
- Reporting layer was rebuilt so that it could work seamlessly in the new data warehouse environment.

After all the phases were implemented, FINSO found that the results were quite promising. Some important results realised by FINSO include the following:

- Increased ease of use
- Improved performance in Cognos reporting

- Time comparison reporting was now possible
- System performance improved drastically. Reports that were previously generated in 2 minutes now took only 2 seconds.

**Source:** *https://www.ironsidegroup.com/2015/12/01/industry-case-study-modernizing-data-warehouse-finance-it/*

## QUESTIONS

1. What are the factors on which financial services organisations depend to gain new customers?

   (**Hint:** Financial services organisations are constantly faced with the task of gaining new customers and providing accurate investment advice. However, all depends heavily on the use of various types of technologies for ensuring that all the fund transactions, client correspondence, market analytics and sales strategies remain reliable and responsive.)

2. How does Ironside help organisations to take better business decisions?

   (**Hint:** Ironside Group is an organisation that provides Business Analytics services and they help organisations in using their data to make better business decisions.)

3. Why did FINSO hire Ironside?

   (**Hint:** FINSO's IT department was having a hard time balancing its planning efforts and the pace of business. Therefore, FINSO hired Ironside to take care of information management.)

4. Discuss the major concerns resolved by the implementation of the new data warehouse solution.

   (**Hint:** Old warehousing infrastructure was to be moved out of the existing obsolete environment, historical data had to be stored so that the past events could be referenced accurately.)

5. What are the benefits realised by FINSO after implementation of the ETL processes?

   **(Hint**: Some important results realised by FINSO include the following:

   - Increased ease of use
   - Improved performance in Cognos reporting)

## 7.9 EXERCISE

1. What is metadata? How is shared metadata defined?
2. What can be the consequences if we remove the data staging layer from the data warehouse system and skip directly to data storage layer?
3. We have a static data mart, where data changes every 4 months, except date and time. Is the change in data capture an efficient extraction implementation?
4. What are some alternate techniques for implementing a change capture on database source systems?
5. What are the challenges faced during multistage data transformation?
6. What makes pipelined data transformation significantly faster than multidimensional data transfer?

## 7.10 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Data Warehousing: An Informational Environment | 1. | transaction |
| | 2. | query |
| | 3. | True |
| | 4. | False |
| | 5. | True |
| | 6. | Strategic |
| | 7. | d. All of these |
| Key Components | 8. | data mart |
| | 9. | True |
| | 10. | fact |
| | 11. | Data source |
| | 12. | metadata |
| Data Warehouse Design Techniques | 13. | bottom-up, top-down |
| | 14. | True |
| | 15. | top-down |
| ETL Processes | 16. | ETL |
| | 17. | extraction, transformation and Loading |
| | 18. | False |
| | 19. | Full |
| | 20. | Incremental |

## 7.11 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. and Becker, B. (2011). *The Data Warehouse Lifecycle Toolkit*. Hoboken: John Wiley & Sons.
- Corr, L. and Stagnitto, J. (n.d.). *Agile Data Warehouse Design*.


### E-REFERENCES


- ETL Database. (2018). *ETL Process Overview*. [online] Available at: http://www.etldatabase.com/etl-process/[Accessed 21 Nov. 2018].
- Blog.panoply.io. (2018). *Data Warehouse Design: The Good, the Bad, the Ugly*.[online] Available at: https://blog.panoply.io/data-warehouse-design-thegood-the-bad-the-ugly [Accessed 21 Nov. 2018].
- Infogoal.com. (2018). *Data Warehousing Metadata*. [online] Available at: http://infogoal.com/datawarehousing/metadata.htm [Accessed 21 Nov. 2018].

CHAPTER 8

# Descriptive, Predictive, Prescriptive and Diagnostic Analytics

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Explain the importance of descriptive analytics
- Discuss the fundamentals of descriptive statistics
- Describe the need of predictive analytics
- Discuss predictive modelling
- Conduct model comparison and improvement
- Elucidate the overview of prescriptive analytics
- Explain the need of diagnostic analysis

## 8.1 INTRODUCTION

In the previous chapter, you have studied about data warehousing and its key components. Next, the chapter has discussed different techniques of designing a data warehouse. The Extract, Transform and Load (ETL) processes have been described at the end of the chapter.

Business analytics provides the insight and value to a business, making it one of the most important IT functions for any running business. It provides the ability to look beyond the heaps of structured and unstructured data and produce meaningful interpretations for strategic decisions, which makes it really powerful and valuable. There are different types of business analytics which helps organisations in knowing their customers or clients more closely.

One such type of business analytics is descriptive analytics. It can be defined as condensing the existing data to get a better understanding of what is going on using business intelligence tools. This helps to get an idea about what happened in the past and if it was as expected or not. For example, a coffee shop may learn how many customers they served in the time duration between 9 a.m. and 11 a.m. and which coffee was ordered the most. So, this analysis answers questions like "What happened?", but is not capable to answer deeper questions like "Why it happened?". Due to this reason, companies which are highly data-driven do not rely just on descriptive analysis. The companies use diagnostic analytics for this purpose.

Another important type of business analytics is predictive analytics. It can be defined as the process of focussing on predicting the possible outcome using machine-learning techniques like support-vector machine (SVM), random forests and statistical models. It tries to forecast on the basis of previous data and scenarios. So, this is used to find the answers to questions like "What is likely to happen?". For example, a hotel chain owner might ramp down promotional offers during a restive season of rains in a coastal area. This is based on the predictions that there are going to be fewer footfalls due to heavy rain. However, it must not be understood that this analysis can predict whether an event will occur in the future or not. It merely is able to predict the probability that an event will occur. If predictive analysis model is tuned properly based on historical data, it can be used to support complex predictions in marketing and sales. It can perform better than standard business intelligence in giving correct forecasts.

Notes

In the field of business analytics, model comparison and improvement play a crucial role in refining predictive analytics. This involves carefully assessing different models and implementing enhancements. By optimising these models based on historical data, businesses can achieve heightened accuracy in forecasting, surpassing traditional business intelligence. This process empowers businesses to make informed decisions, especially in dynamic areas like marketing and sales.

After studying predictive and descriptive analytics, one should be in a good position to take the final step, i.e., prescriptive analytics. This analysis will provide a prediction or a forecast of what future trends in the business may look like.

For example, there can be significant statistical measures of higher or lower sales, profitability trends accurately measured in dollars for new market prospects, or measured cost savings from a future joint venture. In the event that the organisations know where the future lies by foreseeing the patterns, they can best arrange to exploit conceivable plans that the patterns may offer. The third step of the business analytics process is prescriptive analytics, which involves the application of decision science, operations research methodologies and management science to make optimal utilisation of the available resources.

Prescriptive analytics methods and techniques are mathematically-based algorithms designed to take variables and other parameters into a qualitative framework and generate an optimal or real-time solution for complex problems. Such methods can be utilised to ideally distribute a company's limited assets to take the best preferred advantage of opportunities it has found in the anticipated future patterns. The limitations on human and financial assets turn away organisations from pursuing each opportunity. Utilising prescriptive analytics allows an organisation to designate limited assets to accomplish goals as ideally as possible. Prescriptive analytics is simply a computerised method for applying calculation and interpretation and providing valuable insights from various data sources.

This chapter begins by describing descriptive analytics. Further, this chapter explains predictive analytics in detail. Next, the chapter discusses the predictive modelling and its different types. Towards the end, the chapter elucidates model comparison and improvement and prescriptive analytics.

## 8.2 DESCRIPTIVE ANALYTICS

Descriptive analytics involves "What has occurred in the corporation" and "What is going on now?" Let us consider the case of Facebook. Facebook users produce content through comments, posts and picture uploads. This information is unstructured and is produced at an extensive rate. Facebook stats reveal that 2.4 million posts equivalent to around 500 TB of information are produced every minute. These jaw-dropping figures have offered popularity to another term which we know as Big Data. Comprehending the information in its raw configuration is troublesome. This information must be abridged, categorised and displayed in an easy-to-understand way to let the managers comprehend it.

Business intelligence and data mining instruments/methods have been accepted components of doing so for bigger organisations. Practically, every association does

some type of outline and Management Information System (MIS) reporting using the information base or simply spreadsheets. There are three crucial approaches to abridge and describe the raw data:

- **Dashboards and MIS reporting:** This technique provides condensed data giving information on 'What has happene', 'What's been going on' and 'How can it stand with the plan?'
- **Impromptu detailing:** This technique supplements the past strategy in helping the administration to extract the information as required.
- **Drill-down reporting:** This is the most complex piece of descriptive analysis and gives the capacity to delve further into any report to comprehend the information better.

**SELF ASSESSMENT QUESTIONS**

1. Which of the following types of analytics provides a prediction or a forecast of future trends in the business?

   a. Descriptive analytics
   b. Predictive analytics
   c. Prescriptive analytics
   d. None of these

2. Drill-down reporting is the most complex part of descriptive analysis and provides the capability to delve deeper into any report to better understand the information. (True/False)

## 8.3 DESCRIPTIVE STATISTICS

Statistics, as defined by David Hand, the former president of the Royal Statistical Society, UK, is both the science of uncertainty and the technology of extracting information from data. Statistics involves collecting, organising, analysing, interpreting and presenting data. Descriptive statistics refers to the set of statistical techniques and methods used to summarise, organise, and describe the main features of a dataset. The primary goal is to provide a meaningful and concise representation of the data, enabling analysts and decision-makers to gain insights into its characteristics, patterns, and trends.

You are familiar with the concept of statistics in daily life as reported in newspapers and the media, for example, baseball batting averages, airline on-time arrival performance, and economic statistics such as the Consumer Price Index (CPI). Statistical methods are essential to business analytics. Microsoft Excel supports statistical analysis in two ways:

- With statistical functions that are entered in worksheet cells directly or embedded in formulas.
- With the Excel Analysis Toolpak add-in to perform more complex statistical computations.

A population consists of all items of interest for a particular decision or investigation for example, all individuals in the United States of America who do not own cell phones, all subscribers to Netflix, or all stockholders of Google. A company like Netflix keeps extensive records of its customers, making it easy to retrieve data about the entire population of customers. However, it would probably be impossible to identify all individuals who do not own cell phones.

Notes

A sample is a subset of a population. For example, a list of individuals who rented a comedy from Netflix in the past year would be a sample from the population of all customers. Whether this sample is representative of the population of customers—which depends on how the sample data is intended to be used—may be debatable; nevertheless, it is a sample. Most populations, even the finite ones, are usually too large to practically or effectively deal with. For example, it would be unreasonable as well as costly to survey the TV viewers' population of the United States of America. Sampling is also necessary when data must be obtained from destructive testing or from a continuous production process.

Thus, the process of sampling aims to obtain enough information to create a legal interpretation about a population. Market researchers, for example, use sampling to gauge consumer perceptions on new or existing goods and services, auditors use sampling to verify the accuracy of financial statements, and quality control analysts sample production output to verify quality levels and identify opportunities for improvement.

### 8.3.1 UNDERSTANDING STATISTICAL NOTATION

We typically label the elements of a dataset using subscripted variables, $x_1$, $x_2$, …, and so on. In general, $x_i$ represents the i*th* observation. In statistics, it is common to use Greek letters, such as σ(sigma), μ (mu), and π (pi), to represent the population measures and italic letters such as by (x-bar), s, and p for sample statistics. We will use N to represent the number of items in a population and n to represent the number of observations in a sample. Statistical formulas often contain a summation operator, (Greek capital sigma), which means that the terms that follow it are added together. Thus, understanding these conventions and mathematical notations will help you interpret and apply statistical formulas.

### 8.3.2 CENTRAL TENDENCY

Central tendency is the measurement of a single value that attempts to describe a set of data by identifying the central position within that set of data. Measurement of central tendency is also called as measures of central location. Some common terms used as valid measures of central tendency are as follows:

- **Mean:** The mathematical average is called the mean (or the arithmetic mean), which is the sum of the observations divided by the total number of observations. The mean of a population is shown by the μ, and the sample mean is denoted by x. If the population contains N observations $x_1$, $x_2$,…, $x_N$, then the population mean is calculated as:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The sample mean of n sample observations $x_1$, $x_2$, …, $x_n$ is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

One property of the mean is that the sum of the deviations of each observation from the mean is zero:

$$\sum_i (X_i - \bar{X}) = 0$$

This simply means that the sum of deviations above the mean is the same as the sum of deviations below the mean. Thus, the mean 'balances' the values on either side of it. However, it does not suggest that half the data lie above or below the mean.

- **Median:** The measure of location that specifies the middle value when the data is arranged from least to greatest is the median. If the number of observations is odd, the median is the exact middle of the sorted numbers, i.e., the 4th observation. If the number of observations is even, say 8, the median is the mean of the two middle numbers, i.e., the mean of 4th and 5th observations. We can use the Sort option of MS Excel to order the data as per the rank and then find the median. The Excel function MEDIAN (data range) could also be used.
- **Mode:** A third method of measuring the location is called mode. It is the observation/number/series that occurs the maximum number of times in a group of entities. Mode is valuable for data sets containing smaller number of unique values. You can easily identify the mode from a frequency distribution by identifying the value having the largest frequency or from a histogram by identifying the highest bar. You may also use the Excel function, MODE.SNGL (data range). For frequency distributions or grouped data, the modal group is the group with the greatest frequency.
- **Midrange:** A fourth measure of location that is used occasionally is the midrange. This is simply the average of the greatest and least values in the data set.

### 8.3.3 VARIABILITY

A commonly used measure of dispersion is the variance. Basically, variance is the squared deviations average of the observations from the mean. The bigger the variance is, the more is the spread of the observations from the mean. This indicates more variability in the observations. The formula used for calculating the variance is different for populations and samples. The formula for the variance of a population is:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

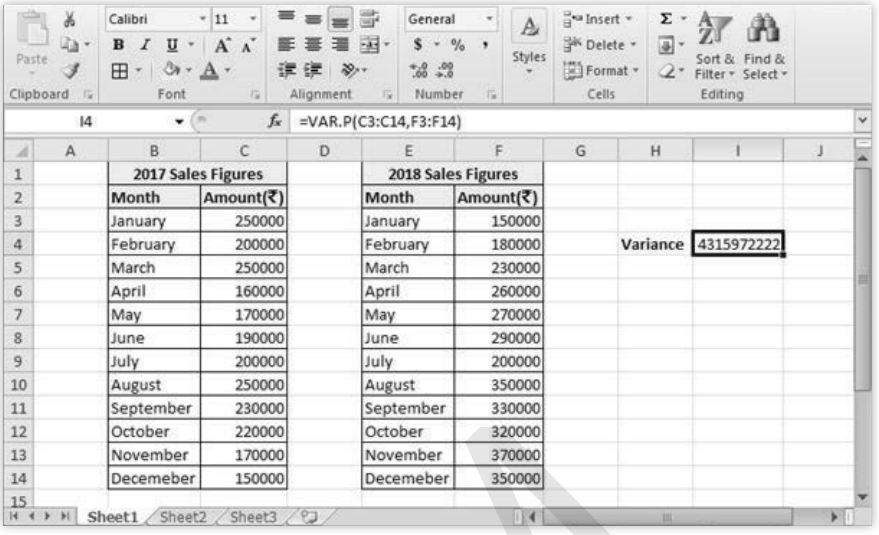
where $x_i$ is the value of the ith item, N is the number of items in the population, and $\mu$ is the population mean. The variance of a sample is calculated by using the formula:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

where n is the number of items in the sample and $\bar{x}$ is the sample mean.

Notes

The monthly sales figures of an organisation for two years is as shown in cells C3-C14 and F3-F14. The variance of the two years' sales figures can be calculated in cell I4 of the spreadsheet by using the VAR.P function, as shown in Figure 1:




| | 2017 Sales Figures | | | 2018 Sales Figures | |
|---|---|---|---|---|---|
| 2 | Month | Amount(₹) | | Month | Amount(₹) |
| 3 | January | 250000 | | January | 150000 |
| 4 | February | 200000 | | February | 180000 |
| 5 | March | 250000 | | March | 230000 |
| 6 | April | 160000 | | April | 260000 |
| 7 | May | 170000 | | May | 270000 |
| 8 | June | 190000 | | June | 290000 |
| 9 | July | 200000 | | July | 200000 |
| 10 | August | 250000 | | August | 350000 |
| 11 | September | 230000 | | September | 330000 |
| 12 | October | 220000 | | October | 320000 |
| 13 | November | 170000 | | November | 370000 |
| 14 | Decemeber | 150000 | | Decemeber | 350000 |

Figure 1: Calculating the Variance

Suppose we want to calculate the variance of heights of adult females in India. It is not possible to measure the height of all women in India. In table 1, a sample comprising 50 females has been extracted from the overall population:

| S. No. | Height (in cms) | S. No. | Height (in cms) |
|---|---|---|---|
| 1 | 155 | 26 | 158 |
| 2 | 157.3 | 27 | 141 |
| 3 | 161.4 | 28 | 141 |
| 4 | 148 | 29 | 143 |
| 5 | 146 | 30 | 143 |
| 6 | 144 | 31 | 142 |
| 7 | 143 | 32 | 145 |
| 8 | 141 | 33 | 147 |
| 9 | 155 | 34 | 147 |
| 10 | 156 | 35 | 148 |
| 11 | 138 | 36 | 148 |
| 12 | 152 | 37 | 149 |
| 13 | 154 | 38 | 149 |
| 14 | 145 | 39 | 150 |
| 15 | 147 | 40 | 150 |
| 16 | 148 | 41 | 151 |
| 17 | 149 | 42 | 152 |
| 18 | 150 | 43 | 153 |
| 19 | 151 | 44 | 144 |
| 20 | 152 | 45 | 154 |
| 21 | 153 | 46 | 157 |
| 22 | 154 | 47 | 160 |
| 23 | 155 | 48 | 165 |
| 24 | 157 | 49 | 170 |
| 25 | 157 | 50 | 158 |

The variance of the sample population of females can be calculated by using the VAR.S function, as shown in Figure 2:

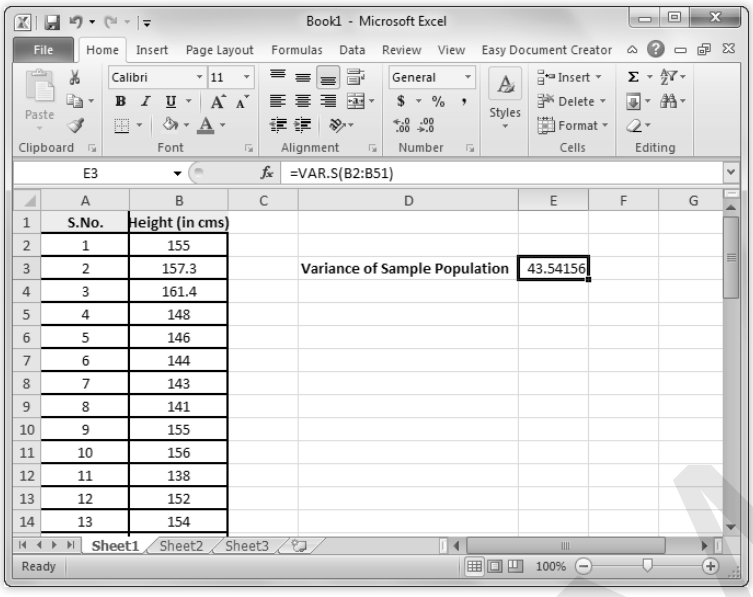



FIGURE 2: Calculating the Variance of Population

## 8.3.4 STANDARD DEVIATION

The square root of the variance is the standard deviation. For a population, the standard deviation is computed as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

and for samples, it is

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

The standard deviation is usually easier to understand than the variance because of similarity in its measure units that are same as the data units. Thus, it can be more easily related to the mean or other statistics measured in the same units.

The standard deviation is a popular measure of risk, particularly in financial analysis, because many people associate risk with volatility in stock prices. The standard deviation measures the tendency of a fund's monthly returns to vary from their long-term average (as Fortune stated in one of its issues, ". . . standard deviation tells you what to expect in the way of dips and rolls. It tells you how scared you'll be.").

For example, a mutual fund's return might have averaged 11% with a standard deviation of 10%. Thus, about two-thirds of the time, the annualised monthly return was between 1% and 21%. By contrast, another fund's average return might be 14% but have a standard deviation of 20%. Its returns would have fallen in a range of –6% to 34% and, therefore, is riskier.

NOTES

You can calculate the standard deviation for heights of a sample of 50 females in Excel by using the STDEV.S function, as shown in Figure 3:

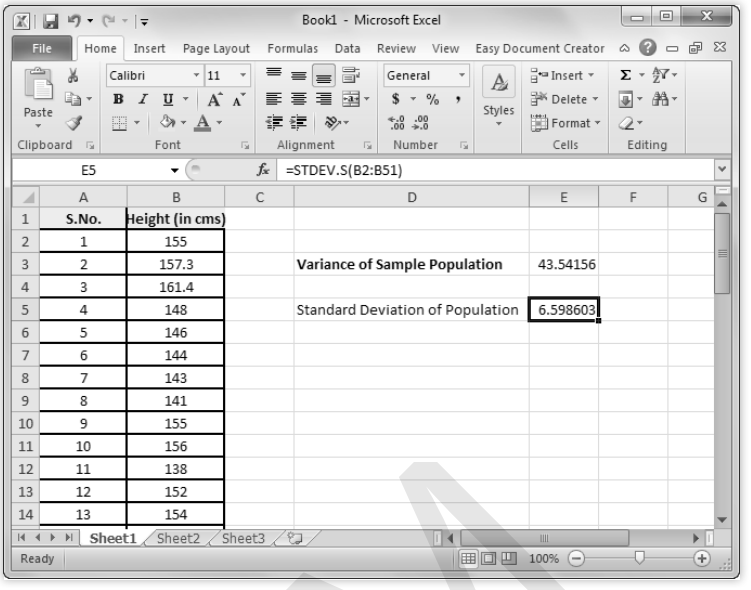



FIGURE 3: CALCULATING THE STANDARD DEVIATION

## Standardised Values

A z-score, or standardised value, provides a measure of the distance of the observation away from the mean, irrespective of the measurement units. In a data set, z-score for the ith observation is calculated as follows:

- We subtract the sample mean from the ith observation, xi, and divide the result by the sample standard deviation. The numerator denotes the distance that xi is away from the sample mean, a negative value designates that $x_i$ is at the left of the mean, and a positive value means it lies at the right. By dividing by the standard deviation, s, we scale the distance from the mean to express it in units of standard deviations.
- Thus, a z-score of 1.0 means that the observation is one standard deviation to the right of the mean and a z-score of –1.5 means that the observation is 1.5 standard deviations to the left of the mean. Thus, even though two data sets may have different means and standard deviations, the same z-score means that the observations have the same relative distance from their respective means.
- z-scores can be computed easily on a spreadsheet; however, Excel has a function that calculates it directly, STANDARDIZE (x, mean, standard_dev).

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Suppose you want to calculate the standardised values on the basis of the marks obtained by 10 students out of 100. You first need to compute the mean of the marks by using the =AVERAGE(B3:B12) formula and then the standard deviation of the population by using the =STDEV.P(B3:B12) formula.

Now, you can compute the z-score in cell C3 by using the =STANDARDIZE(B3,$F$3,$F$4) formula, as shown in Figure 4:

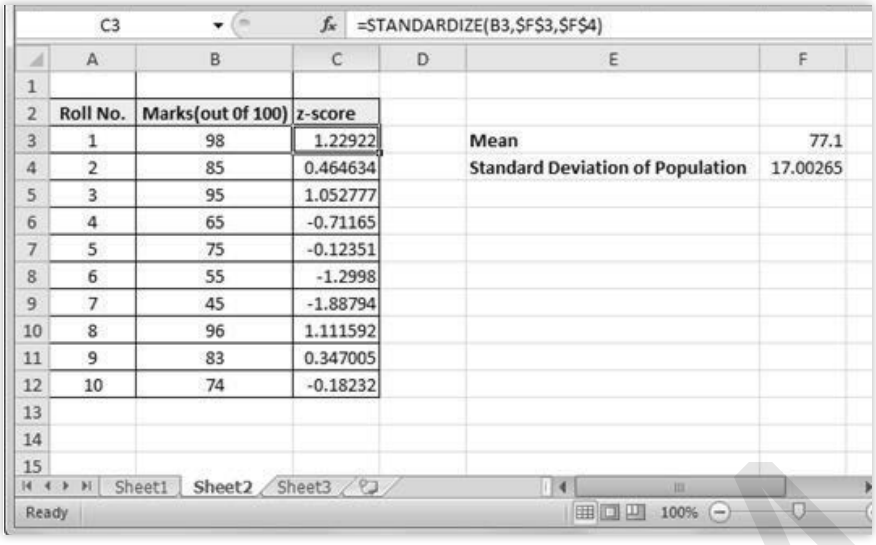



Figure 4: Computing the z-score

You can calculate the z-score for each student in the similar manner.

### Coefficient of Variation

The Coefficient of Variation (CV) provides a relative measure of the dispersion in data relative to the mean and is defined as:

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}}$$

Often, the coefficient of variation is multiplied by 100 to be expressed as a percentage. This statistic is useful when comparing the variability of two or more data sets when their scales differ. The coefficient of variation offers a relative risk to return measure.

The smaller the coefficient of variation, the smaller the relative risk is for the return provided. The reciprocal of the coefficient of variation, called return to risk, is often used because it is easier to interpret. That is, if the objective is to maximise return, a higher return-to-risk ratio is often considered better. A related measure in finance is the Sharpe ratio, which is the ratio of a fund's excess returns (annualised total returns minus Treasury bill returns) to its standard deviation.

If several investment opportunities have the same mean but different variances, a rational (risk-averse) investor will select the one that has the smallest variance. This approach to formalising risk is the basis of the modern portfolio theory, which seeks to construct minimum-variance portfolios.

**Self Assessment Questions**

3. A _____ consists of all items of interest for a particular decision or investigation.
4. A _____ is a subset of a population.
5. Sampling is also necessary when data must be obtained from destructive testing or from a continuous production process. (True/False)

NOTES

**ACTIVITY**

Prepare a report on the relationship between statistical analytical concepts and their usage in analytical sciences in the simplest manner possible.

## 8.4 PREDICTIVE ANALYTICS

Predictive analytics is a branch of business analytics that utilises data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. In other words, it involves the use of statistical models and algorithms to analyse current and historical facts to make predictions about future events or trends. Predictive analytics is widely employed in various industries, including business, finance, healthcare, marketing, and more, to support decision-making and gain a competitive advantage.

Figure 5 shows the steps involved in predictive analytics:

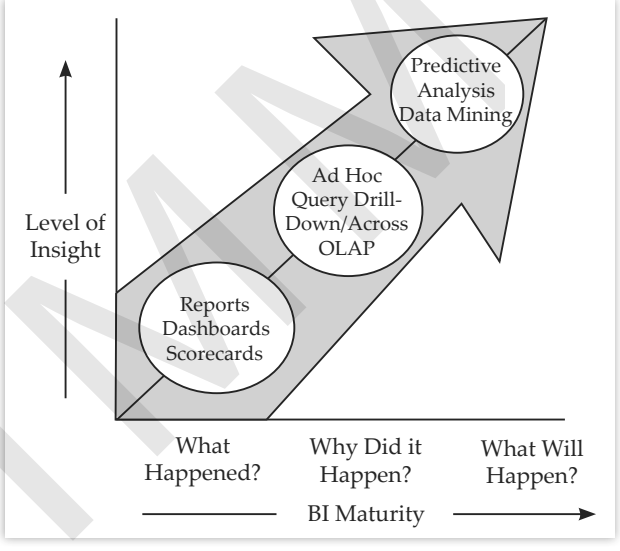



**FIGURE 5: Predictive Analytics**

**Source:** *http://www.witinc.com/predictive-analytics.id.355.htm*

**SELF ASSESSMENT QUESTIONS**

6. _____ analytics performs an in-depth analysis of data to reveal details such as frequency of events, operation costs and the underlying reason for failures.
7. In predictive analytics, we use statistics, data mining techniques and machine learning to analyse the future. (True/False)

**ACTIVITY**

Search and enlist some commercially available predictive analytics tools in the market.

## 8.5 PREDICTIVE MODELLING



Predictive modelling is the method of making, testing, and authenticating a model to best predict the likelihood of a conclusion. Several modelling procedures from artificial intelligence, machine learning and statistics are present in predictive analytics software solutions. The model is selected on the basis of testing, authentication and assessment using the detection theory to predict the likelihood of an outcome in a given input data amount. Models can utilise single or more classifiers to decide the probability of a set of data related to another set. The different models available for predictive analytics software enable the system to develop new data information and predictive models. Each model has its own strengths and weaknesses and is best suited for different types of problems.

Predictive analysis and models are characteristically used to predict future probabilities. Predictive models in business context are used to analyse historical facts and current data to better comprehend customer habits, partners and products and to classify possible risks and prospects for a company. It practices many procedures, including statistical modelling, data mining and machine learning aid analysts to make better future business predictions.

- Predictive modelling is at the heart of business decision making
- Building decision models more than science is an art
- Creating an ideal decision model demands:
  - Good understanding of functional business areas
  - Knowledge of conventional and in-trend business practices and research
  - Logical skillset
- It is always recommended to start simple and keep on adding to the models as required.

The greatest set of changes and advances in predictive modelling are coming to fruition due to the increase in unstructured information such as content archives, video, voice and pictures from which useful information can also be extracted. Basically, predictive modelling requires organised data of the kind that is not found in social networking databases. To make unstructured data indexes valuable for this sort of examination, organised data must be extricated from them first. One case is sentiment analysis from Web posts. It would be about impossible, in any case, to attempt to assemble a predictive model specifically from the content in the posts themselves. An extraction step is required to get usable data as keywords, expressions and importance from the content in the posts.

Predictive models are representations of the relationship between how a member of a sample performs and some of the known characteristics of the sample. The aim is to assess how likely a similar member from another sample is to behave in the same manner. This model is used a lot in marketing. It helps identify implied patterns which indicate customers' preferences. This model can even perform calculations at the exact time when a customer performs a transaction.

NOTES

A predictive analytics model combines many predictors or quantifiable variables. This method allows for the data collection and preparation of a statistical model, to which extra data can be added as and when available.

The accumulation of higher data volumes creates a nifty predictive model, trusting the larger data sets which produce more dependable forecasts based on the data volume examined. Moreover, trusting the actual data to power predictive analytics models marks better accurateness of the predicting process. The various business processes on predictive modelling are as follows:

- **Creating the model:** A software-based solution allows you to make a model of multiple algorithms on the data set.
- **Testing the model:** Test the predictive model on the data set. In some situations, the testing is done on previous data to the effectiveness of a model's prediction.
- **Authenticating the model:** Authenticate the model results by means of business data understanding and visualisation tools.
- **Assessing the model:** Assessing the best suited model from the used models and selecting the appropriate model tailored for the data.

The predictive modelling process includes executing one or more algorithms on the data set subjected to prediction. This is a recurring process and often includes model training, using several models on the same data set and lastly getting the appropriate model based on the business data.

### 8.5.1 LOGIC-DRIVEN MODELS

Logic-driven models are created on the basis of inferences and postulations provided by the sample space and existing conditions. Creating logical models requires solid understanding of business functional areas, logical skills to evaluate the propositions better and knowledge of business practices and research. To understand better, let us take an example of a customer who visits a restaurant around six times a year and spends around ₹5000 per visit. The restaurant gets around 40% margin on per visit billing amount.

The annual gross profit on that customer turns out to be 5000 × 6 × 0.40 = ₹12000.

30% of the customers do not return each year, while 70% do return to provide more business to the restaurant.

Assuming the average lifetime of a customer (time for which a consumer remains a customer) → 1/.3 = 3.33 years. So, the average gross profit for a typical customer turns out to be 12000 × 3.33 = ₹39,960

Armed with all the above details, we can logically arrive at a conclusion and can derive the following model for the above problem statement:

$$\text{Economic Value of each Customer (V)} = \frac{R \times F \times M}{D}$$

where,

R = Revenue generated per customer

F = Frequency of visits per year

M = Profit margin

D = Defection rate (Non-returning customers each year)

So, as you can see, logical driven predictive models can be derived for a number of situations, conditions, problem statements and other scenarios where predictive analytical models provide a futuristic view on the basis of validation, testing and evaluation to guess the likelihood of an outcome in a given set amount of input data.

### 8.5.2 DATA-DRIVEN MODELS

The main aim of the data-driven model concept is to find links between the state system variables (input and output) without clear knowledge of the physical attributes and behaviour of the system. The data-driven predictive modelling derives the modelling method based on the set of existing data and entails a predictive methodology to forecast the future outcomes. A company expecting losses in the current quarter due to the poor market performance and sentiments is a classic example of data-driven predictive modelling. You have the data and you know about data inferences. Here, you are simply predicting the outcomes based on the data.

**Self Assessment Questions**

8. _____ is the method of making, testing and authenticating a model to best predict the likelihood of a conclusion.
   a. Predictive modelling
   b. Descriptive modelling
   c. Prescriptive modelling
   d. None of these
9. Predictive analysis and models cannot be used to predict future probabilities. (True/False)
10. _____ models are created on the basis of inferences and postulations provided by the sample space and existing conditions.
11. The _____ predictive modelling derives the modelling method based on the set of existing data and entails a predictive methodology to forecast the future outcomes.

## 8.6 MODEL COMPARISON AND IMPROVEMENT

Business analytics relies heavily on predictive models to gain insights and make informed decisions. Model comparison and improvement constitute a crucial stage in the lifecycle of these models, playing a key role in enhancing their accuracy, reliability, and overall effectiveness. This phase acknowledges that the real-world performance of models may deviate from expectations due to changing data patterns, external factors, or advancements in analytical techniques.

The process of improvement is not a one-time event; it is an ongoing commitment to refining and optimising models. This involves incorporating new data, updating algorithms, and considering advancements in technology. Continuous refinement ensures that predictive models remain relevant and effective in the face of changing business dynamics.

Notes

Business environments are dynamic, and requirements can shift rapidly. Models that were effective in the past may lose relevance over time. Model comparison and improvement strategies enable organisations to adapt to changing business needs, ensuring that predictive models align with current objectives and contribute meaningfully to decision-making processes. The organisations that prioritise it gain a competitive edge by ensuring their analytics models are at the forefront of precision and reliability. This, in turn, enhances strategic planning, resource allocation, and overall business performance.

This iterative process is a strategic imperative for organisations aiming to fully harness analytics in decision-making.

### 8.6.1 EVALUATING MODEL PERFORMANCE

The evaluation of model performance involves the use of various metrics to quantify how well a model is performing. This includes assessing its ability to correctly predict outcomes, avoid false positives and false negatives, and adapt to different scenarios. Assessing the efficacy of predictive models demands a nuanced understanding of diverse metrics tailored to the specific business context. Some commonly used metrics across various applications are:

- **Accuracy:** Accuracy represents the overall correctness of predictions. It is the ratio of correctly predicted instances (true positives and true negatives) to the total instances.
- **Precision:** Precision focuses on the accuracy of positive predictions. It measures the proportion of true positive predictions among all instances predicted as positive, indicating the model's ability to avoid false positives.
- **Recall (Sensitivity or True Positive Rate):** Recall assesses the model's ability to capture all positive instances. It measures the proportion of true positive predictions among all actual positive instances, highlighting the model's sensitivity to detecting relevant cases.
- **F1 Score:** The F1 score combines precision and recall into a single metric. It provides a balance between the two, giving an overall measure of a model's performance on both false positives and false negatives.
- **Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC):** The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate at various thresholds. AUC-ROC quantifies the model's ability to distinguish between classes, with a higher AUC indicating better performance.
- **Specificity (True Negative Rate):** Specificity measures the ability of a model to correctly identify negative instances. It is particularly relevant in scenarios where avoiding false positives is crucial, such as in medical diagnoses.
- **Mean Absolute Error (MAE):** MAE measures the average absolute difference between predicted and actual values, providing a straightforward assessment of the model's accuracy in terms of the magnitude of errors.
- **Mean Squared Error (MSE):** MSE calculates the average squared difference between predicted and actual values, emphasising larger errors more than MAE. It offers a measure of overall model performance.

- **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, offering an interpretable scale similar to the original target variable. It provides a sense of the average magnitude of errors in the original units.
- **R-squared (Coefficient of Determination):** R-squared signifies the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R-squared value indicates a better-fitting model, capturing a larger share of the variability.

### 8.6.2 TECHNIQUES FOR MODEL COMPARISON

Model comparison entails assessing multiple models to identify the one that best suits the business objectives. Techniques for model comparison involve systematic methods to assess and contrast the performance of different predictive models. Some common techniques include:

- **Cross-validation:** Cross-validation involves dividing the dataset into subsets for training and testing, ensuring thorough model evaluation by rotating these subsets. It assesses a model's generalisation capability across different data partitions.
- **A/B testing:** A/B testing directly compares the performance of two or more models in real-world scenarios, measuring their impact on specific outcomes or key performance indicators. This method aids in selecting the most effective model for deployment.
- **Statistical hypothesis testing:** Statistical hypothesis testing applies tests to compare performance metrics of different models, helping determine if observed differences are statistically significant. It guides the selection of the best-performing model based on statistical evidence.
- **Ensemble methods:** Ensemble methods combine predictions from multiple models, utilising techniques like bagging and boosting to enhance overall performance. This approach leverages the diverse strengths of individual models for improved predictive accuracy.
- **Receiver Operating Characteristic (ROC) curve analysis:** ROC curve analysis plots the trade-off between true positive rate and false positive rate at different classification thresholds. AUC-ROC quantifies a model's discriminative ability, facilitating comparison in binary classification tasks.

### 8.6.3 MODEL IMPROVEMENT STRATEGIES

Model improvement strategies in business analytics encompass deliberate actions aimed at refining and enhancing predictive models for superior performance. This iterative process involves continuous monitoring, updating with new data, and adapting to evolving business dynamics.

By incorporating user feedback and considering ethical considerations, models can align with societal values and user expectations. Leveraging advanced techniques such as ensemble methods and hyperparameter tuning further refines models, optimising their accuracy and relevance. Additionally, a focus on explainability and interpretability ensures that models remain transparent and understandable.

Notes

Overall, model improvement strategies play a pivotal role in ensuring that predictive models remain robust, reliable, and aligned with the dynamic needs of the business environment, ultimately contributing to more informed decision-making processes.

**Self Assessment Questions**

12. Which metric emphasises the average absolute difference between predicted and actual values?
    a. Precision
    b. Recall
    c. Mean Squared Error (MSE)
    d. Mean Absolute Error (MAE)
13. Cross-validation assesses a model's generalisation capability across the same data partitions. (True/False)
14. Model improvement strategies involve continuous monitoring, updating with _____, and adapting to evolving business dynamics.

## 8.7 OVERVIEW OF PRESCRIPTIVE ANALYTICS

By using the optimisation technique, prescriptive analytics determines the finest substitute to minimise or maximise some equitable finance, marketing and many other areas. For example, if we have to find the most economical way of shipping goods from a factory to a destination, we will use prescriptive analytics. Figure 6 shows a diagrammatic representation of the stages involved in the prescriptive analytics:

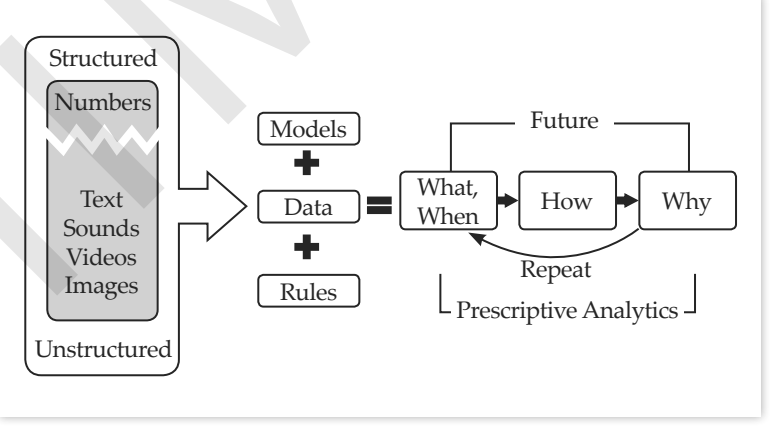



Figure 6: Prescriptive Analytics

Data, which is available in abundance, can be streamlined for growth and expansion in technology as well as business. When data is analysed successfully, it can become the answer to one of the most important questions: how can businesses acquire more customers and gain business insight? The key to this problem lies in being able to source, link, understand and analyse data.

All companies need to address their data challenges to support their decision-making capabilities or may risk themselves in falling behind in this highly competitive landscape. Today, businesses are collecting, storing, analysing and interpreting more

data as compared to the previous years, and this trend is continuing at an alarming rate to gain momentum. According to many leading professors and researchers, this is the era of a Big Data revolution. In any case, it is not the amount of information that is progressive. Rather, the revolution has got something to do with the volume, variety and velocity of data.

Since a lot has been written on Big Data, we will focus on analytics, which will help companies transform the finance function by offering forward looking insights and help them devise a solution appropriate for the optimal course of action, improve the ability to communicate and collaborate with other companies at a lower cost of ownership.

These transformative characteristics will lead to better performance improvements in business sectors. Prescriptive analytics go beyond predictions, workforce optimisations and decision options. It is usually used to analyse complex data to analyse huge complex data to forecast outcomes, offer decision options and show alternative business impact. This method also consists of many scientific and mathematical methods used for understanding how alternative learning investments impact the bottom line. Moreover, this analytics can also help enterprises to take decisions on how to take advantage of a future scenario or reduce a risk in the future course of time and represent the implication of each decision option.

In real life, prescriptive analytics can automatically and continuously process new data to improve forecast accuracy and offer better decision options. For instance, prescriptive analytics can be utilised to profit healthcare key arranging. By utilising data analytics, one can harness operational information which includes population statistic patterns, financial information, and population health patterns to a more exact arrangement and contribute future capital, such as equipment usage and new facilities.

### 8.7.1 HOW PRESCRIPTIVE ANALYTICS FUNCTIONS

Utilising prescriptive analytics is a complex and time-taking process that investigates all viewpoints to sustain the decision-making process, including:

- Identifying and breaking down every single potential choice
- Defining potential connections and associations between each of these choices with each other
- Identifying variables that could affect each of these choices (positively or negatively)

Prescriptive analytics handle processes from either one of these viewpoints and maps out (at least one) potential results for each of the choices, bringing about a customised model. Elements sustaining the model, including information volume and quality, could affect the exactness of the model (as they would in descriptive and predictive analytics).

Prescriptive analytics utilises procedures like optimisation, game theory, simulation, and decision-analysis techniques. Prescriptive analytics can constantly and consequently prepare new information to enhance predictive precision and give better decision choices.

Notes

### 8.7.2 COMMERCIAL OPERATIONS AND VIABILITY

Most organisations have concentrated intensely on finding the correct cost levels and working model to adequately empower them for their development. Prescriptive analytics adds another measurement to operational and business adequacy by giving directors a chance to foresee what structures, messages and targets will yield ideal outcomes, given the organisation's remarkable parameters, and after that choose which way will give the biggest returns. There are numerous other business applications of prescriptive analytics, such as:

- Optimising spend and rate of profitability (ROI) through exact customer profiling
- Providing important data for brand planning and go-to-market procedures
- Maximising campaign productivity, sales force arrangement and promotional activities
- Predicting and proactively overseeing market events
- Providing significant data for territory examination, customer deals and medical data

A one-estimate fits-all business model is no longer reasonable. The eventual fate of a focused sales model is focused on customised messaging.

### 8.7.3 RESEARCH AND INNOVATION

Prescriptive analytics can be a noteworthy differentiator for any organisation occupied with Research and Development exercises in a competitive industry including:

- Demonstrating, anticipating and enhancing results from item utility
- Understanding sickness (or different zones of intrigue) patterns/movement
- Establishing ideal trial conditions through focused patient cohorts
- Increasing customer adherence to the item and diminishing compliance
- Understanding necessities for customised drug and different advancements
- Determining and setting up focused items and interventions
- Determining and setting up an ideal trial conditions through focused patient cohorts

### 8.7.4 BUSINESS DEVELOPMENT

Understanding what new items are required, what differentiating components will make one item sell better than the other, or which markets are demanding which items are key zones for prescriptive analytics including:

- Identifying and settling on choices about circumstances/rising ranges of unmet needs
- Predicting the potential advantage

- Following industry trends proactively and actualising techniques to get an advantage
- Exploiting data analytics to distinguish particular buyer populations and regions that ought to be focused on
- Leveraging data analytics to distinguish key advancements for item improvement that will produce the biggest return for the investment
- Identifying likely purchasers to cut business improvement costs altogether, and imagine a scenario where situations for items, markets and purchasers could be an unmistakable differentiator for developing organisations

### 8.7.5 CONSUMER EXCELLENCE

Prescriptive analytics can be utilised to improve purchaser excellence in a huge number of ways including:

- Predicting what purchasers will need and settling on key choices that address those necessities
- Segmenting purchasers and recognising and focusing on custom fitted messages to them
- Staying on top of competition and deciding (e.g., marketing, branding) about items that will prompt more desirable items and higher sales

### 8.7.6 CORPORATE ACCOUNTS

Corporate account functions can immensely use prescriptive analytics to improve their capacity to settle on choices that help drive internal excellence and outer strategy:

- Internal excellence
  - Viability and direction for non-item related activities; what choices ought to be made and what is the effect
  - Viability and direction for item related activities; what choices ought to be made and what is the effect
- External-facing key direction
  - Utilising important data to demonstrate item esteem and build up a market valuing
  - Utilising examination to build up a targeted on coupon strategy
  - Recognising ideal price point alternatives and the effect of those choices on the income model for the item
  - Better understanding the whole price cycle from rundown cost to repayment (counting all rebates and refunds) to inform the ideal pricing system
  - Utilising an important competitor data to build up estimates and get market access

Notes

### 8.7.7 SUPPLY CHAIN

Prescriptive analytics can likewise furnish supply chain capacities with an upper hand through the capacity to predict and make decisions in a few basic areas including:

- Forecasting future demand and pricing (e.g., supplies, material, fuel and different components affecting cost to guarantee proper supply)
- Utilising prescriptive analytics to illuminate stock levels, schedule plants, route trucks and different components in the supply chain cycle
- Modifying supplier threat by mining unstructured information regarding value-based information
- Understanding historical demand examples and product course through supply chain channels, anticipating future examples and settling on choices on future state procedures

### 8.7.8 GOVERNANCE, RISK, AND COMPLIANCE

Governance, risk and compliance (GRC) is a strategy used to manage an organisation's overall governance, risk and compliance with regulations. Prescriptive analytics can help associations accomplish consistence through the capacity to anticipate upcoming dangers and settle on the proper mitigation choices.

Governance, risk and compliance are functions of increasing importance across almost every industry. Prescriptive analytics can help organisations achieve compliance through the ability to expect forthcoming risks and make proper mitigation decisions. Utilisation of prescriptive analytics in the region of governance, hazard and compliance incorporates:

- Improving internal review effectiveness
- Notifying third-party arrangement and management
- Classifying patterns related with outlandish spend (e.g., total spend working on this issue of pharma)
- Applying learned compliance controls

**Self Assessment Questions**

15. By using the _____ technique, prescriptive analytics determines the finest substitute to minimise or maximise some equitable finance, marketing and many other areas.
16. Prescriptive analytics is usually used to analyse complex data to analyse huge complex data to forecast outcomes, offer decision options and show alternative business impact. (True/False)
17. Which of the following types of procedures are used by prescriptive analytics?
    a. Optimisation
    b. Game theory
    c. Simulation
    d. All of these

18. Corporate account functions can immensely use prescriptive analytics to improve their capacity to settle on choices that help drive internal excellence and outer strategy. (True/False)

**Activity**

Create a PowerPoint presentation on descriptive, predictive and prescriptive analytics. Show the presentation in your class.

## 8.8 DIAGNOSTIC ANALYTICS

Diagnostic analytics is used to find the root cause of a given situation. It can also be used to find the casual relationships between two or more data sets if the root cause is not detectable. The analytics team or person must be careful about selecting relevant data for analysis or for finding relation among more than one data set.

Let us take an example where you have done descriptive analytics and it shows low sales on your online grocery store website. Followed by some event checks and analysis, it occurs to you that users are adding items in the card but are not checking out. You now come to a conclusion that there is some issue with user experience on your website, but what is it precisely? There exists many factors which could be affecting sales, such as the payment page is not using more secure https Web service, the payment options form does not work or an unexpected charged amount appears on the page. Hence, diagnostic analysis enables you to present a picture and the cause behind it, which is not apparent in the presented data. The following are the functions of diagnostic analytics:

- **Identifying the problem and events worth investigating:** Using results of descriptive analysis, the analyst must identify the areas which require further analysis and investigation since they are the ones which raise questions whose answers cannot be found by just looking at the data provided. It may be anything from falling sales to unexpected performance boost. Every one of these causes then can be analysed further using diagnostic analytics to find the root problems or causes.
- **Drilling into the analytics:** Once anomalies are distinct and recognised, the analyst must identify the data source which might be able to direct for the root cause of the anomalies. During this process, the analyst may have to look outside the selected data sets to find patterns and directions. This also may require pulling data from other data sources which can be used to identify correlation between data sets and checking if these correlations are casual in nature.
- **Identifying casual relationships:** To explain the cause of identified anomalies, unseen relationships are identified by closely observing the events. Techniques and concepts such as probability theory, time-series data analytics filtering and regression analysis can be implemented and prove useful for unravelling the true nature of data. Since data volume, variety and velocity have increased drastically in past years, manual methods, which were used by analysts for diagnostic analytics produced results that were highly dependent on abilities of the analyst. However,

even those analysts did not guarantee the consistency of results. Modern methods for diagnostic analytics use machine learning since machines are far more capable than humans in recognising and clustering patterns. An intelligent implementation and use of diagnostic analytics is imperative as it answers very specific questions such as 'How to avoid a given problem?' or finding ways to replicate a solution for other similar problems. Also, documentation of the diagnostic analysis must be done in a meaningful way which must state which issue was identified, what data sources were used to analyse and eliminate the issue and which casual relationships between data sets were identified during the analysis. The vitality of diagnostic analysis cannot be ignored as the data collected by IDC in the survey states that 25% of large organisations will have supplement data scientists by 2021 who will provide contextual data interpretation using qualitative research methods which would enable the organisations to dig deep inside the understandings of human's emotion, perception and stories of their world.

**Self Assessment Questions**

19. _____ analytics can be defined as a type of advanced analytics which is performed to answer more complex questions related to a project or event, such as 'Why did it happen?'
20. Diagnostic analytics can also be used to find the casual relationships between two or more data sets if the root cause is not detectable. (True/False)

## 8.9 SUMMARY

- Business analytics provides insight and value to a business, making it one of the most important IT functions for any running business.
- Descriptive analytics is the most essential type of analytics and establishes the framework for more advanced type of analytics.
- Statistics involves collecting, organising, analysing, interpreting and presenting data.
- A sample is a subset of a population.
- The process of sampling aims to obtain enough information to create a legal interpretation about a population.
- Central tendency is the measurement of a single value that attempts to describe a set of data by identifying the central position within that set of data.
- A commonly used measure of dispersion is the variance. Basically, variance is the squared deviations average of the observations from the mean.
- Standard deviation is usually easier to understand than variance because of similarity in its measurement units that are same as the data units.
- A z-score, or standardised value, provides a measure of the distance of the observation away from the mean, irrespective of the measurement units.
- Descriptive analytics analyses a database to provide information on the trends of past or current business events that can help managers, planners, leaders, etc., to develop a road map for future actions.

- Model comparison and improvement in business analytics involve assessing and refining predictive models continuously to enhance their accuracy, reliability, and alignment with evolving business.

## 8.10 KEY WORDS

- **Statistics:** It involves collecting, organising, analysing, interpreting and presenting data.
- **Population:** It consists of all items of interest for a particular decision or investigation.
- **Sample:** It is a subset of a population.
- **Central tendency:** It is the measurement of a single value that attempts to describe a set of data by identifying the central position within that set of data.
- **Mean:** It is the sum of the observations divided by the total number of observations.
- **Median:** It is the measure of location that specifies the middle value when the data is arranged from least to greatest.
- **Mode:** It is the observation/number/series that occurs the maximum number of times in a group.
- **Variance:** It is the squared deviations average of the observations from the mean.
- **Hyperparameter:** It is a configuration setting external to the model, influencing its training process and performance, determined before training.
- **Diagnostic analytics:** It is used to find the casual relationships between two or more data sets if the root cause is not detectable.

## 8.11 CASE STUDY: PNT MARKETING SERVICES USED PREDICTIVE ANALYTICS FOR EMAIL MARKETING CAMPAIGNS

PNT Marketing Services is an organisation that provides marketing and analytics services to its clients, most of which included digital marketing agencies and for-profit educational institutions. Advanced analytics was required to enhance their existing email and online marketing channels. The e-mail marketing predictive analytics was required to improve the lead generation for PNT's clients.

PNT was faced with the challenge of deploying a platform using which it could create and deploy various predictive models. It was required that the platform be easy to use for the business users and should not require statistical programming. In addition, it was necessary that the platform should also provide powerful and accurate models which could be deployed for their clients. PNT researched well to find out which platform could best serve its needs and it finalised on LityxIQ. PNT started using LityxIQ platform to support complex modelling requirements for its diverse clients.

PNT implemented LityxIQ in the following ways:

- **Data:** First of all, a data warehouse was constructed to track all the contacts and a year's history of emails that were sent. For each contact to whom an e-mail was

Notes

sent, various data were recorded and these included email clicks, opens, click-throughs, page visit patterns, conversion data and call centre data. All this data or information was further used to generate other RFM (recency, frequency, monetary) metrics. The RFM-type metrics include the number of contacts made in the last one year, time elapsed since last contact, and percentage of emails clicked out of the total emails sent. All this data was fed into the LityxIQ system to support advanced predictive analytics.

- **Predictive Analytics:** PNT started building various models in a rapid fashion using PredictIQ. Different models were built for different clients using the given client's data. PNT also updated the predictive models regularly because in today's business environment, the nature of market and email programs keep changing frequently. PredictIQ allowed PNT to test results of different models without having to do any coding.

As a result of implementing predictive analysis, PNT realised the following benefits:

- 116% increase in the click-through rate
- 57% increase in the click-to-lead rate
- Various predictive models could be deployed easily by PNT staff who were non-statisticians
- A weekly database of more than 5 million records was created to support the e-mail campaigns.

**Source:** *https://lityx.com/case-studies/predictive-analytics-email-campaigns/*

## QUESTIONS

1. What kind of challenge was PNT facing?

   (**Hint:** PNT was faced with the challenge of deploying a platform using which it could create and deploy various predictive models.)

2. Why was advanced analytics required in PNT?

   (**Hint:** Advanced analytics was required to enhance existing email and online marketing channels.)

3. What is the purpose of performing e-mail marketing predictive analytics?

   (**Hint:** The e-mail marketing predictive analytics was required to improve the lead generation for PNT's clients.)

4. Why did PNT choose LityxIQ?

   (**Hint:** PNT was faced with the challenge of deploying a platform using which it could create and deploy various predictive models. PNT researched well to find out which platform could best serve its needs and it finalised on LityxIQ.)

5. How did PNT use data for implementing its advanced predictive analytics?

   (**Hint:** A data warehouse was constructed to track all the contacts and a year's history of emails that were sent. For each contact to whom an email was sent, various data was recorded, which included e-mail clicks, opens, click-throughs, page visit patterns, conversion data and call centre data. All this data or information was further used to generate other RFM metrics.)

## 8.12 EXERCISE

1. Explain the importance of descriptive analytics for an organisation.
2. Discuss the fundamentals of descriptive statistics.
3. Describe the need of predictive analytics in an organisation.
4. Discuss predictive modelling with suitable examples.
5. Write a short note on prescriptive analytics.
6. Explain the contribution of model comparison and improvement to enhancing predictive models.

## 8.13 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Overview of Descriptive Analytics | 1. | b. Predictive analytics |
| | 2. | True |
| Understanding Descriptive Statistics | 3. | population |
| | 4. | sample |
| | 5. | True |
| Predictive Analytics | 6. | Descriptive |
| | 7. | True |
| Types of Predictive Modelling | 8. | a. Predictive modelling |
| | 9. | False |
| | 10. | logic-driven |
| | 11. | data-driven |
| Model Comparison and Improvement | 12. | d. Mean Absolute Error (MAE) |
| | 13. | False |
| | 14. | New data |
| Prescriptive Analytics | 15. | optimisation |
| | 16. | True |
| | 17. | d. All of these |
| | 18. | True |
| Diagnostic Analytics | 19. | Diagnostic |
| | 20. | True |

## 8.14 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Bari, A., Chaouchi, M. and Jung, T. (n.d.). *Predictive analytics for dummies*.
- Farmer, L. and Safer, A. (n.d.). *Library improvement through data analytics*.

Notes

## E-REFERENCES


- Beattie, A. (2018). *Descriptive Analytics*. [online] Investopedia. Available at: https://www.investopedia.com/terms/d/descriptive-analytics.asp [Accessed 4 Dec. 2018].
- Sas.com. (2018). *Predictive Analytics: What it is and why it matters*. [online] Available at: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html [Accessed 4 Dec. 2018].
- SearchBusinessAnalytics. (2018). *Prescriptive analytics takes analytics maturity model to a new level*. [online] Available at: https://searchbusinessanalytics.techtarget.com/feature/Prescriptive-analytics-takes-analytics-maturity-model-to-a-new-level [Accessed 4 Dec. 2018]

CHAPTER 9

# Data Representation and Visualisation

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the ways of representing visual data
- Discuss the techniques used for visual data representation
- Explain the different types of data visualisation
- Define the applications of data visualisation
- Explain the significance of visualising Big Data
- Describe the tools used in data visualisation
- Discuss how the data visualisation is useful for managers
- Elucidate the concept pf data is visualised in Excel

## 9.1 INTRODUCTION

In the previous chapter, you studied about the descriptive statistics and predictive modelling. The chapter explained in detail about logic-driven models and data-driven models of predictive modelling. You also studied about prescriptive analytics.

Data is everywhere, but to represent the data in front of users in such a way that it communicates all the necessary information effectively is important. Data visualisation can be understood as a technique which can be used to communicate data or information by transforming it into pictorial or graphical format. The main purpose of data visualisation is to make users understand the information clearly and efficiently. It is one of the important steps in data analysis or data science.

Depending upon the complexity of data and the aspects from which it is analysed, visuals can vary in terms of their dimensions (one-/two-/multi-dimensional) or types, such as temporal, hierarchical, network, etc. All these visuals are used for presenting different types of datasets. Different types of tools are available in the market for visualising data. But what is the use of data visualisation in Big Data? Is it necessary to use it? Let us first track down the real meaning of visualisation in the context of Big Data analytics.

The chapter begins by explaining the concept of visual data and techniques used for visual representation. It also covers the types and applications of data visualisation. Next, it covers about various types of tools using which data or information can be presented in a visual format. Moreover, the chapter explains the concept of visualising and exploring data in Excel.

## 9.2 WAYS OF REPRESENTING VISUAL DATA

The data is first analysed and then the result of that analysis is visualised in different ways as discussed above. There are two ways to visualise a data—infographics and data visualisation:

- Infographics are the visual representations of information or data rapidly and accurately. The use of colourful graphics in drawing charts and graphs helps in improving the interpretation of a given data.

NOTES

❍ Data visualisation is a different approach from the infographics. It is the study of representing data or information in a visual form. With the advancement of digital technologies, the scope of multimedia has increased manifold. Visuals in the form of graphs, images, diagrams, or animations have completely proliferated the media industry and the Internet. It is an established fact that the human mind can comprehend information more easily if it is presented in the form of visuals. Instructional designers focus on abstract and model-based scientific visualisations to make the learning content more interesting and easy to understand. Nowadays, scientific data is also presented through digitally constructed images. These images are generally created with the help of computer software.

Figure 1 shows an example of infographics:

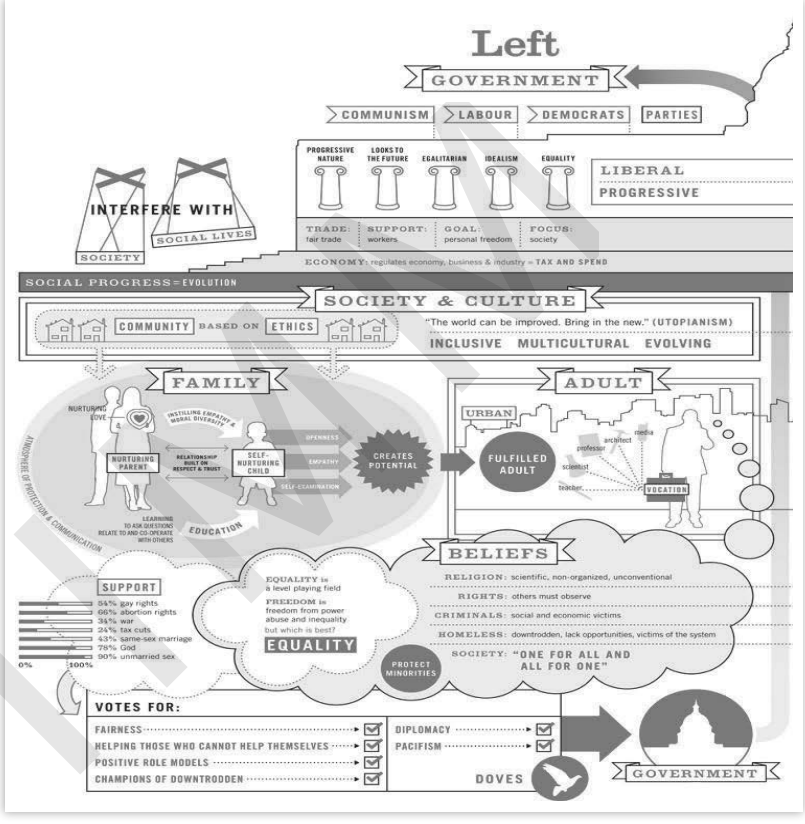



FIGURE 1: An Example of Infographics

**Source:** *http://www.jackhagley.com/What-s-the-difference-between-an-Infographic-and-a-Data-Visualisation*

❍ Visualisation is an excellent medium to analyse, comprehend and share information. Let us see why:

- Visual images help to transmit a huge amount of information to the human brain at a glance.
- Visual images help in establishing relationships and distinction between different patterns or processes easily.
- Visual interpretations help in exploring data from different angles, which help gain insights.

- Visualisation helps in identifying problems and understanding trends and outliers.

❍ Data can be classified on the basis of the following three criteria irrespective of whether it is presented as data visualisation or infographics:

- **Method of creation**: It refers to the type of content used while creating any graphical representation.
- **Quantity of data displayed**: It refers to the amount of data which is represented. For example, geographical map, companies financial data, etc.
- **Degree of creativity applied**: It refers to the extent to which the data is created graphically or designed in a colourful way or it is just showing some important data in black and white diagrams.

❍ On the basis of above evaluation, we can understand which is the correct form of representation for a given data type. Let us discuss the various content types:

- **Graph**: A representation in which X and Y axes are used to depict the meaning of the information.
- **Diagram**: A two-dimensional representation of information to show how something works.
- **Timeline**: A representation of important events in a sequence with the help of self-explanatory visual material.
- **Template**: A layout is a design for presenting information.
- **Checklist**: A list of items for comparison and verification.
- **Flowchart**: A representation of instructions which shows how something works or a step-by-step procedure to perform a task.

**SELF ASSESSMENT QUESTIONS**

1. It is a representation of important events in a sequence with the help of self-explanatory visual material.

   a. Timeline b. Template

   c. Flowchart d. Checklist

2. _____ are the visual representations of information or data rapidly and accurately.

## 9.3 TECHNIQUES USED FOR VISUAL DATA REPRESENTATION

Data can be presented in various visual forms, which include simple line diagrams, bar graphs, tables, matrices, etc. Some techniques used for a visual presentation of data are as follows:

❍ **Direct Volume Rendering (DVR):** It is a method used for obtaining a 2D projection for a 3D dataset. A 3D record is projected in a 2D form through DVR for a clearer and more transparent visualisation.

Figure 2 shows a 2D DVR of a 3D image:

Figure 2: 2D Image DVR

- **Isoline**: It is a 2D data representation of a curved line that moves constantly on the surface of a graph. Figure 3 shows a set of isolines:

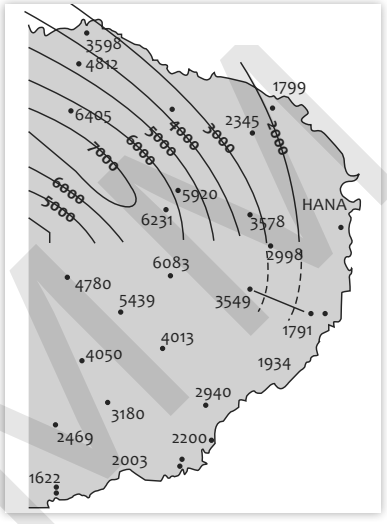



Figure 3: Isolines

- **Isosurface:** It is a 3D representation of an isoline. Isosurfaces are created to represent points that are bounded in a volume of space by a constant value, that is, in a domain that covers 3D space. Figure 4 shows how isosurfaces look like:

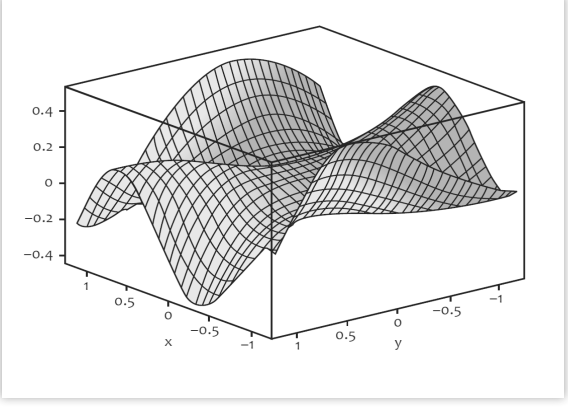



Figure 4: Isosurfaces

- **Streamline:** It is a field line that results from the velocity vector field description of the data flow. Figure 5 shows a set of streamlines:

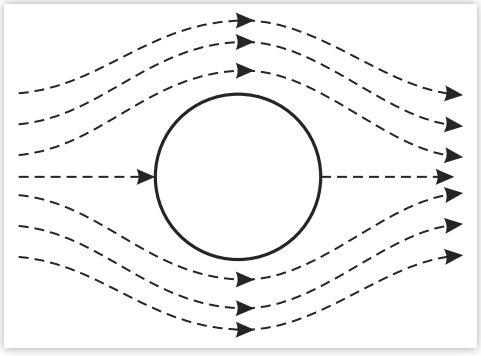


Figure 5: Streamlines

- **Map:** It is a visual representation of locations within a specific area. It is depicted on a planar surface.
- **Parallel coordinate plot:** It is a visualisation technique of representing multidimensional data. In parallel coordinate plot, each row of the data table is mapped as a line or profile. Each attribute related to a row is denoted by a point on the line. Figure 6 shows a parallel coordinate plot:

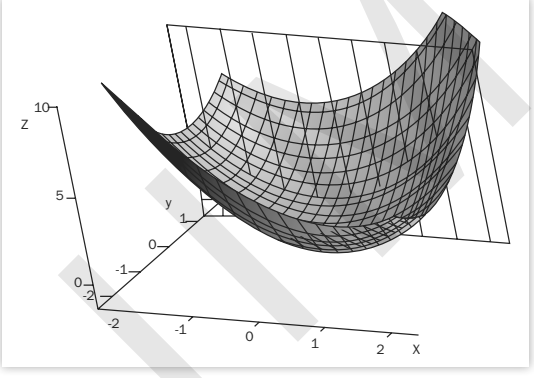



Figure 6: Parallel Coordinate Plot

- **Venn diagram:** It is used to represent logical relations between finite collections of sets. Figure 7 shows a Venn diagram for a set of relations:

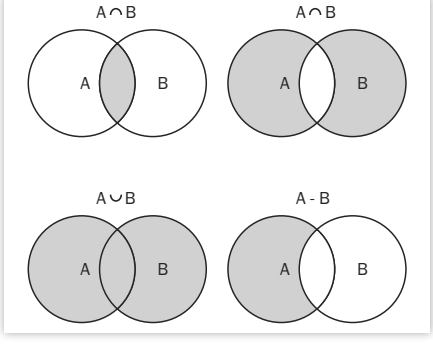



Figure 7: Venn Diagrams

Notes

- **Timeline:** It is used to represent a chronological display of events. Figure 8 shows an example of a timeline for some critical event sets:

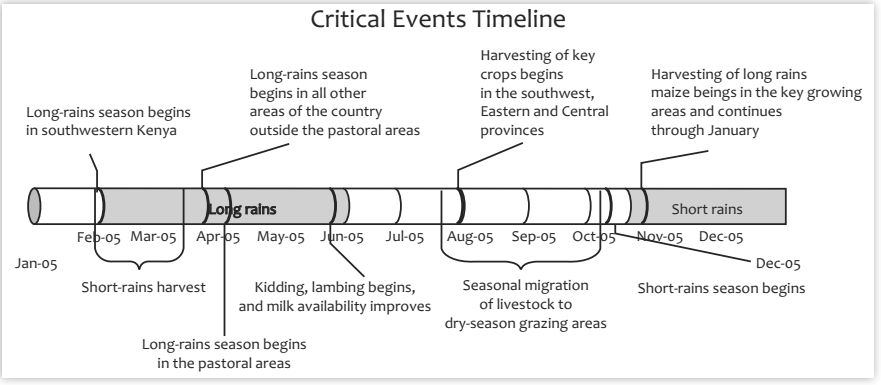



Figure 8: Timeline for Some Critical Events

- **Euler diagram:** It is a representation of the relationships between sets. Figure 9 shows an example of an Euler diagram:

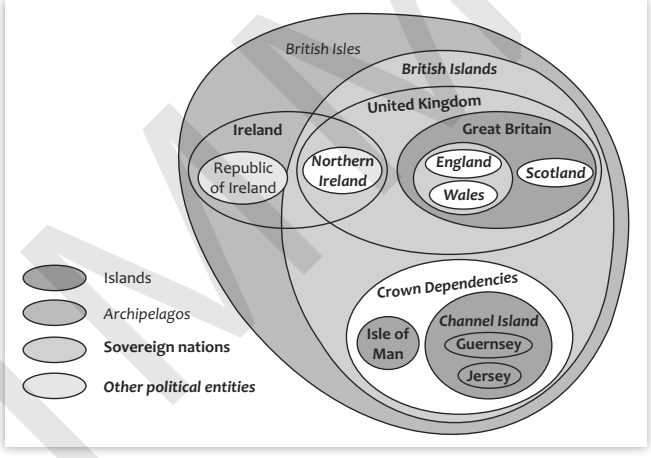



Figure 9: Euler Diagram

- **Hyperbolic trees:** This method is based on hyperbolic geometry and used for visualising information and drawing graphs. In other words, it is used to display hierarchical data in the form of tree. Figure 10 shows a hyperbolic tree:

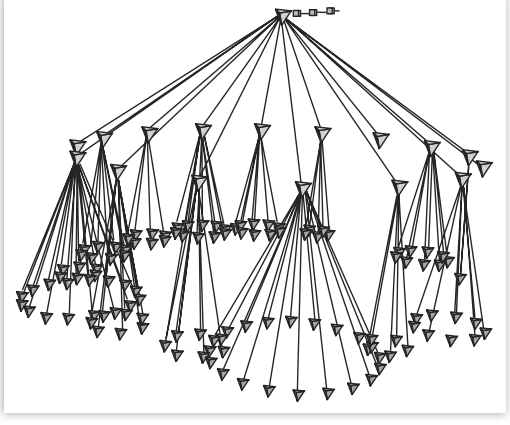


Figure 10: Hyperbolic Tree

- **Cluster diagram:** It represents a cluster, such as a cluster of astronomic entities. Figure 11 shows a cluster diagram:

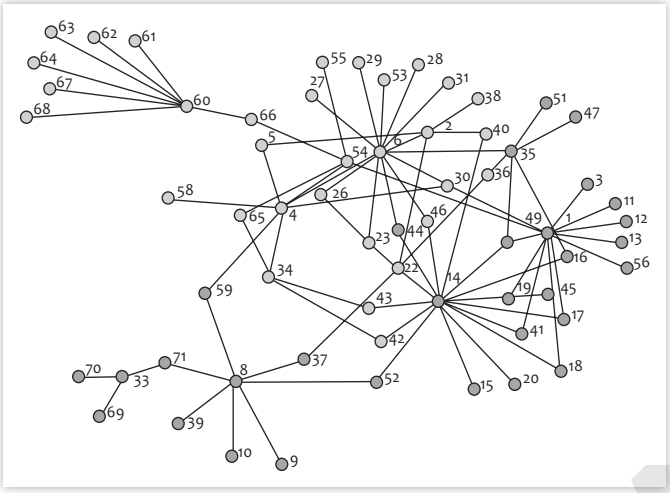


Figure 11: Cluster Diagram

**Self Assessment Questions**

3. _____ is a 2D data representation of a curved line that moves constantly on the surface of a graph.
4. It is a visual representation of locations within a specific area.
   a. Streamline
   b. Isosurface
   c. Venn diagram
   d. Map
5. Timeline is used to represent a chronological display of events. (True/False)

## 9.4 TYPES OF DATA VISUALISATION

You know that data can be visualised in many ways, such as in the forms of 1D, 2D, or 3D structures. Table 1 briefly describes the different types of data visualisation:

Table 1: Data Visualisation Types

| Name | Description | Tool |
|---|---|---|
| 1D/Linear | A list of items organised in a predefined manner | Generally, no tool is used for 1D visualisation |
| 2D/Planar | Choropleth, cartogram, dot distribution map and proportional symbol map | GeoCommons, Google Fusion Tables, Google Maps API, Polymaps, Many Eyes, Google Charts and Tableau Public |
| 3D/Volumetric | 3D computer models, surface rendering, volume rendering, and computer simulations | AC3D, AutoQ3D, TrueSpace |
| Temporal | Timeline, time series, Gantt chart, sanky diagram, alluvial diagram, and connected scatter plot | TimeFlow, Timeline JS, Excel, Timeplot, TimeSearcher, Google Charts, Tableau Public and Google Fusion Tables |

Notes

| Name | Description | Tool |
|---|---|---|
| Multidimensional | Pie chart, histogram, tag cloud, bubble cloud, bar chart, scatter plot, heat map, etc. | Many Eyes, Google Charts, Tableau Public and Google Fusion Tables |
| Tree/Hierarchical | Dendogram, radial tree, hyperbolic tree and wedge stack graph | d3, Google Charts, and Network Workbench/Sci2 |
| Network | Matrix, node link diagram, hive plot and tube map | Pajek, Gephi, NodeXL, VOSviewer, UCINET, GUESS, Network Workbench/Sci2, sigma.js, d3/Protovis, Many Eyes and Google Fusion Tables |

As shown in Table 1, the simplest type of data visualisation is 1D representation and the most complex data visualisation is the network representation.

**Self Assessment Questions**

6. A list of items organised in a predefined manner is called
   a. 2D/Planar
   b. 1D/Linear
   c. Multidimensional
   d. Network
7. The most complex data visualisation is the _____ representation.

**Activity**

Search and prepare a report on animated charts that are used for data visualisation.

## 9.5 APPLICATIONS OF DATA VISUALISATION

Data visualisation tools and techniques are used in various applications. Some of the areas in which we apply data visualisation are as follows:

- **Education:** Visualisation is applied to teach a topic that requires simulation or modelling of any object or process. Have you ever wondered how difficult it would be to explain any organ or organ system without any visuals? Organ system or structure of an atom is best described with the help of diagrams or animations.
- **Information:** Visualisation is applied to transform abstract data into visual forms for easy interpretation and further exploration.
- **Production:** Various applications are used to create 3D models of products for better viewing and manipulation.
- **Science:** Every field of science including fluid dynamics, astrophysics, and medicine use visual representation of information. Isosurfaces and direct volume rendering are typically used to explain the scientific concepts.
- **Systems visualisation:** Systems visualisation is a relatively new concept that integrates visual techniques to better describe complex systems.
- **Visual communication:** Multimedia and entertainment industry use visuals to communicate their ideas and information.

- **Visual analytics:** It refers to the science of analytical reasoning supported by the interactive visual interface. The data generated by social media interaction is interpreted using visual analytics techniques.

**Exhibit**

**"DATA VISUALISATION APP" BY UNICEF**

The UNICEF has released a 'data visualisation application' which provides a user-friendly visual representation of Indian education scenario's complex analytics. The application uses government database for schools, known as UDISE (Unified District Information System for Education) and NAS (National Assessment survey). This can be used as a visual tool by government officials, policy makers, research scholars to resolve issues and make the education system successful and future ready.

**Self Assessment Questions**

8. _____ is a relatively new concept that integrates visual techniques to better describe complex systems.
9. Isosurfaces and direct volume rendering are typically used to explain scientific concepts. (True/False)
10. It refers to the science of analytical reasoning supported by the interactive visual interface.
    a. Education
    b. Production
    c. Visual analytics
    d. Science

## 9.6 VISUALISING BIG DATA

Visual analysis of data is not a new thing. For years, statisticians and analysts have been using visualisation tools and techniques to interpret and present the outcomes of their analyses.

Almost every organisation today is struggling to tackle the huge amount of data pouring in every day. Data visualisation is a great way to reduce the turn-around time consumed in interpreting Big Data. Traditional visualisation techniques are not efficient enough to capture or interpret the information that Big Data possesses. For example, such techniques are not able to interpret videos, audios and complex sentences. Apart from the type of data, the volume and speed with it is generating pose a great challenge in data visualisation. Most of the traditional analytics techniques are unable to cater to any of these problems.

Big Data comprises both structured as well as unstructured forms of data collected from various sources. Heterogeneity of data sources, data streaming and real-time data are also difficult to handle by using traditional tools. Traditional tools are developed by using relational models that work best on static interaction. Big Data is highly dynamic in function and therefore, most traditional tools are not able to generate quality results. The response time of traditional tools is quite high, making it unfit for quality interaction.

Notes

### 9.6.1 DERIVING BUSINESS SOLUTIONS

Nowadays, almost every company is working with Big Data and facing the following challenges:

- Most data is in unstructured form
- Data is not analysed in real time
- The amount of data generated is huge
- There is a lack of efficient tools and techniques

Considering all these factors, IT companies are focusing more on research and development of robust algorithms, software and tools to analyse the data that is scattered in the Internet space. Tools such as Hadoop are providing the state-of-the-art technology to store and process Big Data. Analytical tools are now able to produce interpretations on smartphones and tablets because of advanced methods of dimensionality reduction, advanced algorithms for various data (such as streaming data) and advanced analytics. It is possible because of advanced methods of dimensionality reduction, advanced algorithms for various data (such as streaming data) and advanced analytics that is enabling business owners and researchers to explore data for finding out trends and patterns.

### 9.6.2 TURNING DATA INTO INFORMATION

The most exciting part of any analytical study is to find useful information from a plethora of data. Visualisation facilitates identification of patterns in the form of graphs or charts, which in turn helps to derive useful information.

Visual data mining also works on the same principle as simple data mining; however, it involves the integration of information visualisation and human–computer interaction. Visualisation of data produces cluttered images that are filtered with the help of clutter-reduction techniques. Uniform sampling and dimension reduction are two commonly used clutter-reduction techniques.

Visual data reduction process involves automated data analysis to measure density, outliers, and their differences. These measures are then used as quality metrics to evaluate data-reduction activity. Visual quality metrics can be categorised as:

- Size metrics (e.g. number of data points)
- Visual effectiveness metrics (e.g. data density, collisions)
- Feature preservation metrics (e.g. discovering and preserving data density differences)

In general, we can conclude that a visual analytics tool should be:

- Simple enough so that even non-technical users can operate it
- Interactive to connect with different sources of data
- Competent to create appropriate visuals for interpretations
- Able to interpret Big Data and share information

Apart from representing data, a visualisation tool must be able to establish links between different data values, restore the missing data, and polish data for further analysis.

**Self Assessment Questions**

11. _____ tools are now able to produce interpretations on smartphones and tablets.
12. Tools such as Hadoop are providing the state-of-the-art technology to store and process Big Data. (True/False)
13. Visual quality metrics can be categorised as Size metrics. (True/False)
14. _____ involves the integration of information visualisation and human–computer interaction.

**Activity**

Suppose you are a data analyst in an organisation. What are the challenges you have faced while analysing Big Data using traditional analytics techniques? What techniques will you use to overcome these challenges?

## 9.7 TOOLS USED IN DATA VISUALISATION

Some generalised visualisation tools are listed as follows:

- **Excel:** It is a tool that is used for data analysis. It helps you to track and visualise data for deriving better insights. This tool provides various ways to share data and analytical conclusions within and across organisations.
- **Last.Forward:** It is open-source software provided by last.fm for analysing and visualising social music network.
- **Digg.com:** Digg.com provides some of the best Web-based visualisation tools.
- **Pics:** This tool is used to track the activity of images on the website.
- **Arc:** It is used to display the topics and stories in a spherical form. Here, a sphere is used to display stories and topic, and bunches of stories are aligned at the outer circumference of sphere. Larger stories have more diggs. The arc becomes thicker with the number of times users dig the story.
- **Google Charts API:** This tool allows a user to create dynamic charts to be embedded in a Web page. A chart obtained from the data and formatting parameters supplied in a HyperText Transfer Protocol (HTTP) request is converted into a Portable Network Graphics (PNG) image by Google to simplify the embedding process.
- **TwittEarth:** TwittEarth is tool which is capable of mapping location of tweets from all over the globe on a 3d representation of globe and show it. It is an effort to improve social media visualisation and provide a global image mapping in tweets.
- **Tag Galaxy:** Tag Galaxy provides a stunning way of finding a collection of Flickr images. It is an unusual site which provides search tool which makes the online combing process a memorable visual experience. If you want to search a picture,

Notes

you have to enter a tag of your choice and it will find the picture. The central (core) star contains all the images directly relating to the initial tag and the revolving planets consist of similar or corresponding tags. Click on a planet and additional sub-categories will appear. Click on the central star and Flickr images gather and land on a gigantic 3D sphere.

- **D3:** With D3, you get the ability to attach DOM (Document Object Model) with random data and then apply transformations which are data driven, on the document. Also, you can use the same data to design and develop an interactive SVG have features like smooth transition and interactions.
- **RootzMap mapping the Internet:** It is a tool to generate a series of maps on the basis of the datasets provided by the National Aeronautics and Space Administration (NASA).

### 9.7.1 OPEN-SOURCE DATA VISUALISATION TOOLS

We already know that Big Data analytics requires the implementation of advanced tools and technologies. Due to economic and infrastructural limitations, every organisation cannot purchase all the applications required for analysing data. Therefore, to fulfill their requirement of advanced tools and technologies, organisations often turn to open-source libraries. These libraries can be defined as pools of freely available applications and analytical tools. Some examples of open-source tools available for data visualisation are VTK, Cave5D, ELKI, Tulip, Gephi, IBM OpenDX, Tableau Public and Vis5D.

Open-source tools are easy to use, consistent, and reusable. They deliver high-quality performance and are compliant with the Web as well as mobile Web security. In addition, they provide multichannel analytics for modeling as well as customised business solutions that can be altered with changing business demands.

**Self Assessment Questions**

15. _____ helps you to track and visualise data for deriving better insights.
16. Google Charts API tool allows a user to create dynamic charts to be embedded in a Webpage.(True/False)
17. It is an effort to improve social media visualisation and provide a global image mapping in tweets.
    a. TwittEarth
    b. Tag Galaxy
    c. D3
    d. Arc

**Activity**

Find the information about the following open source data visualisation tools:

- VTK
- IBM OpenDX
- ELKI
- Tableau Public

## 9.8 DATA VISUALISATION FOR MANAGERS

As a product manager, one must understand the power of data visualisation. Analysing the amassed data, drawing conclusions from it and creating a compelling data visualisation is the art a manager should know.

Data visualisation can be used as a method to convey data or information by transforming it visually, so that it is more accessible by the people receiving it. Best visual representations give the reader a clearer idea and let them draw conclusion based on data that they might otherwise have missed. Also, presenting data in too much detail may confuse the consumers and they may lose interest. So, a strike of balance is needed while modelling data visualisation.

Organisations can also assess their direction of business using data visualisation. For example you may find yourself looking for answers that how you can boost sales in a particular demography or hoe you can cut down the operational costs. You can authenticate your future business moves by backing it up with the data you have.

In 2013, a survey was conducted by Aberdeen group in which it was found that the rate of finding relevant information by managers in organisation increased by more than 28% using data visualisation tools as compared to their peers who relied simply on dashboards and managed reporting.

**Self Assessment Questions**

18. Organisations can also assess their direction of business using data visualisation. (True/False)
19. Data visualisation can be used as a method to convey data or information by transforming it visually, so that it is more accessible by the people receiving it. (True/False)

## 9.9 VISUALISING AND EXPLORING DATA IN EXCEL

Data visualisation is the method of depicting data (typically in larger quantities) in graphical or visual form. The researchers observed that data visualisation improves decision making, provides managers with better analytic capabilities that reduce the dependence on IT professionals and improves collaboration and information sharing. Raw data is important, particularly when one needs to identify accurate values or compare individual numbers. However, it is quite difficult to identify trends, patterns and find exceptions, or compare groups of data in tabular form. The human brain does a surprisingly good job in processing visual information—if presented in an effective way.

Data visualisation is also important both for building decision models and for interpreting their results. To identify the appropriate model to use, we would normally have to collect and analyse data to determine the type of relationship (for example, linear, non-linear) and estimate the values of the parameters in the model. Visualising the data will help to identify the proper relationship and use the

Notes

appropriate data analysis tool. Furthermore, complex analytical models often yield complex results. Visualising the results helps in understanding and gaining insight about model output and solutions.

## 9.9.1 DASHBOARDS

Making data visible and accessible to employees at all levels is a hallmark of effective modern organisations. A dashboard is a visual picture of a group of specific business measures. It is similar to the dashboard of an automotive, such as a car, which displays fuel level, speed, seat signs, temperature, and so on. Dashboards deliver important key synopses of valuable business data to efficiently manage a business function or process.

## 9.9.2 COLUMN AND BAR CHARTS

MS Excel refers to the vertical bar charts as column and horizontal bar charts as bar charts. Column and bar charts are valuable for equating categorical or series specific data, for demonstrating differences between value sets and for displaying percentages or proportions of a whole.

Figure 12 shows column and bar charts:


Figure 12: Column and Bar Chart

**Source:** *https://www.aploris.com/support/documentation/bar-and-line-charts*

## 9.9.3 DATA LABELS AND DATA TABLES CHART OPTIONS

MS Excel provides options for including the numerical data on which charts are based within the charts. Data labels can be added to chart elements to show the actual value of bars. Data tables can also be added; these are usually better than data labels, which can get quite messy. Both can be added from the Add Chart Element Button in the Chart Tools Design tab, or also from the Quick Layout button, which provides standard design options.

Figure 13 shows data labels and data tables chart:

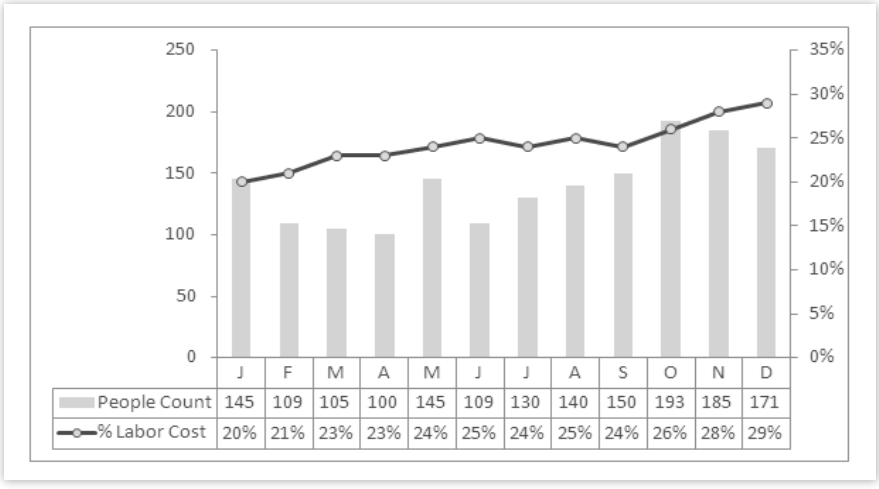



Figure 13: Data Labels and Data Tables Chart

**Source:** *http://datapigtechnologies.com/blog/index.php/the-trouble-with-chart-data-tables/*

## 9.9.4 LINE CHARTS

Line charts are a useful way of displaying data for a given period. You may enter multiple series of data in line charts; however, it can become difficult to interpret if the size of data values differs exponentially. In such a case, it would be advisable to create individual charts for different data series. Figure 14 shows line charts:

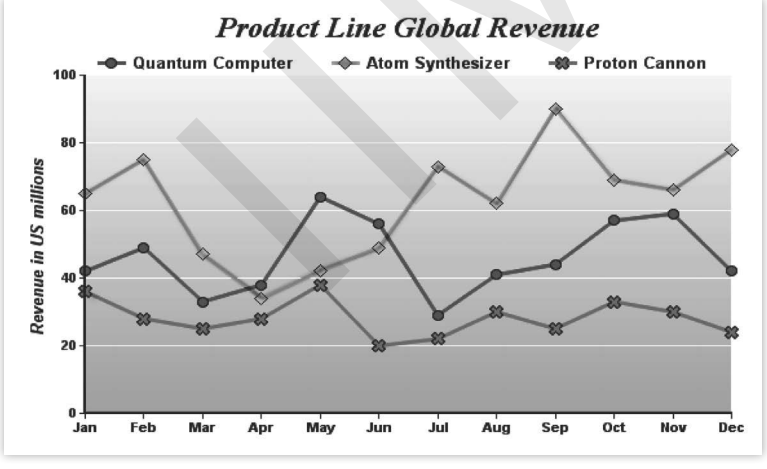



Figure 14: Line Charts

**Source:** *http://www.advsofteng.com/gallery_line.html*

## 9.9.5 PIE CHARTS

For many types of data, we are interested in understanding the relative proportion of each data source to the total. A pie chart shows this by dividing a circle into pie-shaped areas displaying the relative part. New age 3D pie charts can get confusing at times because of their narrow representation in case of huge data variables. This is because the third dimension also represents something especially on a coordinate

Notes

graph. Hence, pie charts are preferred only in two dimensional form for effective and simpler data representation. Figure 15 displays a pie chart:

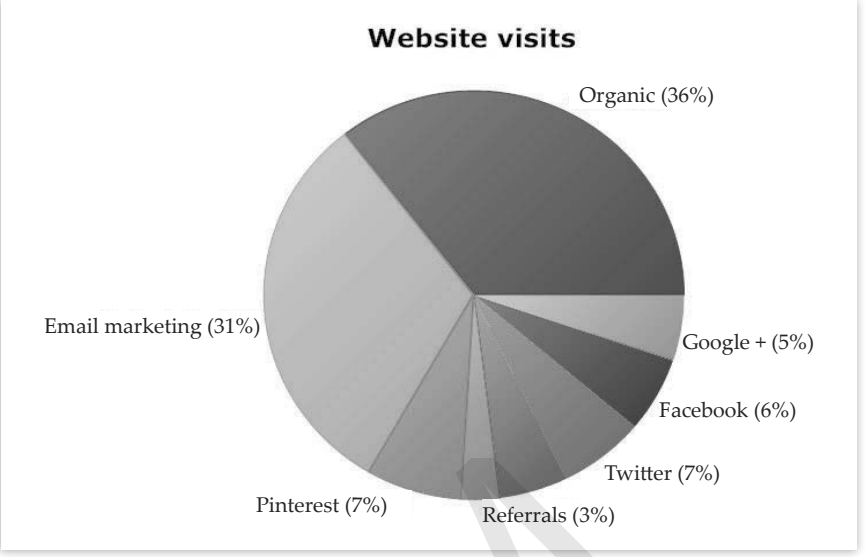



Figure 15: Pie Charts

**Source:** *http://www.f1f9.com*

## 9.9.6 SCATTER CHART

Scatter charts demonstrate the connection between two variables. To create a scatter chart, we require variable pairs and observations related to them. For example, students in a class might have grades for both a midterm and a final exam. Figure 16 shows a scatter chart:

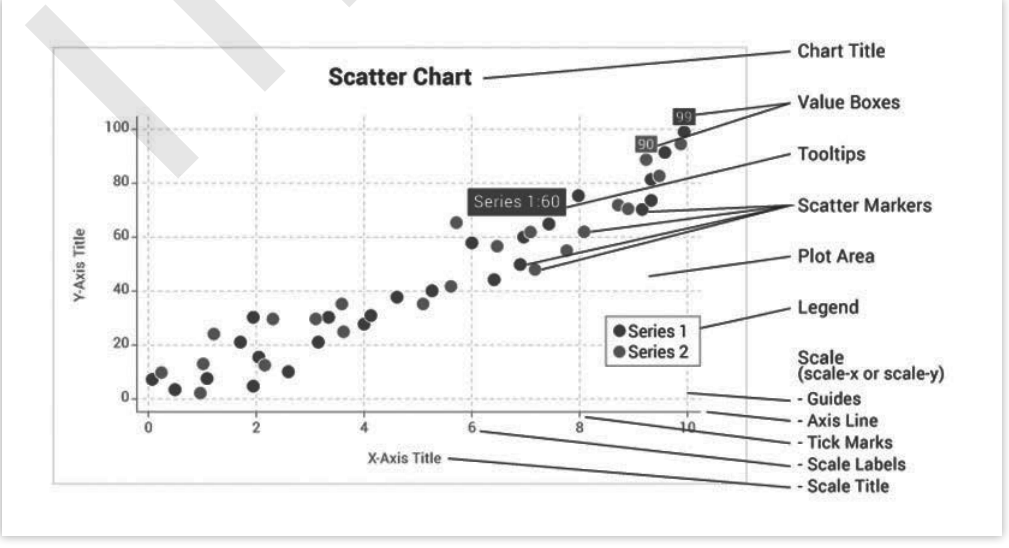



Figure 16: Scatter Chart

**Source:** *https://www.zingchart.com/docs/chart-types/scatter-plots/*

## 9.9.7 BUBBLE CHARTS

A bubble chart is a chart related to scatter chart, in which the data marker size corresponds to a third variable; thus, it is a method to display three variables in 2D space. Figure 17 shows a bubble chart:

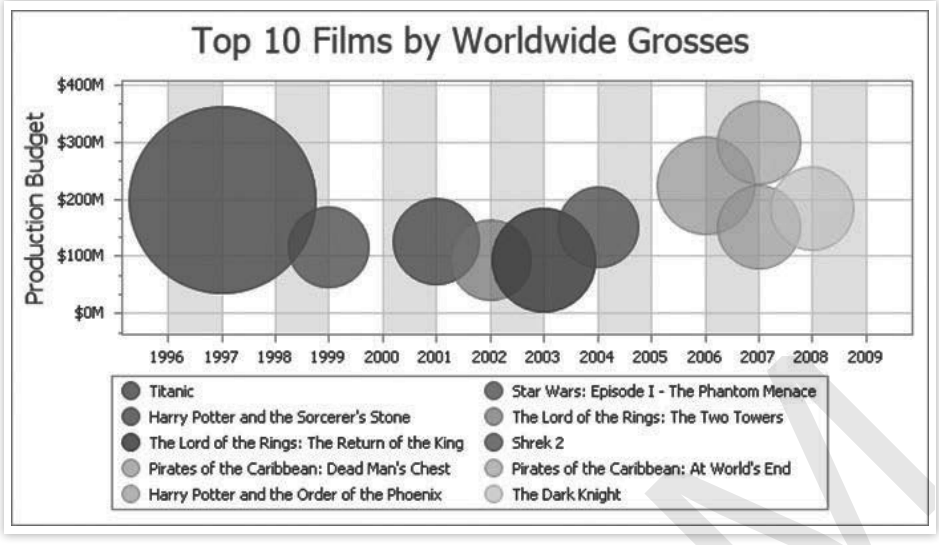



FIGURE 17: Displaying Bubble Charts

**Source:** *https://community.devexpress.com/blogs/ctodx/archive/2008/10/28/dxperience-v2008-vol-3-bubble-charts-for-winforms-and-asp-net.aspx*

## 9.9.8 MISCELLANEOUS EXCEL CHARTS

Excel provides several additional charts for special applications. These additional types of charts (including bubble charts) can be selected and created from the Other Charts button in the Excel ribbon. These include the following:

- A stock chart allows you to plot stock prices, such as the daily high, low and close. It may also be used for scientific data such as temperature changes.
- A surface chart shows 3-D data.
- A doughnut chart is similar to a pie chart but can contain more than one data series.
- A radar chart allows you to plot multiple dimensions of several data series.

## 9.9.9 PARETO ANALYSIS

Pareto analysis is a term named after Vilfredo Pareto, an Italian economist. In 1906, he realised that a large portion of the total wealth is held by a comparatively small number of the people in Italy. The Pareto principle is often seen in many business situations. For example, higher percentage of sales may come usually from a small percentage of customers, a higher percentage of defects originate from relatively smaller batches of the product or a high percentage of stock value belongs to a small percentage of selective items.

Notes

**Self Assessment Questions**

20. A bar chart depicts data in the form of horizontal bars. (True/False)
21. A _____ is used to show the relationship between numeric values in several data series.
22. It refers to a type of chart that displays a surface which is three dimensional (3D) and joins a group of data points.
    a. Surface chart
    b. Bubble chart
    c. Area chart
    d. Pie chart

**Activity**

Prepare a report on data visualisation tools available on the Web other than the tools discussed in the chapter.

## 9.10 SUMMARY

- Data visualisation can be understood as a technique which can be used to communicate data or information by transforming it into pictorial or graphical format.
- Data visualisation represents the data as visual objects, with the help of visual aids such as graphs, bar, histograms, tables, pie charts, mind maps, etc.
- Visualisation facilitates identification of patterns in the form of graphs or charts, which in turn helps to derive useful information.
- Data visualisation can be used as a method to convey data or information by transforming it visually, so that it is more accessible by the people receiving it.
- Infographics are the visual representations of information or data rapidly and accurately.
- Direct Volume Rendering (DVR) is a method used for obtaining a 2D projection for a 3D dataset.
- Isoline is a 2D data representation of a curved line that moves constantly on the surface of a graph.
- Venn diagram is used to reprsent logical relations between finite collections of sets.
- Visual analytics refers to the science of analytical reasoning supported by the interactive visual interface.
- Excel is a tool that is used for data anlysis. It helps you to track and visualise data for deriving better insights.
- Google Charts API is a tool that allows a user to create dynamic charts to be embedded in a Web page.
- TwittEarth is tool which is capable of mapping location of tweets from all over the globe on a 3rd representation of globe and show it.
- RootzMap mapping the interest is a tool in gernate a series of maps on the basis of the datasets provided by the National Aeronautics and Space Administration (NASA).

- A dashboard is a visual picture of a group of specific business measures.
- Column and bar charts are valuable for equating categorical or series specific data, for demonstrating differences between value sets and for displaying percentages or proportions of a whole.
- Data labels can be added to chart elements to show the actual value of bars.

## 9.11 KEY WORDS

- **Graph:** A representation in which X and Y axes are used to depict the meaning of the information.
- **Diagram:** A two-dimensional representation of information to show how something works.
- **Timeline:** A representation of important events in a sequence with the help of self-explanatory visual material.
- **Template:** A layout is a design for presenting information.
- **Direct Volume Rendering (DVR):** It is a method used for obtaining a 2D projection for a 3D dataset. A 3D record is projected in a 2D form through DVR for a clearer and more transparent visualisation.
- **Parallel coordinate plot:** It is a visualisation technique of representing multidimensional data.
- **Hyperbolic trees:** They represent graphs that are drawn using the hyperbolic geometry.
- **Systems visualisation:** Systems visualisation is a relatively new concept that integrates visual techniques to better describe complex systems.
- **Visual communication:** Multimedia and entertainment industry use visuals to communicate their ideas and information.
- **Visual analytics:** It refers to the science of analytical reasoning supported by the interactive visual interface. The data generated by social media interaction is interpreted using visual analytics techniques.
- **Excel:** It is a new tool that is used for data analysis. It helps you to track and visualise data for deriving better insights. This tool provides various ways to share data and analytical conclusions within and across organisations.
- **Last.Forward:** It is open-source software provided by last.fm for analysing and visualising social music network.
- **Digg.com:** Digg.com provides some of the best Web-based visualisation tools.
- **Pics:** This tool is used to track the activity of images on the website.
- **Isoline:** It is a 2D data representation of a curved line that moves constantly on the surface of a graph.
- **Streamline:** It is a field line that results from the velocity vector field description of the data flow.
- **Map:** It is a visual representation of locations within a specific area.
- **Venn diagram:** It is used to represent logical relations between finite collections of sets.

Notes

- **Euler diagram:** It is a representation of the relationships between sets.
- **Cluster diagram:** It represents a cluster, such as a cluster of astronomic entities.
- **Visual data reduction:** It is a process which involves automated data analysis to measure density, outliers, and their differences.
- **Scatter charts:** These demonstrate the connection between two variables.

## 9.12 CASE STUDY: SAVYVA CUSTOMISES ITS BI TOOL

SAVYVA GmbH is a German organisation that provides different services to its small and mid-sized client organisations. The services pertain to management of risks, transfer pricing and international tax. It has been able to provide its clients with excellent control over data and processes because it has been able to develop a unique and business intelligent Operational Transfer Pricing and Tax Technology platform CROSSVIEW, which has been powered by Dundas BI, a powerful BI tool. CROSSVIEW provides in-depth insights into their cross-border business and financial transaction data. This platform also provides views from multiple dimensions.

During its course of business, SAVYVA realised that there are various complexities in dealing with transfer pricing, such as varied processes, managing large quantities of data and making all such data more actionable. SAVYVA had the expertise pertaining to its business domain and they wanted to build a new software solution either from scratch or by creating it using some existing BI tool as a base. So, SAVYVA started looking for a partner who had expertise in the multi-dimensional analysis. SAVYVA looked for various tools which were quick and efficient and able to do advanced calculations. However, they did not have the extensibility and permissibility desired by SAVYVA.

SAVYVA wanted to build or acquire a BI tool that possessed quantitative analysis capabilities and a solution in which the organisation can embed its own components as well. Advanced visualisations and system integration ability were crucial for SAVYVA in order that all the business processes could be managed seamlessly using a single application. For this purpose, they decided to tie up with Dundas BI as it allowed system integration. The combination of Dundas BI platform along with the system components was developed into a product called CROSSVIEW.

CROSSVIEW platform was developed by SAVYVA team which included professionals from diverse fields, such as tax technology, Business Intelligence, SQL Server implementation, transfer pricing, multi-dimensional modeling, IT and database. SAVYVA illustrated a few reasons for choosing Dundas BI as its base platform. They are as follows:

- **Platform extensibility:** SAVYVA used open APIs for achieving their desired level of customisation and they modified it as per their specific needs. As an example, text narrative reports creation was enabled to satisfy advanced documentation and regulatory requirements.
- **Enterprise level governance:** Dundas had in-built support for configuring multiple projects on a single deployment, which enabled SAVYVA to provide customised interactive solutions for tax consultants and large multinational groups.
- **Advanced data visualisations:** Advanced data visualisations, such as Sankey, chord, relationship diagrams, etc., were unique and influenced SAVYVA's decision

to adopt Dundas BI. Sankey diagram is visualisation for transfer pricing, which displays the flow of inter-company transactions between multiple legal entities.

- **Customised navigation components:** SAVYVA designed customised navigation for CROSSVIEW which enabled it to better manage transfer pricing business processes. The end users could navigate among different dashboards in different view modes.
- **Narrative reporting:** SAVYVA integrated a narrative reporting component with Dundas BI which enabled end users to embed, format and update files, dashboards and scoreboards within the CROSSVIEW's online document report file editor. Different files of different types can be linked to each other.
- **Process and transaction flowcharting:** CROSSVIEW allowed end users to import and visualise complex process diagrams for RACI analysis.

**Source:** *https://www.dundas.com/learning/savyva*

### QUESTIONS

1. What type of challenges were faced by SAVYA?

   (**Hint:** SAVYVA realised that there are various complexities in dealing with transfer pricing, such as varied processes, managing large quantities of data and making all such data more actionable.)

2. What kind of partener SAVYA was looking for?

   (**Hint:** SAVYVA started looking for a partner who had expertise in the multi-dimensional analysis.)

3. Why did SAVYVA feel the need to develop a stronger BI tool?

   (**Hint:** SAVYVA wanted to build or acquire a BI tool that possessed quantitative analysis capabilities and a solution in which the organisation can embed its own components as well.)

4. How was CROSSVIEW platform developed?

   (**Hint:** The combination of Dundas BI platform along with the system components was developed into a product called CROSSVIEW.)

5. Why did SAVYVA choose Dundas BI?

   (**Hint:** SAVYVA chose Dundas BI because of reasons, such as Platform Extensibility and Enterprise Level Governance.)

## 9.13 EXERCISE

1. What do you understand by data visualisation? List the different ways of data visualisation.
2. Describe the different techniques used for visual data representation.
3. Discuss the types and applications of data visualisation.
4. Describe the importance of Big Data visualisation.
5. Elucidate the transformation process of data into information.
6. Write a short note on the tools used in data visualisation.

## 9.14 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Ways of Representing Visual Data | 1. | a. Timeline |
| | 2. | Infographics |
| Techniques used for Visual Data Representation | 3. | Isoline |
| | 4. | d. Map |
| | 5. | True |
| Types of Data Visualisation | 6. | b. 1D/Linear |
| | 7. | Network |
| Applications of Data Visualisation | 8. | Systems visualisation |
| | 9. | True |
| | 10. | c. Visual analytics |
| Visualising Big Data | 11. | Analytical |
| | 12. | True |
| | 13. | True |
| | 14. | Visual data mining |
| Tools used in Data Visualisation | 15. | Excel |
| | 16. | True |
| | 17. | a. TwittEarth |
| Data Visualisation for Managers | 18. | True |
| | 19. | True |
| Visualising and Exploring Data in Excel | 20. | True |
| | 21. | Scatter chart |
| | 22. | a. Surface chart |

## 9.15 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Yau, N. (2013). *Data Points. Indianapolis*, IN: J. Wiley & Sons.
- Yuk, M. and Diamond, S. (2014). *Data Visualization for Dummies*. Hoboken, New Jersey: Wiley.


### E-REFERENCES


- Data visualization. (2018, November 22). Retrieved November 30, 2018, from https://en.wikipedia.org/wiki/Data_visualization
- Suda, B., & Hampton-Smith, S. (2017, February 07). The 38 best tools for data visualization. Retrieved November 30, 2018, from https://www.creativebloq.com/design-tools/data-visualization-712402
- 50 Great Examples of Data Visualization. (2009, June 01). Retrieved November 30, 2018, from https://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/

CHAPTER 10

# Quantitative Techniques

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Explain the concept of linear programming techniques
- Describe the different problems solved using quantitative techniques
- Describe the game theory and dynamic programming

## 10.1 INTRODUCTION

In the previous chapter, you studied about data representation and visualisation. The chapter explained various ways and techniques to represent visual data. You also learned about different types of data visualisation tools and applications.

In any organisational setup, decision making is considered a very important process, which is generally taken care of by the people in the higher management, who may or may not be the owners of the organisation. Decision makers often adopt different types of quantitative techniques to take timely and correct decisions in different kinds of situations. Some common quantitative techniques used by decision makers to resolve complex organisational problems are linear programming, capital budgeting, game theory, decision tree, etc. Quantitative techniques work upon the past data of an organisation and emphasise on creating a mathematical expression to state the objectives and constraints of a problem to find the most appropriate solutions for achieving organisational goals. Some of the basic steps involved in resolving the problem by using quantitative techniques are planning, leading, organising and controlling.

The chapter begins by explaining the concept of linear programming. Further, it explains different techniques to solve complex organisational problems, such as graphical method, simplex method, duality concept and sensitivity analysis. You will also learn about different concepts such as transportation problem, shortest path problem, assignment problem, minimum spanning tree, network models with yield, game theory, dynamic programming and neural networks.

## 10.2 OVERVIEW OF LINEAR PROGRAMMING (LP)

Linear programming is the simplest way to understand complex relationships or problems through linear functions and find optimum solutions. These methods can be applied for finding solutions of problems in both personal and professional lives. For example, linear programming can help you find the shortest route while moving from one place to another. At work, it helps you create strategies to improve the efficiency of your team and deliver projects on time.

To solve a problem in linear programming, you have to be familiar with some common terminologies:

- **Decision variables:** To solve a problem, you first have to identify certain quantities that affect the output of the problem. These quantities are called decision variables. For example, the total number of items produced by manufacturing units, P and

Q, of an organisation are X and Y, respectively. These X and Y will be the decision variables for the organisation.

- **Objective function:** It is important to define an objective to solve a problem or to take the right decision. The objective function is mainly used to maximise or minimise a numerical value. The numerical value could be the cost of the raw material, cost of the project, profit margin, etc. For example, a company ABC wants to increase the sale of its products so 'Increment in sale' is the objective and the mathematical representation of this objective is the objective function.
- **Constraints:** These refer to the restrictions or limitations that exist for decision variables. In other words, the constraints restrict the value of decision variables. For example, in an organisation, the limit on the availability of certain resources is known as constraints.
- **Non-negativity restriction**: The values for decision variables should be greater than or equal to 0.

Let us now learn about certain methods and concepts required to solve linear programming problems.

### 10.2.1 LINEAR PROGRAMMING FORMULATIONS

A problem is called a linear programming problem if the objective function, decision variables and constraints are all linear functions. The steps commonly used for defining a linear programming problem involve:

- Identifying decision variables
- Writing an objective function
- Mentioning constraints
- Stating the non-negativity restriction explicitly

Consider a linear programming problem in which a watch manufacturer wants to maximise his profit. For this, he first needs to determine the sales and the profit earned individually from both ladies and gents watches.

Let $x_1$= Optimal production of ladies watches

$P_1$= Profit from each ladies watch sold

$x_2$= Optimal production of gents watches

$p_2$= Profit from each gents watch sold.

Hence, total profit from ladies watches = $p_1 x_1$

Total profit from gents watches = $p_2 x_2$

The objective function of the problem for maximising profit can be formulated as:

maximise $Z = p_1x_1 + p_2x_2$

Let w be the total amount of material available to produce both ladies and gents watches. Each unit of ladies watch consumes $w_1$ unit of material and each unit of gents watch consumes $w_2$ units of material.

In this problem, constraints like availability of raw material can be expressed by the following mathematical expression:

$w_1 x_1 + w_2 x_2 \leq w$.

In addition to raw material, other resources such as labour, machinery and time are also considered in the preceding expressions.

## 10.2.2 GRAPHICAL SOLUTION

If you want to find the values of two decision variables, a graphical method is the best option to find an optimal solution. In this method, you can explicitly visualise both the procedure and the solution.

Let us understand this with the help of an example in which we assume the following objective function, constraints and non-negativity restriction:

(1) **Objective method:** Maximum Z = 50x + 18y

(2) **Constraints:** 2x + y <= 100 -------(1)

x + y <= 80 ----------(2)

(3) **Non-negativity restriction**: x1 and x2 > = 0

After making the preceding assumptions,

Consider 2x + y = 100 ----------(1)

Put x =0 in eq. (1) to find the value of y.

2(0) + y =100

Put y = 100 in eq. (1)

2x + 0 = 100

2x = 100

x = 100/2

x = 50

The values of x and y are as follows:

| | | |
|---|---|---|
| *x* | 0 | 50 |
| y | 100 | 0 |

Now, consider equation 2 to find another set of values of x and y.

x + y = 80 -------(2)

when x = 0, y = ?

Notes

0 + y = 80

y = 80

When y = 0, x = ?

x + 0 = 80

x = 80

The values of x and y are as follows:

| | | |
|---|---|---|
| x | 0 | 80 |
| y | 80 | 0 |

The feasible region formed by equations 1 and 2 is shown in Figure 1:

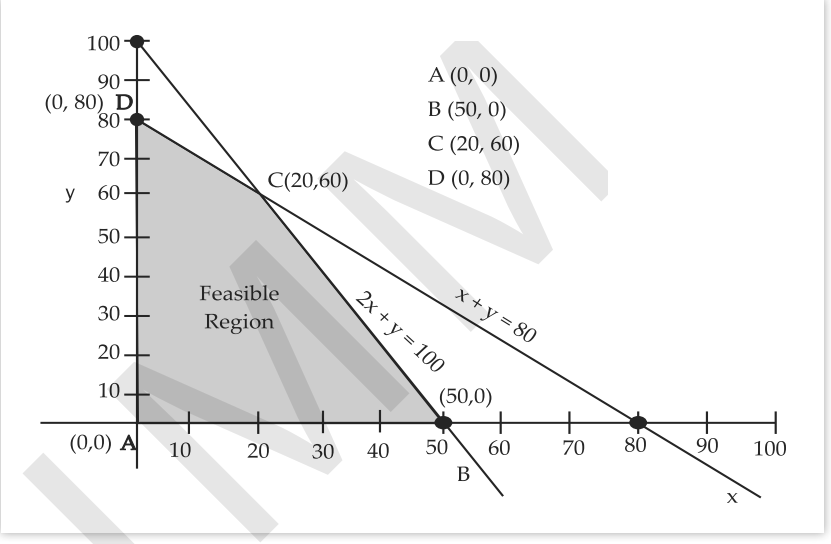



Figure 1: Displaying the Feasible Region

In Figure 1, the shaded area on the graph or the area which is under both the lines is called the feasible region or solution space. The four corner points of the feasible region, A(0,0), B(50,0), C(20,60) and D(0,80), are used to find the maximum value for an objective function.

Z = 50x + 18y

A(0,0) Z(A) = 50(0) + 18 (0) = 0

B(50,0) Z(B) = 50(50) + 18 (0) = 50(50) + 0 = 2500

C(20,60) Z(C) = 50(20) + 18 (60) = 100 + 1080 = 2080

D(0,80) Z(D) = 50(0) + 18 (80)= 0 + 1440 = 1440

Now, choose the maximum value as it will be the optimal solution for the objective function.

The maximum value in this case is B(50,0) Z(B) = 2500.

### 10.2.3 SIMPLEX METHOD

Any type of linear programming problem involving only two variables can be easily solved using the graphical method. If the variables are more than two, then solving a linear programming problem using the graphical method becomes difficult. The simplex method was developed to overcome the limitations of the graphical method. It is an iterative procedure for getting the most feasible solution. This method involves a function and several constraints stated as inequalities. These inequalities create a polygon region and the solution lies on one of the vertices or corner points of the polygon region.

Due to the increase in the number of equations and variables, these corner points can be very large in number. An effective algorithm can be used to reduce the corner points to be tested and reach an optimal solution in just a few iterations.

### 10.2.4 ARTIFICIAL VARIABLES

Artificial variables are created with the only purpose of using the simplex method on problems involving mixed constraints. These variables have no physical existence in the problem and are only used for the objective of finding the basic feasible solution so that the simplex method can be applied.

In order to reach the only basic feasible solution, the artificial variable introduced must satisfy the non-negative constraint. Consider the following equation in which an artificial variable 'a' is introduced containing the surplus variable, v:

$y_1 + y_2 - v + a = 2$

Now, to restrict the artificial variable to become a part of the final optimal solution of the given problem, a positive constant is introduced in the solution which forces the artificial variable to become zero.

### 10.2.5 SPECIAL CASES: ALTERNATIVE OPTIMA, INFEASIBILITY, UNBOUNDED

In linear programming, you have so far learned about the graphical solution, the simplex method and the concept of artificial variables. In this section, you will learn about special cases in linear programming like alternative optima, infeasibility and unbounded to understand and solve different types of linear programming problems:

- **Alternative optima:** A solution is called the alternative optima solution when a linear programming problem has more than one optimal solution which satisfies the set of constraints of the problem. The objective function for the problem can be either maximised or minimised.
- **Infeasibility:** In a problem, if no solution exists which satisfies all the constraints, then this situation is called infeasibility and the problem is called an infeasible problem. A linear programming problem (LPP) has no feasible solution if one artificial variable is positive in the optimum iteration. This situation never occurs when all the constraints on the right-hand side are non-negative.

NOTES

- **Unbounded:** An unbounded solution of a linear programming problem refers to a situation in which an objective function is enhanced indefinitely without ignoring its constraints and bounds. In case of the simplex method, the unbounded solution indicates the existence of infinite or negative values of the replacement ratio. In case of the graphical method, if the feasible region has no boundaries or constraints defining its maximum limit, then the solution obtained is considered as unbounded. Generally, real-life situations have bounded solutions. If any such situation of unbounded solution arises, there might be chances that some errors have been committed in the formulation of the problem.

## 10.2.6 USING EXCEL SOLVER TO SOLVE LP PROBLEMS

Solver is a tool in Excel that has the capability to solve linear programming problems and allows integers or binary restrictions to be placed on decision variables. You can use this tool to solve problems with up to 200 decision variables. Before using the Solver to resolve a problem, you need to install it in the Excel application. In our case, we have used Excel 2010 to solve linear programming problems. Perform the following steps to add the Solver add-in in Excel:

1. *Open* Excel 2010 and click the File → Options option.

   The Excel Options dialog box appears (Figure 2).

2. *Select* the Add-Ins option in the left pane (Figure 2).
3. *Select* the Excel Add-Ins option in the Manage dropdown list (Figure 2).
4. *Click* the Go button, as shown in Figure 2:

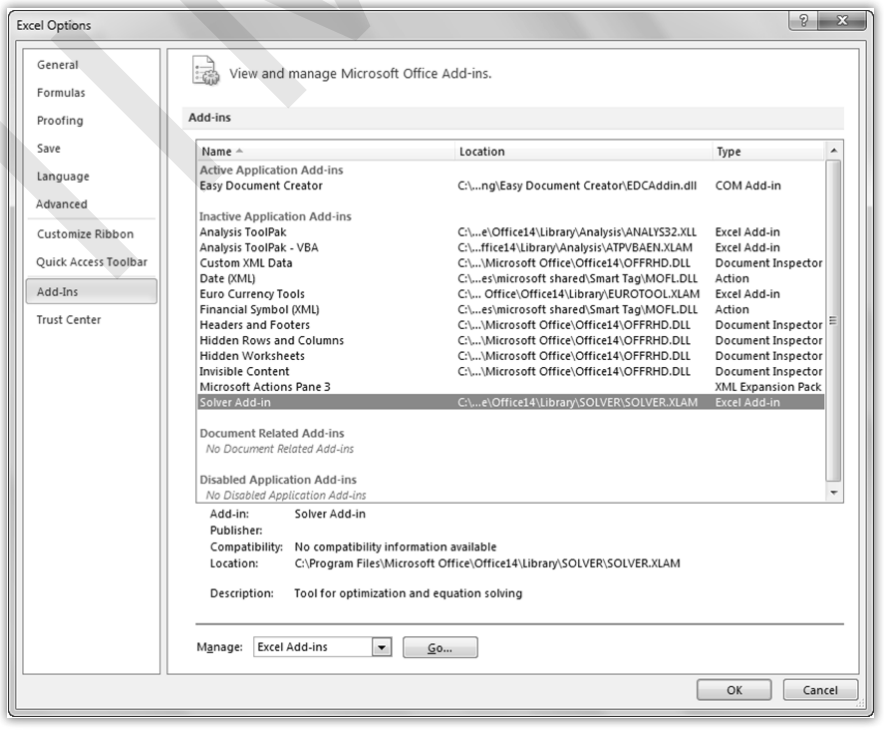



Figure 2: Excel Options Dialog Box

The Add-Ins dialog box appears (Figure 3).

5. *Select* the Solver Add-in option and click the OK button, as shown in Figure 3:

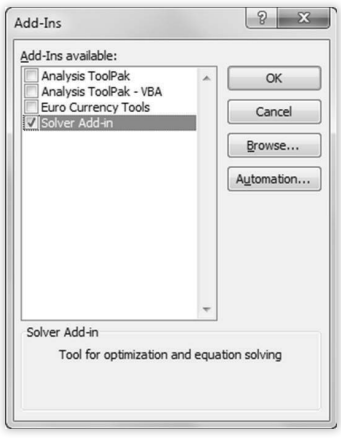



FIGURE 3: Adding Solver Add-in

After successful loading of the Solver Add-in, the Solver command gets displayed in the Analysis group on the Data tab.

Now, consider an example of a food chain restaurant, which wants to open 9 new branches in a large city. They have three variations of the location-convenience food corner, standard food corner and expanded food corner. The convenience food corner requires 5.125 crores to build and 35 employees for its operation. The standard food corner requires 7.25 crores to build and 15 employees for its operation. The expanded food store requires 13.375 crores to build and 50 employees for its operation.

The corporation can spend 83.5 crores in construction and employ 250 people. The estimated annual revenues from the convenience food corner, standard food corner and expanded food corner are 1 crore, 4 crores and 2.8 crores, respectively. How many of each variation should be built by the food chain restaurant to maximise its revenue?

### Solution:

The solution of the preceding problem is as follows:

1. *Create* the decision variables:

   x = convenience food corner

   y = standard food corner

   z = expanded food corner

2. *Identify* the constraints of the problem as:

   $x + y + z \leq 9$

   $5.125x + 7.25y + 13.375z \leq 83.5$

   $35x + 15y + 50z \leq 250$

   $x \geq 0$, $y \geq 0$, and $z \geq 0$

3. *Write* the objective function as:

   $P(x, y, z) = 1x + 4y + 2.8z$

NOTES

4. *Enter* the preceding data in an Excel sheet (Figure 4).
5. *Enter* the following formulas in the Excel sheet (Figure 4):

| | Cell | Formula |
|---|---|---|
| constraint $a_1$ | E15 | = E6 + G6 + I6 |
| constraint $a_2$ | E16 | = 5.125*E6 + 7.25*G6 + 13.375*I6 |
| constraint $a_3$ | E17 | = 35* E6 + 15*G6 + 50*I6 |
| Maximise | E19 | = 1* E6 + 4*G6 + 2.8*I6 |

6. *Click* the Solver button in the Data tab, as shown in Figure 4:

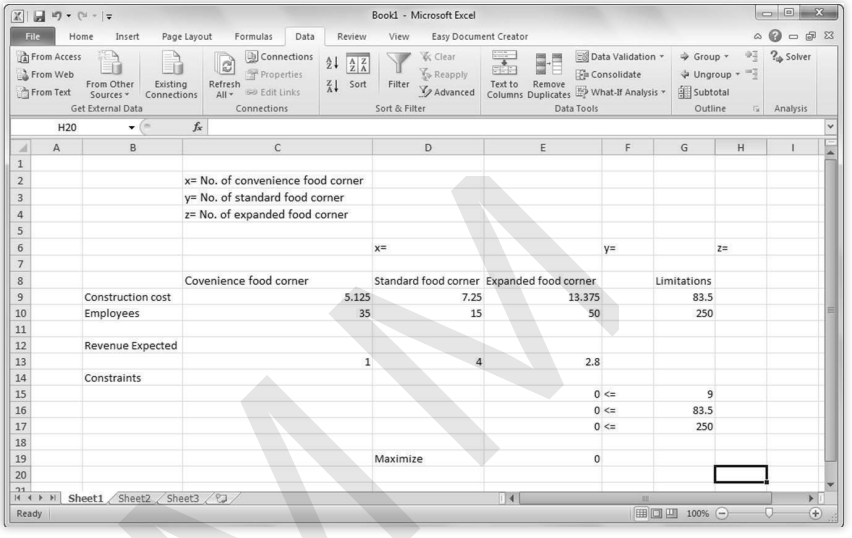



FIGURE 4: Data in Excel Sheet

The Solver Parameters dialog box appears (Figure 5).

7. *Enter* the cell addresses and constraints (Figure 5).
8. *Select* the Simplex LP in Select a Solving Method dropdown list (Figure 5).
9. *Click* the Solve button, as shown in Figure 5:

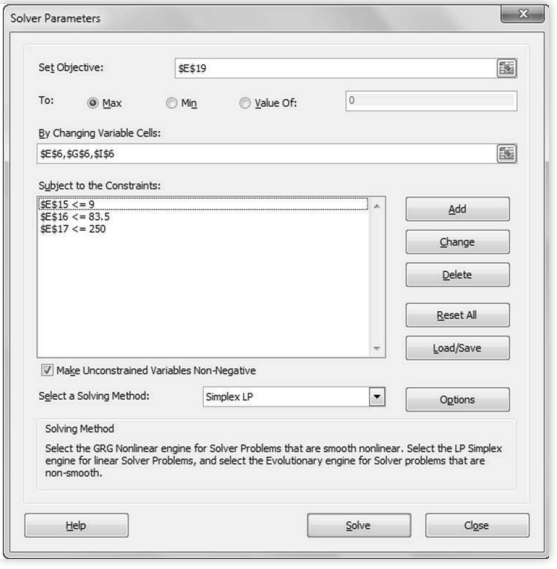



FIGURE 5: Solver Parameters Dialog box

The Solver Results dialog box appears (Figure 6).

10. *Click* the OK button to save the results in the sheet, as shown in Figure 6:

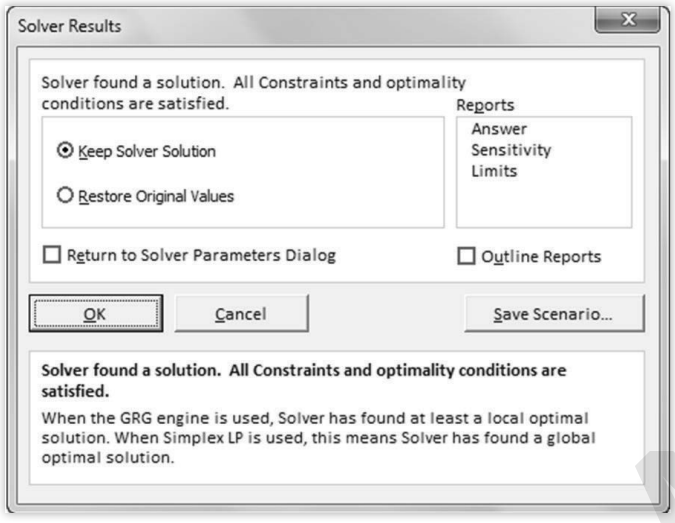



**Figure 6: Solver Results Dialog box**

The results appear in the Excel sheet, as shown in Figure 7:

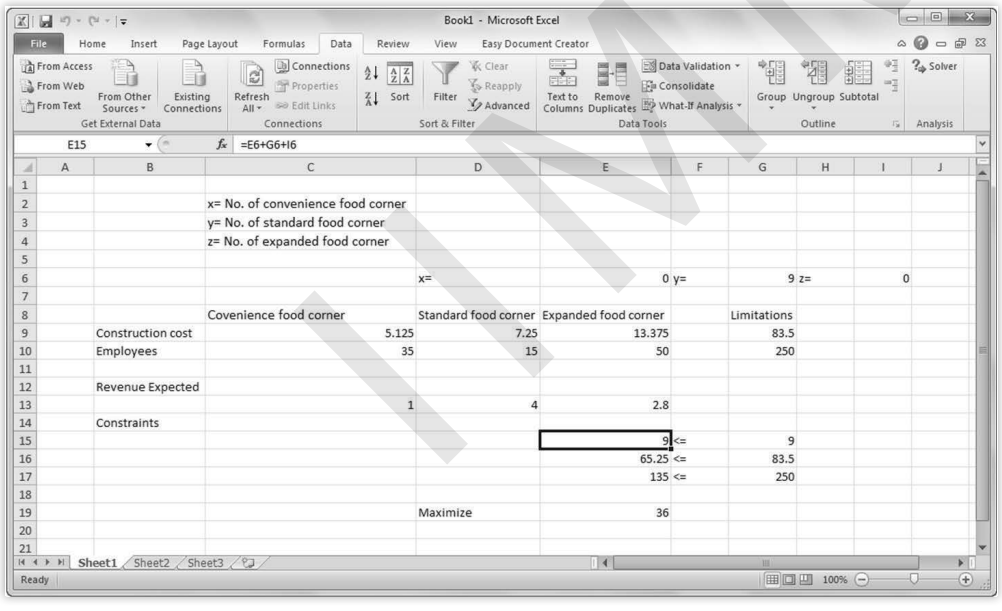



**Figure 7: Solver Results**

In Figure 7, the value of x and z is 0; whereas, the value of y = 9. It suggests that the food chain restaurant must open 9 standard food corners to maximise its revenue.

## 10.2.7 DUALITY CONCEPTS

According to the duality principle in accounting, all aspects of a transaction must be recognised. This principle states that all transactions must have two aspects: one debit and one credit. In accounting, there must be a giver and a receiver of value. Double entry book keeping system is the best implementation example of the duality concept. It is the fact that every transaction has a dual effect in a business and this is

Notes

recorded in accounts. For example, when a sale is made, the asset of stock is reduced as goods leave the business and the asset of cash is increased as cash comes into the business. Every financial transaction behaves in this dual way.

In linear programming, the duality concept states that every linear programming problem has a related linear programming problem which describes the original linear programming problem. The original linear programming problem is known as primal and the related or derived problem is known as dual.

In duality, the maximisation problem in the primal transforms into the minimisation problem and vice versa. If the primal problem comprises x variables and y constraints, then the dual will comprise y variables and x constraints.

In matrix form, you can express the primal problem as:

Maximise $c^T x$ subject to $Ax \leq b$, $x \geq 0$;

With the corresponding symmetric dual problem,

Minimise $b^T y$ subject to $A^T y \geq c$, $y \geq 0$.

On the other hand, an alternative primal formulation is:

Maximise $c^T x$ subject to $Ax \leq b$;

With the corresponding asymmetric dual problem,

Minimise $b^T y$ subject to $A^T y = c$, $y \geq 0$.

Some advantages of using the duality concept are as follows:

- It is helpful in case of the presence of a large number of constraints and a small number of variables in primal. The computation of the primal can be done by first converting the problem into dual and then solving it. This considerably reduces the computation.
- The duality concept helps managers in finding alternative actions to a problem.
- Calculation of the dual verifies the accuracy of the primal solution.

### 10.2.8 SENSITIVITY ANALYSIS

Sensitivity analysis refers to the study and causes of the uncertainty in the output obtained of a mathematical model or system that relies on inputs from different sources. An error or uncertainty in any of these inputs may result in an uncertain output. Uncertainties in the input can be due to the errors made during the measurements or poor understanding of the procedures.

Consider an example of different mathematical models such as the economic model, climate model, finite element model, etc., which are usually very complex. In these models, the exact relationship between the inputs and outputs are difficult to understand. This type of model in which the output is regarded as an opaque function depending upon the inputs provided is also known as a black box.

The following can help minimise the uncertainties in the models:

- Quantification of uncertainty in the model
- Evaluation of the amount of input that is contributing to uncertainty

Sensitivity analysis is one of the tools that give a clear vision and right utilisation of decision models to derive the solution of a problem. It helps decision analysts to understand the limitations, advantages, disadvantages and scope in a decision model. Finally, decision makers get a clear idea that how sensitive an optimum solution is and how to choose input values for one or more parameters to make changes. In sensitivity analysis, you have to follow one principle that is to change your decision model and observe the action.

Sensitivity analysis is also used in financial modeling where a financial analyst wants to find out the effect of a company's net working wealth on its profit margin so analysis will involve all the variables that have an impact on the company's profit margin (i.e., cost of goods sold, workers' wages and managers' wages, etc.).

There are many important reasons to perform sensitivity analysis:

- Sensitivity analysis adds credibility to any type of financial model by testing the model across a wide set of possibilities.
- Financial sensitivity analysis allows the analyst to be flexible with the boundaries within which to test the sensitivity of the dependent variables against the independent variables. For example, the model to study the effect of a 5-point change in interest rates on bond prices would be different from the financial model that would be used to study the effect of a 20-point change in interest rates on bond prices.

Let us now learn to perform sensitivity analysis by using Data Tables in Excel. Consider an example of a furniture shop which sells fancy wooden almirahs. Now, the owner of the shop wants to analyse how the change in price and sales volume affects his profit. This will enable him to adjust his sales strategy to earn better profit.

Figure 8 shows the data related to the sales of the furniture shop:

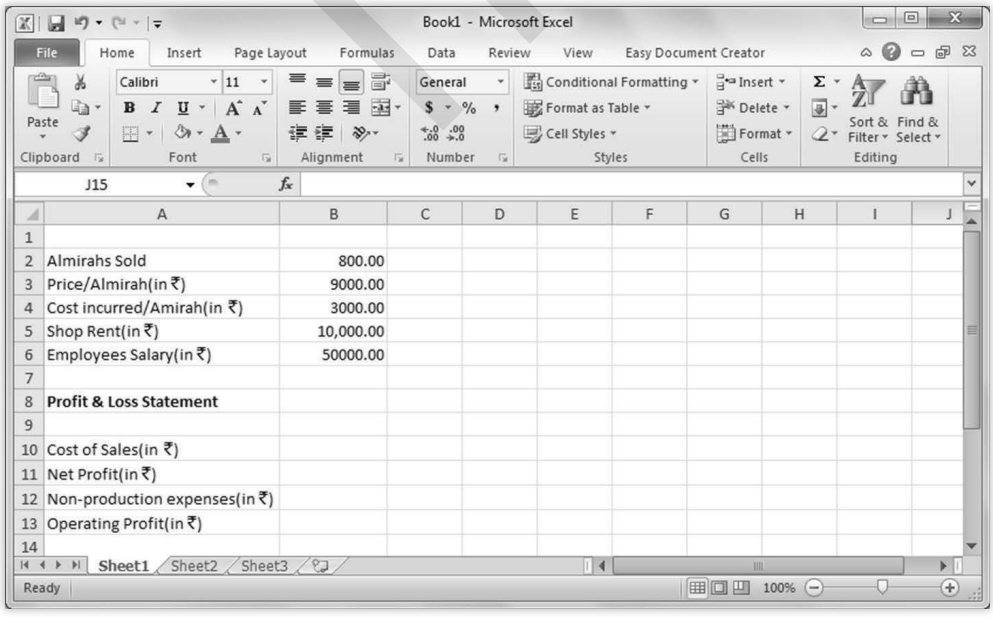



FIGURE 8: Displaying Sales Data of the Furniture Shop

NOTES

Now, type the following formulas in different cells of the Excel Sheet:

1. Type = B3*B2 in cell B10
2. Type = B4*B2 in cell B11
3. Type = B10 – B11 in cell B12
4. Type = B12 – B5 – B6 in cell B13

The Excel Sheet displays the values of different variables after typing the formulas, as shown in Figure 9:

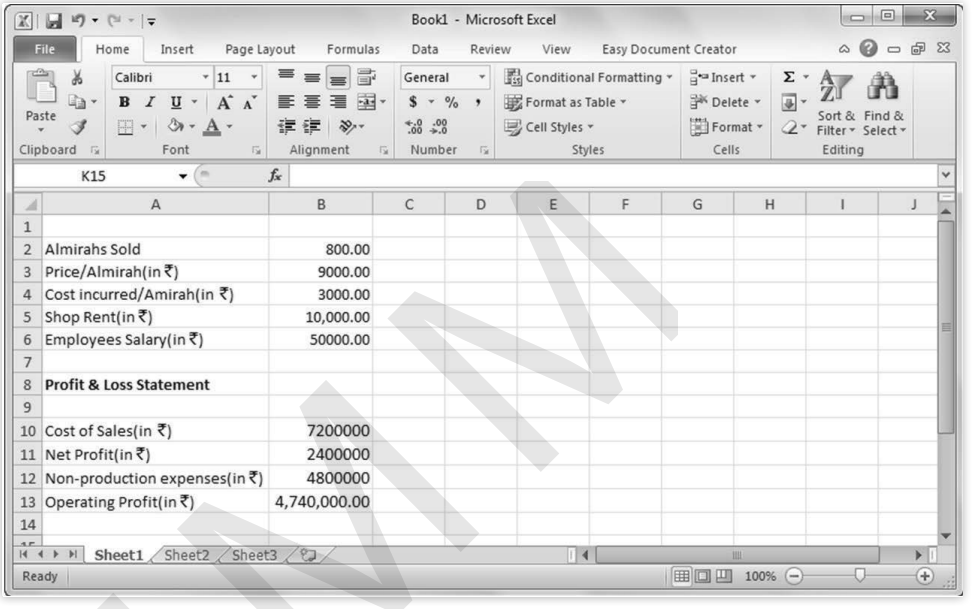



FIGURE 9: Entering Formulas in the Excel Sheet

Enter and select the data related to the number of chairs sold at different prices. Now, select the Data Table option from the What-If Analysis drop-down list in the Data tab, as shown in Figure 10:

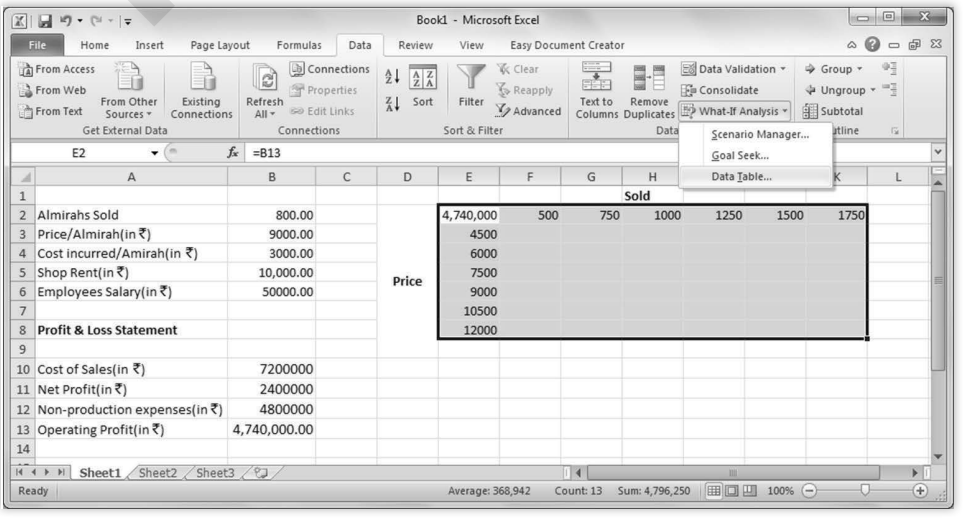



FIGURE 10: Selecting the Data Table Option

The Data Table dialog box appears. Now, enter the reference of B2 and B3 cells in the Row and Column input cell text boxes. Next, click the OK button in the Data Table dialog box, as shown in Figure 11:

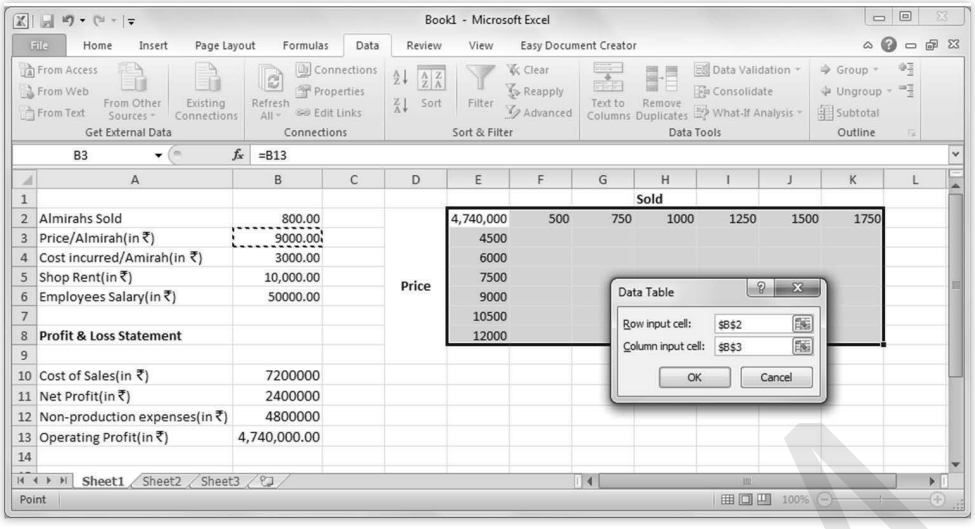



FIGURE 11: ADDING REFERENCES IN DATA TABLE DIALOG BOX

Figure 12 shows that the amount of profit changes with the change in the values of sales and price volume:

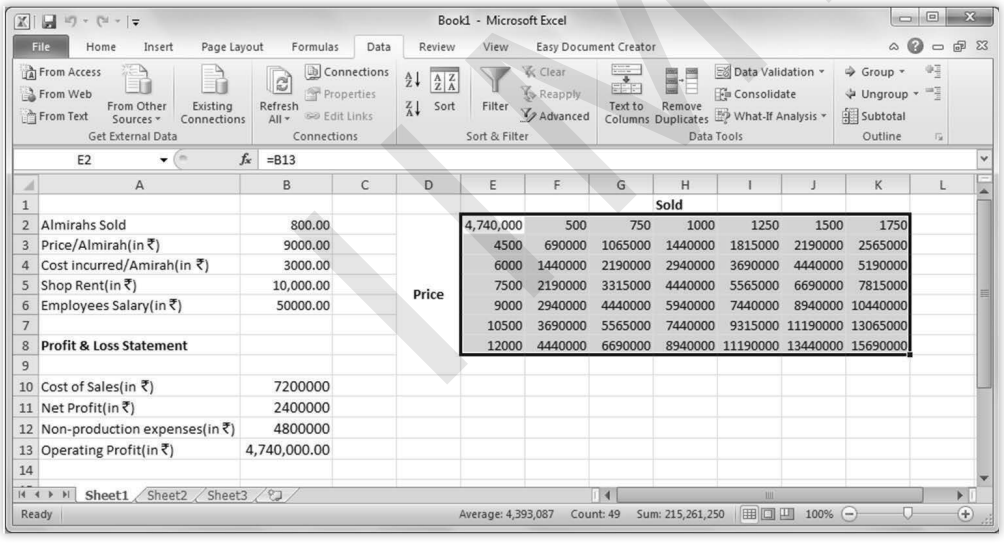



FIGURE 12: DISPLAYING THE CHANGE IN PROFIT

In Figure 12, you can notice that the amount of profit varies when the sales and price volume are changed. For example, when 750 almirahs are sold at a price of ₹ 6000, the profit is about 2190000 as compared to the profit of 4440000 made on the sale of 1500 almirahs.

## 10.2.9 CHI-SQUARED TEST

The chi-squared test is used to determine whether a substantial difference exists between the expected frequencies and the observed frequencies in one or multiple

Notes

categories. The chi-squared test is denoted by the symbol $\chi^2$. This test performs successfully with categorical data and not with numerical data. The categorical data is the one which can be counted and divided into categories. For example, the chi-squared test cannot be used to know how attendance of a student in a class influences his performance in exams by using his score or percentage in exams.

However, chi-squared test can be used if students are categorised into 'Pass' and 'Fail'.

In other words, this test is used for measuring how well the observed distribution of data fits with the expected distribution of data in case of independent variables. The formula for calculating chi-square is as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

In the preceding formula, the O signifies the observed or actual value whereas E signifies the expected value.

Consider an example in which a researcher wants to investigate whether taking a particular tonic every day for a certain period protects people from catching cold during that particular term. The data is shown in the following table:

| | Number of People Who Took the Tonic | Number of People Who Did Not Take the Tonic |
|---|---|---|
| Caught cold during the term | 14 | 21 |
| Did not catch cold during the term | 100 | 93 |

**Solution:**

The null hypothesis is as follows:

$H^o$ = Taking the tonic does not affect the probability of catching cold. Assume the level of significance (P) as 0.05.

Considering the null hypothesis to be true, we expect the number of people not taking the tonic and catching cold is equal to the number of people catching cold after taking the tonic.

| | Number of People Who Took the Tonic | Number of People Who Did Not Take the Tonic | Total Number of People |
|---|---|---|---|
| Caught cold during the term | 14 | 21 | 35 |
| Did not catch cold during the term | 100 | 93 | 193 |
| Total | 114 | 114 | 228 |

The proportion of people who caught cold is 35/228 = 0.15

It is expected that 15% of people taking the tonic may catch cold, i.e., 17.1 persons (15% of 114).

It is expected that 85% of people taking the tonic may not catch cold, i.e., 96.9 persons (85% of 114).

It is expected that 15% of people not taking the tonic may catch cold, i.e., 17.1 persons (15% of 114).

It is expected that 85% of people not taking the tonic may not catch cold, i.e., 96.9 persons (85% of 114).

Now, consider Table 1:

**Table 1: Calculating the Value of Chi-square**

| | Number of People (O) | Expected (E) | O-E | (O-E)-0.5 (Yates correction) | Square of Corrected Difference | Square of Corrected Difference/E |
|---|---|---|---|---|---|---|
| Took the tonic, caught cold | 14 | 17.1 | -3.1 | -3.6 | 12.96 | 0.76 |
| Took the tonic, did not catch cold | 100 | 96.9 | 3.1 | 2.6 | 6.76 | 0.07 |
| Did not take the tonic, caught cold | 21 | 17.1 | 3.9 | 3.4 | 11.56 | 0.67 |
| Did not take the tonic, did not catch cold | 93 | 96.9 | -3.9 | -4.4 | 19.36 | 0.2 |
| Total | | | | | | 1.70 |

Now, calculate the degrees of freedom (DF).

DF = (number of rows – 1) x (number of columns – 1)

DF = (2–1) x (2–1)

DF = 1

Consider Table 2 for determining the critical value:

**Table 2: Displaying the Critical Value**

| Degrees of Freedom | Significance Level | | |
|---|---|---|---|
| | 5% | 2% | 1% |
| 1 | 3.84 | 5.41 | 6.64 |
| 2 | 5.99 | 7.82 | 9.21 |

The critical value of the chi-squared test at 5% significance and 1 degree of freedom is 3.84.

The calculated value is 1.70.

The calculated value is smaller than the critical value at 5% level of probability.

Notes

### Conclusions

We cannot reject the null hypothesis because there is no significant difference between the observed and expected results at the 5% level of probability.

There is no significant difference between the number of people taking the tonic and of those not taking the tonic.

**Self Assessment Questions**

1. _____ is the simplest way to understand complex relationship or problems through linear functions and find the optimum solution.
2. To solve a problem, first you have to identify the decision variables because your output is dependent on these variables. (True/False)
3. The values for decision variables should be greater than or equal to _____.
4. A problem is called _____ problem if the objective function, decision variables and constraints are all linear functions.
5. Simplex method was developed by _____ in 1947.

**Activity**

Search and write a short note on the use of Excel Solver in Data Envelopment Analysis (DEA) which is a method used to measure the productive efficiency of the decision-making module.

## 10.3 PROBLEMS SOLVED USING QUANTITATIVE TECHNIQUES

As a company grows, different types of problems and challenges arise in handling its diversified operations, some of which are:

- Meeting the deadline of production
- Supplying of goods on time
- Calculating transportation cost
- Handling wages

These are the basic operations performed in an organisation on a routine basis. These operations sometimes become very complex, prompting managers to perform a detailed analysis to take appropriate decisions. Let us learn how different types of problems are handled in an organisation.

### 10.3.1 TRANSPORTATION PROBLEM

In transportation problem, the objective is to minimise the cost of distribution of products or services from various sources to destinations. The source or origin refers to the location from where the products or services are delivered to a particular location or destination. Transportation involves cost. The unit transportation cost refers to the cost of dispatching a unit of product from the source to destination.

Figure 13 shows how goods can be sent from a number of sources to a number of destinations:

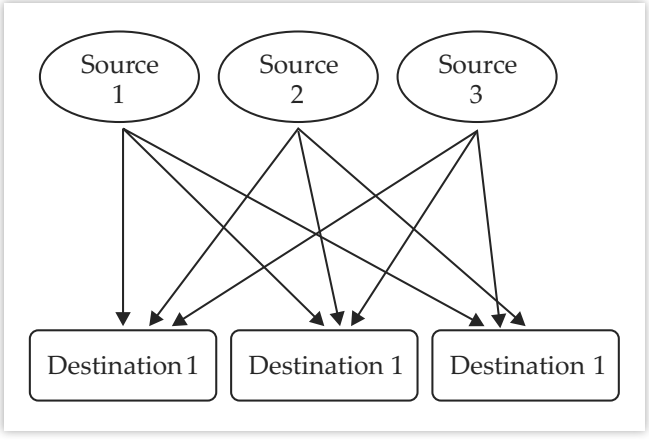



FIGURE 13: Displaying the source to destination connectivity

**Source:** *https://study.com/academy/lesson/using-the-transportation-simplex-method-to-solve-transportation-problems.html*

There are two types of transportation problems:

- **Balanced:** The balanced transportation problem is the problem in which the number of products at source equals the number of products required at the destination. For example, the total production is 30000 units at three factories and requirement at three warehouses is 10000 units each, then the transportation problem is considered as a balanced one.
- **Unbalanced:** The unbalanced transportation problem is the one in which the total availability of products at source is not equal to the total requirement of products at the destination. For example, the total production is 30000 units at three factories and requirement at three warehouses is 15000 units or 8000 units each, then the transportation problem is considered as an unbalanced one. An unbalanced transportation problem can be changed into a balanced transportation problem by adding dummy source(s) or dummy destinations as per the requirement having zero transportation cost per unit.

## 10.3.2 ASSIGNMENT PROBLEM

Management often has to deal with issues like allocating resources that may be limited in number to different activities simultaneously. This must be done in a way not only to ensure the completion of the task on time but also in a manner that restrains costs and maximises profits. This is known as the assignment problem. This problem can be resolved by the simplex method or by the transportation method but a more simple way to resolve the assignment problem is to use the assignment model.

Consider an example of a factory in which a manager needs to assign $n$ jobs to $m$ workers. He will have to take a decision regarding which job should be given to which worker. Each worker can take one job at a time. However, there must be some way so that the profit can be maximised and cost or time can be minimised.

Notes

The following actions or assumptions are taken into account to resolve such assignment problems:

- The number of workers and number of tasks are equal or not.
- One task must be assigned to each worker.
- Cost Cwt associated with worker w(m = 1, 2, 3,…., k) performing task t (n = 1, 2, 3, …., k).
- Analyse how all n tasks should be done to minimise the total cost.

### 10.3.3 TRANSSHIPMENT PROBLEM

In the transportation problem, shipments deliver consignments directly from the source to destination points but in many situations, it is not possible to deliver consignment directly to the destination. There is an intermediate point or location through which a shipment is transferred to a destination from a source. Figure 14 displays a shipment from a source to a destination and an intermediate point (transshipment point) from where the shipment is transshipped:

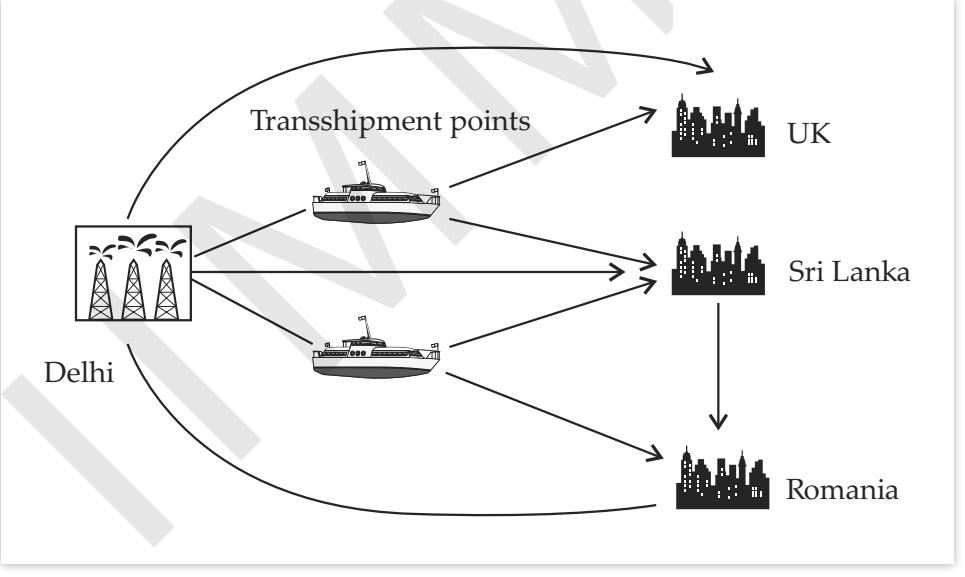



Figure 14: Displaying Transshipment Points

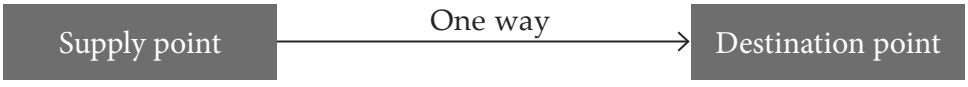
**Source:** *https://www.linearprogramming.info/model-of-transportation-with-transshipment-solved-with-excel-solver/*

In case of the transshipment problem, the cheapest route from source to destination is investigated in advance. The transshipment problem is introduced by a famous Physicist, Alex Orden, in 1956. The concept of the original transportation problem was based on direct shipping from source to destination but Orden extended this and included the possibility of transshipment.

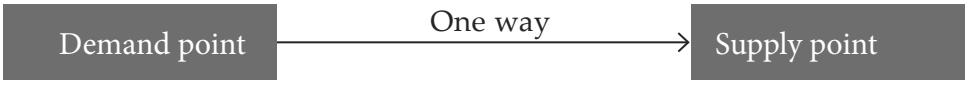
Resolving the transshipment problem does not appear to be a daunting task but becomes complicated and time-consuming when there are a number of intermediate transfer points. It is convenient to complete this task with the help of computer algorithms to minimise the total shipping cost.

The concept of the transshipment problem is as follows:

- **Supply point:** It is the sending point of a shipment. It is a one-way route.




- **Demand point:** It is the receiving point of a shipment. It is a one-way route.




- **Transshipment point:** It can both receive and send shipments. It is a two-way route.


### 10.3.4 SHORTEST PATH PROBLEM

In the shortest path problem, you search for the shortest path and the minimum cost of a path from a source to a destination node pair in a weighted graph network. The shortest path problem is used in the following areas or fields:

- Vehicle routing in the transportation system
- Path planning in the robotics system
- Scheduling
- Traffic routing in communication networks
- Video image analysis
- Virtual endoscopy
- Energy minimisation in vision

Some more variations of the shortest path problem are the single-source shortest path problem, the single-destination shortest path problem and the all-pairs shortest path problem. To solve the shortest path problem, you need an algorithm which best suits in a fast and real-time processing environment. Some key features required in an algorithm to solve the shortest path problem are:

- Fast computing
- Low memory consumption
- Robust
- High accuracy
- Dynamic/runtime reconfigurability

The shortest path algorithm is used in real-time applications like Google Maps for automatic search of directions in different situations between two physical locations. Some methods to address the shortest path problem are as follows:

- Scaling technique
- Fast integer sorting technique

Notes

- Integer matrix multiplication technique
- Component hierarchy technique

### 10.3.5 MAXIMUM FLOW PROBLEM

The maximum flow problem deals with the production of products at source(s) and delivery of the maximum amount of products towards destination(t) without incurring any holding cost for those products. The maximum flow problem finds its application in the following areas or fields:

- Internet traffic routing
- Railroad freight transportation
- Trucking
- Open-pit mining
- Sports team elimination
- Airline scheduling
- Numerical linear algebra

We can resolve maximum flow problem by using the Ford-Fulkerson algorithm which states that "as long as there exists a path from the start node(s) to the destination node (t), with available capacity on all edges in the path, we send flow along one of the paths. Then, we search for another path, and so on. A path with available capacity is known as an augmenting path".

### 10.3.6 MINIMUM SPANNING TREE

Consider an undirected and connected graph G = (V, E). A spanning tree of the graph G is a tree that includes every vertex of G (or spans G) and is a subgraph of G. In the subgraph of G, every edge in the tree belongs to G. The sum of the weights of all the edges in the tree refers to the cost of the spanning tree. There can be numerous spanning trees.

The minimum spanning tree is the spanning tree which involves the minimum cost among all the spanning trees. Like spanning trees, there can also be numerous minimum spanning trees. Minimum spanning trees find their applications in the following areas:

- Computer networks
- Telecommunications networks
- Transportation networks
- Water supply networks
- Electrical grids
- Cluster Analysis
- Handwriting recognition
- Image segmentation

The minimum spanning tree problem also finds its application in algorithms approximating the travelling salesman problem, multi-terminal minimum cut problem and minimum-cost weighted perfect matching. The two famous algorithms for finding a minimum spanning tree are Kruskal's and Prim's algorithms. In Kruskal's algorithm, the spanning tree is built by adding edges one by one into a growing spanning tree. In each iteration, this algorithm finds an edge which has minimum weight and adds that weight to the growing spanning tree. On the other hand, in Prim's algorithm, the minimum spanning tree is determined by growing

the spanning tree from an initial position. Next, we add a vertex to the growing spanning tree to find the minimum spanning tree.

## 10.3.7 NETWORK MODELS WITH YIELD

In the network model, the flow (material, people, or funds, etc.) is described from source to destination.

In the network model, the flow is subjected to positive or negative yield. To understand positive and negative yields, consider an example of a manufacturing company making metallic bowls. In the manufacturing process, waste is a negative yield. The amount of raw material is always greater than the final production because of waste. The following processes are involved in the manufacturing of metallic bowls by using the raw materials:

- Grinding
- Polishing
- Drilling
- Packing

One thing that needs to be observed and compared is the amount of the metal left at the end of all these processes to the amount of the metal in the beginning. The reduction in the amount of metal (or flow) is called the process yield.

In the production processes, the yield is generally less than 1 or negative as input is greater than the final output. However, in some other types of processes, yields may be greater than 1. For example, interest yield on bank balance is always greater than 1 or positive.

## 10.3.8 INTEGER PROGRAMMING (IP) FORMULATIONS

Integer programming states the optimisation of a linear function subjected to a set of linear constraints over integer variables. Integer programming is a powerful technique for the formulation of a wide variety of problems. The following are the areas in which IP is used:

- To blend with a limited number of ingredients
- To impose logical conditions in linear programming problems
- To schedule jobs
- To balance assembly line
- To schedule airline crew
- To dispatch vehicles

Some common problems in which integer programming formulation is commonly used are as follows:

- **Warehouse location problem:** It is a problem which is formulated as an integer programme. To understand this problem, let us take the example of an organisation which produces goods. To send goods from a factory to a customer's location, you need to take care of the transportation system. In this system, you have to involve locations of warehouses to minimise the overall cost of transportation.

Notes

- **Knapsack-capital budgeting problem:** In the typical capital-budgeting problem, decisions need to be taken in order to select the number of potential investments. The investment decisions may involve the selection of potential plant locations, the configuration of capital equipment, etc. The aim is to maximise the total contribution from all investments without crossing the limited availability of any resource.

**Self Assessment Questions**

6. In the transportation problem, the objective is to minimise the cost of distribution of products or services from various sources to destinations. (True/False)
7. The _____ refers to the cost of dispatching a unit of product from source to destination.
8. The unbalanced transportation problem is the problem in which the number of products at the source equals the number of products required at the destination. (True/False)
9. In the transshipment problem, shipments deliver consignments directly from the source to destination points. (True/False)
10. The shortest path problem is also known as the _____ problem.

**Activity**

Search and explain the two famous algorithms used for finding the minimum spanning tree with suitable diagrams.

## 10.4 ADDITIONAL PROBLEMS

In this chapter, you have learned about different statistical problems like transportation, assignment, transshipment, shortest path, maximum flow, minimum spanning tree, etc. In these statistical problems, data is analysed for interpretation. Now, you will learn about some more interesting concepts like game theory and dynamic programming in detail which are also used for decision making in organisations.

### 10.4.1 GAME THEORY

Game theory is the study of a group of people when they are in a competition. It tries to discover the actions that a player must perform which would maximise his/her probability of winning. Game theory can be applied in almost all fields such as social science, economics, science, computer science, etc. In other words, it is the science of making logical decisions by humans, animals, and computers.

The main idea behind game theory is the game in which if one player knows the other player's game strategy, he/she can easily win the game. The game emphasises on identifying the players' identities, likings, and available tactics and how these tactics can affect the result.

In game theory, one player analyses the following traits of other players to become successful:

- **Overconfidence:** Sometimes, people are overconfident and believe that only they can win a game. Overconfidence comes from the underestimation of enemy's skills; whereas, confidence comes from the belief in your own skills. Do not be overconfident because that is the biggest reason of losing a game. Come up with a plan and remember that everyone else has a plan too.
- **Logical attitude:** People need to be logical in certain situations. For example, chess is a logical game and if you want to win this game, you have to play logically because your every move is very important and the other player is also playing this game in the same way. Imagine, if in a game, every person is programmed like a machine, then one can easily judge the moves the other players may make.
- **Unpredictability:** People always feel comfortable with people who are predictable. Everything comes with some advantages and disadvantages. If you play a game logically, it becomes very easy for other players to predict your next move, and of course, this will help your opponent to win the game. Therefore, it is good to be unpredictable to play a game with different and unique ideas to make it difficult for your opponent to judge your game.
- **Psychology:** One of the most important factors to win a game is the knowledge of human psychology. If you have a good idea about human psychology, you can easily analyse other player's game in one or two moves and plan your moves accordingly to win the game.

In business, game theory is widely used to predict the behaviors of competitors. Often, businesses have numerous choices that affect their capability of gaining profitability. For example, an organisation may face difficulty in deciding to retire the existing products or to develop the new ones, reduce prices relative to the competitor, implement new marketing strategies, etc. Economists generally use game theory to understand the organisational behavior when engaged in different types of situations.

### 10.4.2 DYNAMIC PROGRAMMING

To solve a problem using dynamic programming, you need to break down the problem into subproblems recursively. After processing these subproblems, you need to save the outcomes for future references. The overall strategy employed in dynamic programming is to divide and conquer. Suppose we need to write a programme for finding a simple recursive solution for fibonacci numbers. Now if the exponential time complexity encountered while writing such a programme is optimised by using dynamic programming, then the time taken to write this programme gets much reduced.

Every dynamic programming problem has a schema to be followed:

- Show that the problem can be broken down into optimal subproblems.
- Define the value of the solution by expressing it in terms of optimal solutions for smaller subproblems recursively.

- Compute the value of the optimal solution in a bottom-up fashion.
- Construct an optimal solution from the computed information.

The two approaches used in dynamic programming are as follows:

- **Bottom-up approach:** In this approach, the result for the subproblem is computed, which is then used to solve another subproblem and finally the whole problem gets resolved.
- **Top-down approach:** In this approach, a large problem gets broken down into multiple subproblems. If the subproblem is already solved, then the outcome of the problem can be reused. Otherwise, you need to solve the subproblem and save the result. This approach uses the memorisation strategy to avoid the recomputation of the same subproblem again and again.

### 10.4.3 NEURAL NETWORKS

Neural networks, a fundamental component of artificial intelligence, have emerged as powerful tools in the realm of business analytics. These computational models are inspired by the human brain's intricate network of interconnected neurons, enabling them to recognise patterns, make predictions, and learn from data. Neural networks are employed to extract meaningful insights from vast datasets, aiding decision-making processes. By leveraging their ability to adapt and improve over time, neural networks can uncover hidden correlations and trends within the data, offering businesses a competitive edge. Applications range from customer segmentation and demand forecasting to fraud detection and sentiment analysis, allowing organisations to optimise operations, enhance customer experiences, and make data-driven strategic decisions. The versatility and robustness of neural networks make them an invaluable asset in the rapidly evolving landscape of business analytics, enabling enterprises to harness the full potential of their data for sustainable growth and innovation.

Integral components of neural networks include neurons, synapses (connections), weights, biases, propagation functions, and a learning rule. Neurons receive input from preceding neurons, undergo activation, and employ functions to produce output. Learning entails adjusting free parameters, specifically weights and biases. The learning process involves simulation, parameter adjustment, and the adaptation of neural network responses.

The evolutionary path of neural networks includes Hebbian learning, emphasising neural plasticity and long-term potentiation for unsupervised pattern recognition. Backpropagation addresses exclusive challenges, facilitating the efficiency of multi-layer networks by addressing errors layer by layer. This led to the development of support vector machines, linear classifiers, and max-pooling. The vanishing gradient problem in deep learning impacted feedforward and recurrent neural networks, prompting the introduction of hardware-based designs like biophysical simulation and neurotropic computing. Convolutional networks, featuring alternating convolutional and max-pooling layers, effectively resolved issues in unsupervised pre-training, where each filter represents a trained weight vector. Shift variance for small and large networks is tackled in development networks. Additional learning

techniques include error correction, memory-based learning, and competitive learning. Figure 15 shows the neural networks:

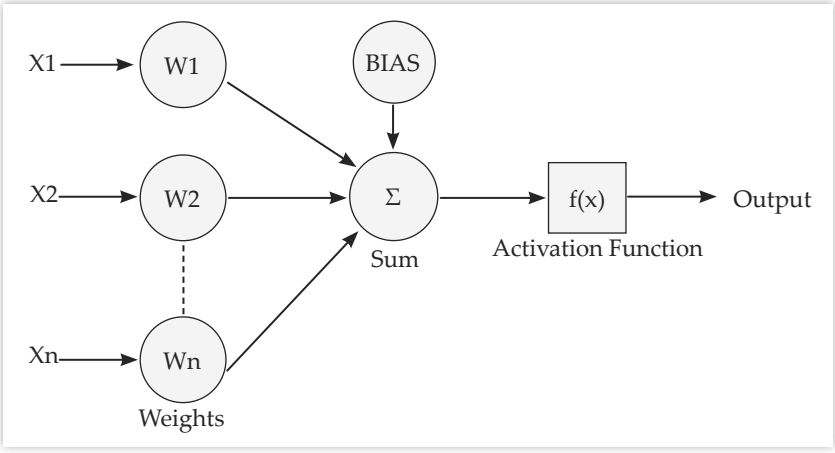



Figure 15: Neural Networks

### Types of Neural Networks

There are seven distinct types of neural networks:

- **Multilayer Perceptron (MLP):** An MLP is a type of artificial neural network characterised by multiple layers of interconnected nodes (neurons).
- **Convolutional Neural Network (CNN):** A CNN is a specialised type of neural network designed for processing and analysing structured grid data, such as images. CNNs use convolutional layers to automatically and adaptively learn spatial hierarchies of features from the input data.
- **Recursive Neural Network (RNN):** An RNN is a type of neural network designed for processing sequences of data. It maintains a hidden state that is updated at each step of the sequence, allowing it to capture temporal dependencies in the data.
- **Recurrent Neural Network (RNN):** An RNN is a more general term referring to neural networks that incorporate the concept of recurrence, allowing them to maintain and utilise information from previous steps in a sequence.
- **Long Short-Term Memory (LSTM):** LSTM is a type of RNN architecture designed to address the vanishing gradient problem and capture long-term dependencies in sequences. LSTMs use memory cells and gating mechanisms to selectively retain and update information over time, making them effective for tasks involving sequences with long-range dependencies.
- **Sequence-to-Sequence (Seq2Seq):** Seq2Seq refers to a type of model architecture that takes a sequence of data as input and produces a sequence of data as output. This architecture is commonly used in tasks such as machine translation, text summarisation, and speech recognition.
- **Shallow Neural Network:** A shallow neural network refers to a type of artificial neural network with a limited number of layers between its input and output layers.

Notes

**Self Assessment Questions**

11. Game theory was developed by John von Neumann. (True/False)
12. _____ comes from an underestimation of the enemy's skills; whereas, _____ comes from the belief in your own skills.
13. In _____, you break down the problem into sub-problems.
14. Which layer is responsible for receiving the initial input data in a neural network?

    a. Output layer
    b. Hidden layer
    c. Input layer
    d. Processing layer

## 10.5 SUMMARY

- Linear programming is the simplest way to understand complex relationships or problems through linear functions and find the optimum solution.
- A problem is called a linear programming problem if the objective function, decision variables and constraints are all linear functions.
- The graphical solution method is mostly used to solve a problem in which you have to find the highest or lowest point of intersection graphically.
- Artificial variables have no physical existence in a problem and are only used for finding the basic feasible solution so that the simplex method can be applied.
- Artificial variables are introduced in the equations that contain surplus variables.
- Duality theory tells us that if the primal is unbounded then the dual becomes infeasible by the weak duality theorem.
- The chi-squared test is used to determine whether a substantial difference exists between the expected frequencies and the observed frequencies in one or multiple categories.
- The transportation problem is related to the issues that occur in management of transporting the goods from one location to another.
- In case of the transshipment problem, the cheapest route from source to destination is investigated in advance.
- In the network model, the flow (material, people, or funds, etc.) is described from source to destination.
- Game theory is the study of a group of people when they are in a competition.
- Neural networks, a fundamental component of artificial intelligence, have emerged as powerful tools in the realm of business analytics.
- An MLP is a type of artificial neural network characterised by multiple layers of interconnected nodes (neurons).
- A CNN is a specialised type of neural network designed for processing and analysing structured grid data, such as images.

- An RNN is a type of neural network designed for processing sequences of data.
- LSTM is a type of RNN architecture designed to address the vanishing gradient problem and capture long-term dependencies in sequences.
- Seq2Seq refers to a type of model architecture that takes a sequence of data as input and produces a sequence of data as output.

## 10.6 KEY WORDS

- **Decision variables:** These are the variables that can be referred to as the quantities which decision makers want to determine.
- **Objective function:** It refers to the mathematical representation of a problem.
- **Linear programming problem:** It refers to a problem in which the objective function, decision variables and constraints all have to be linear functions.
- **Graphical solution:** It refers to the method that is mostly used to solve a problem in which you have to find the highest or lowest point of intersection graphically.
- **Infeasibility:** It refers to a situation in a problem when no solution is satisfying all the constraints.
- Activation Function: An activation function in a neural network determines the output of a neuron, given its input or set of inputs. It introduces non-linearities to the network, allowing it to learn and represent complex relationships in the data.
- Summation Function: The weighted sum is a linear transformation of the input data, and the activation function introduces non-linearity, enabling the neural network to model more complex functions and relationships in the input data.

## 10.7 CASE STUDY: INDIVIDUAL CALLER ASSIGNMENT AND IDENTIFYING A BASIC FEASIBLE SOLUTION FOR A CALL CENTRE

A call centre named XYZ currently employs 1,000 people (telecallers). Each person has been trained on different sets of trainings, namely A, B, C, D, E and F. For instance, set A people are those who have undertaken trainings 1, 2, 3, 4, 5, and 6. Likewise, set B people are those who have undertaken another set of trainings. The numbers of employees belonging to each set are as follows:

| | | | |
|---|---|---|---|
| A | 200 | D | 100 |
| B | 200 | E | 250 |
| C | 100 | F | 150 |

The call centre manager has categorised the different types of calls into four major categories, namely X, Y, Z, and W. The IVR system has been designed in a way that allows allotment of a caller to the customers. The average daily distribution of calls is:

| | | | |
|---|---|---|---|
| X | 20% | Z | 15% |
| Y | 30% | W | 35% |

Notes

As per the training and skill sets possessed by an employee, he/she is able to resolve various calls in different time periods.

The resolution time taken by different employees belonging to different training sets for each type of call is as follows:

| Time (min) | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| X | 10 | 11 | 10 | 1 | 12 | 5 |
| Y | 8 | 4 | 9 | 12 | 7 | 5 |
| Z | 2 | 11 | 4 | 11 | 6 | 11 |
| W | 3 | 12 | 7 | 14 | 4 | 11 |

The objective of the call centre's manager is to complete the work of all callers in the least aggregate time. To achieve this, it is necessary to train all resources in different skills and decrease the resolution time dynamically. For this, ABC requires that such allocation of telecallers is done in real time. The manager, at present, is faced with a problem of minimising the total time taken by all the callers. This has to be done by making an assignment. To solve the current problem using an assignment, the manager assumes that the actual inflow of calls is somewhat constant. Therefore, we have a balanced transportation problem having the following distribution of calls:

| | |
|---|---|
| X | 200 calls |
| Y | 300 calls |
| Z | 150 calls |
| W | 350 calls |
| Total | 1000 calls |

Here, the total number of telecallers and the total number of calls are equal. However, it must be remembered that even if the total number of calls increases, the solution would still hold good.

This problem can be treated as a case of transportation problem where cost has been substituted by cost. In this problem, the decision-maker has to make the right set of assignments in order to minimise the total cost. This problem can be solved using simplex or tabular methods. Any transportation problem is solved in two steps which include identifying a basic feasible solution followed by finding an optimal solution. The basic feasible solution can be found by using one approach out of three approaches, namely North-West Corner Method, Minimum Cost Method and Penalty Cost Method. The manager used the North-West Corner Method to derive the basic feasible solution. In the North-West Corner Method, the maximum possible values are assigned on the North-West corner till all the callers have been exhausted. The steps involved are as follows:

| Time (min) | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| X | 20 | | | | | | 200 |
| Y | | | | | | | 300 |
| Z | | | | | | | 150 |
| W | | | | | | | 350 |
| | 200 | 200 | 100 | 100 | 250 | 150 | |

Notice that in the first assignment, the supply and demand are equal. If this was not the case, we would have assigned a value that is the minimum of the two. After this assignment, both the first row and the first column have been exhausted. Therefore, to make the next assignment, it is necessary to move to the cell (B, Y). Similarly, we will keep moving. The final assignment is as follows:

| Time (min) | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| X | 20 | | | | | | 200 |
| Y | | 200 | 100 | | | | 300 |
| Z | | | | 100 | 50 | | 150 |
| W | | | | | 200 | 150 | 350 |
| | 200 | 200 | 100 | 100 | 250 | 150 | |

**Source:** *https://www.analyticsvidhya.com/blog/2016/06/operations-analytics-case-study-level-hard/*

## QUESTIONS

1. What is the objective of the call centre's manager?

   (**Hint:** The objective of the call centre's manager is to complete the work of all callers in the least aggregate time.)

2. Why the problem encountered by an the manager is treated as a transportation problem?

   (**Hint:** The problem can be treated as a case of transportation problem where cost has been substituted by cost.)

3. How transportation problem can be resolved?

   (**Hint:** In transportation problem, the decision-maker has to make the right set of assignments in order to minimise the total cost. This problem can be solved using simplex or tabular methods.)

4. What conclusion can you draw from the final assignment?

   **(Hint:** First calculate the total time which will come out to be 7,550 minutes. Assign all As for all calls of type X. B will be assigned 33% of calls of type Y and the rest 66.66% calls of type Y will be assigned to C.)

5. You have to solve the same assignment problem using Excel Solver. Make all the calculations and comment on the difference between the total time taken for assignment calculated using Excel Solver and using North-West Corner method.

   **(Hint:** Using Excel Solver, we get total time = 3,750 minutes, and using North-West Corner Method, we get total time = 7,550 minutes.)

## 10.8 EXERCISE

1. What is linear programming? Explain its formulation.
2. Why financial sensitivity analysis is important for a business?
3. Which methods can we use to find the highest or lowest points of intersection in a graph?
4. Explain the concept of the transshipment problem.

5. Write short notes on the following concepts:
   a. Game Theory
   b. Dynamic Programming

## 10.9 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Overview of Linear Programming (LP) | 1. | Linear programming |
| | 2. | True |
| | 3. | 0 |
| | 4. | linear programming |
| | 5. | G.B. Dantzig |
| Problems Solved using Quantitative Techniques | 6. | True |
| | 7. | unit transportation cost |
| | 8. | False |
| | 9. | False |
| | 10. | single-pair shortest path |
| Additional Problems | 11. | True |
| | 12. | Overconfidence and confidence |
| | 13. | dynamic programming |
| | 14. | c. Input layer |

## 10.10 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Dantzig, G.B. Linear Programming and Extensions. Princeton, NJ: Princeton University Press, 1963.
- Karmarkar, N. "A New Polynomial-time Algorithm for Linear Programming." Combinatorica 4, 373-395, 1984.
- Klee, V.; Minty, G.J.; and Shisha, O. (Eds.). "How Good is the Simplex Algorithm?" In **Inequalities 3.** New York: Academic Press, 159-175, 1972.


### E-REFERENCES


- En.wikipedia.org. (2018). *Game theory.* [online] Available at: https://en.wikipedia.org/wiki/Game_theory
- People.brunel.ac.uk. (2018). *Linear programming - formulation.* [online] Available at: http://people.brunel.ac.uk/~mastjjb/jeb/or/lp.html
- Roberts, D. (2018). Systems of Linear Equations - Graphical Solution - MathBitsNotebook(A1 - CCSS Math). [online] Mathbitsnotebook.com. Available at: https://mathbitsnotebook.com/Algebra1/Systems/SYlinearGraphic.html

CHAPTER 11

# System Management and KPI

## Table of Contents

# Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the need of a system management
- Explain the concept of data quality
- Describe the concept of business metrics
- Discuss the importance of KPIs

## 11.1 INTRODUCTION

In the previous chapter, you have learned the concept of linear programming. Further, it has explained different techniques to solve complex organisational problems, such as graphical method, simplex method, duality concept and sensitivity analysis. The chapter has also explained different concepts such as transportation problem, shortest path problem, assignment problem, minimum spanning tree, network models with yield, game theory and dynamic programming.

System management signifies the management of Information Technology (IT) assets of an organisation in a centralised manner. It not only manages the IT assets but also tackles and resolves the problems related to IT assets. There are a number of system management solutions available that help small or big organisations in addressing their requirements which include monitoring of organisational network, management of servers, monitoring storage of data and handling organisational and client's devices, such as printer, laptop, mobile phone, etc. System management also includes sending or generating of notifications in case of failures, issues related to capacity of data and other events taking place over a network. Effective system management also handles compliance issues and is capable of enforcing company policies on employees regarding the usage of IT assets of the organisation. Like system management, Key Performance Indicators (KPIs) are also used by the organisations for achieving their business goals.

KPIs are used by organisations to measure the business goals in order to check the performance and determine whether the organisation is on success track or any improvement is required for its success. KPIs vary from organisation to organisation as some organisations focus on certain aspects of business while others on some other aspects of business. Each department of an organisation might also have different KPIs as per their tracking of their specific goals. KPIs help an organisation not only in tracking their goals, but also in determining the health of their practices to get the best results. KPIs can also be used outside the company. For example, an organisation can also use particular KPIs with the customers while creating a contract with them. The decided KPIs help both the organisation and the customer in tracking the success of their contract at present or in future. Sometimes, KPIs used by a department of an organisation are useful for another department of the same organisation. With the help of KPI-tracking software, companies can view the results after using KPIs on a single dashboard in real time.

This chapter first discusses about the need of system management and the data quality. Next, it explains business metrics and their importance. Further, this chapter discusses the KPI solution and types of KPIs. Towards the end, it discusses the benefits of the KPI software.

Notes

## 11.2 NEED FOR A SYSTEM MANAGEMENT

When the business of an organisation grows, its IT requirements also grow consequently. It is very hard to find an organisation that does not depend upon IT for its business. Therefore, it becomes very important for an organisation to effectively manage and provide a safeguard to its assets. For example, in order to keep systems in the running condition, an organisation uses management solutions, such as service desk management, patch management, etc. Sometimes, the organisation also uses single sign on authentication as a management solution to authenticate its employees while accessing the organisational resources to protect them from any unauthorised access. These management solutions help an organisation in enhancing the productivity of IT assets and its employees. System management solutions also help an organisation in protecting it against the following:

- Fallout from downtime or threats caused due to improper functioning of the systems
- Lost or stolen devices
- Sabotage in network
- Power outages
- Security breaches
- Identity theft
- Errors due to human actions
- Natural or man-made disasters

If any of the previously stated events occurs in an organisation, it may lead to any of the following:

- Financial loss
- Damage of brand
- Legal liabilities
- Other extremely unpleasant situations

Figure 1 shows the various system management solutions:

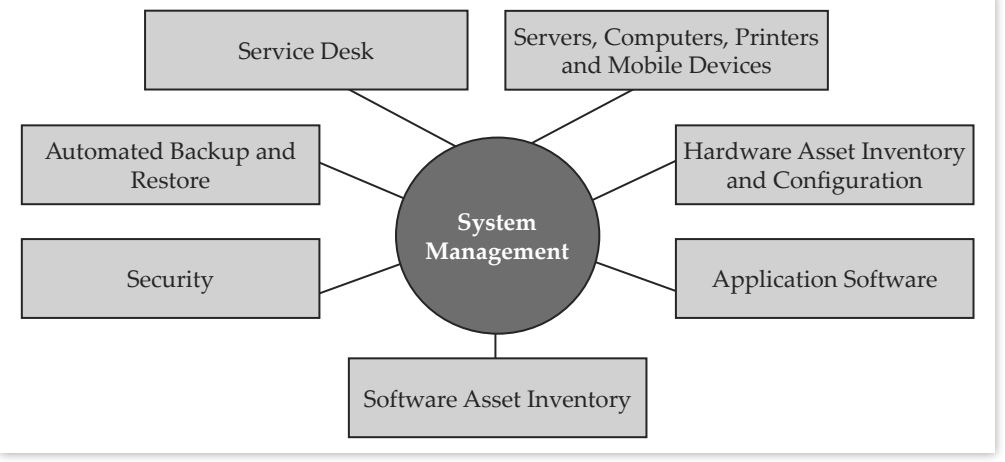



Figure 1: Functions of System Management

**Source:** *smallbusinesscomputing.com*

System management solutions also handle hardware inventory and their configuration. It also focuses on the security of IT assets by installation of anti-virus software in the devices and constantly updating the anti-virus software over a certain period of time. Besides security of devices, system management also focuses on the security of data stored in the devices. It emphasises on the back-up and restoring of data in case of failure of a storage device. The data gets restored from the central data repository or from the location where it is backed up. The quality of data is judged on several dimensions, which include completeness, consistency, conformity, accuracy, integrity and timeliness. It is not wrong to conclude that system management helps in achieving organisational goals by managing the IT assets of an organisation in every possible manner.

Besides so much usefulness of system management, its implementation in an organisation depends upon several factors, some of which are as follows:

- Size of an organisation
- Complexity of IT infrastructure
- Number of devices
- IT resources and expertise

Consider an example of a small organisation which has a small number of computers. System management for this organisation requires investment of more amount of money and time in comparison to the investment needed in managing each system individually. However, due to the lack of system management, the small and medium organisations are vulnerable to security risks. But, small and medium organisations can follow the following tips to safeguard themselves to a certain extent:

- Proper assessment of bottlenecks, gaps and vulnerabilities residing in an organisation's IT set-up.
- Search for vendors and solutions that are helpful in addressing the immediate IT issue, but are also capable of providing help that might be required by an organisation after a certain period of time.
- Search for vendors that may provide customised solutions as per your budget in managing the IT assets. Some popular vendors for small and medium organisations may include Dell KACE, HP Insight Manager, IBM Service Manager for Smart Business, etc.
- Evaluate the packages or offerings of vendors that suit your organisational needs and level of support vendors can provide. For example, the best vendor can be the one who can provide high level of IT expertise and 24/7 support at a lesser price as copared to other vendors.

**Self Assessment Questions**

1. System management solutions help organisations in protecting them against which of the following?
   a. Lost or stolen devices
   b. Power outages
   c. Security breaches
   d. All of these
2. The configuration of hardware devices includes _____ present in it and the type of software installed in it.

Notes

**Activity**

You are working as a cyber security expert in an organisation. You are assigned the role of safeguarding the IT assets of the organisation from different types of threats. Enlist the strategies that you will implement to safeguard the assets of the organisation.

## 11.3 DATA QUALITY

The quality of data is extremely important for an organisation. Therefore, improvement in the data quality is of utmost importance for an organisation in order to take the accurate decisions. The clearing out of bad data from the available data helps in its improvement. You can only be successful in your business when you make the right decisions. The right decisions are taken only on the availability of the right information. The availability of the right information also makes the decision-making process faster.

The organisational data is mainly stored in a data warehouse. Business intelligence solutions are used by organisations to access the data from the data warehouse to gain better insight of their business at any point of time. The data accessed must be of good quality to make faster decisions by the executives of an organisation. If the data available in the data warehouse is of bad quality, then the same set of data might give inaccurate results when accessed at different intervals of time. In this case, the executives have to hold the decision-making process which has to be performed on the basis of the results and have to work in the direction of finding the cause of the different results over the same set of data. The data quality issue might arise because of the following reasons:

- Due to patchwork between enterprise applications and operational systems
- Due to scattered or misplaced values in the data
- Due to outdated and duplicate records in the data
- Due to inconsistent (or undefined) standards and formats in the stored data
- Due to merger or acquisition of the data
- Due to human error while entering, editing, maintaining, manipulating and reporting of the data

In order to avoid inaccuarcy in maintaining data, business organisations have to implement data quality strategy which includes techniques for maintaining the data quality during business processes going on in the organisation. You can conclude that the data quality is all about cleaning of bad data which might be incorrect or invalid in some way. If an organisation wants to make sure that the data available is trustworthy, it has to understand the key dimensions of the quality of the data. Data quality dimensions are used by organisations to measure the level of accuracy of the data from time to time.

Figure 2 shows the key dimensions of the data quality:


**FIGURE 2: Dimensions of the Data Quality**

**Source:** *smartbridge.com*

Let us discuss each dimension in detail:

- **Completeness:** Data is considered complete even if the optional data is not present. For example, if the client's first name and last name are present in the data, but the middle name is not present as it is optional to provide the same, even in that case the record of a client is considered complete despite the fact that the middle name is not available in the company's database.
- **Consistency:** Consistency of the data means that it must reflect the same type of information across all the systems or units of an organisation. For example, an office of an organisation has already been closed, but the sales figures are getting reflected for it in the database. Another example of inconsistency is that an employee of an organisation has left it many years ago, but the salary status of that employee is still getting reflected in the organisation's database. If the employee has left the organisation, then his status across all the offices of the organisation must be consistent .
- **Conformity:** Conformity of the data means that the same set of standards have been followed for entering the data. The standard data definitions include data type, format and size of the data. For example, there must be conformity while entering the date of birth of an employee working in the organisation. The date of birth of the employee must be entered in the 'dd–mm–yyyy' format across all the offices of the organisation.
- **Accuracy:** Accuracy of the data reflects the degree of correctness of the data entered in the database. For example, the profit earned or the sales figures of an organisation must be entered correctly. These profit or sales figures are mathematical values

and reflect the business growth. Therefore, these must be entered with high level of accuracy.

- **Integrity:** Integrity of data means that the data entered in the database must have relationships or are connected appropriately. For example, the employees have several attributes and address is one of them. Therefore, an address relationship must exist with the employee records. If the addresses exist in the database without any employee record, then these are considered orphaned records in the database. When the related records are not linked properly, then this may lead to the duplication of the data in the database.
- **Timeliness:** The timeliness majorly relies on the expectations of the users. The availability of the data for an organisation in a timely manner is considered very important because of the following reasons:
  - Data must be available when an organisation wants to provide its quarterly business results.
  - Data must be available when an organisation wants to provide the correct information to its clients.
  - Data must be available when an organisation wants to check its financial activity at any point of time.

### 11.3.1 DIMENSIONS OF DATA QUALITY

Data quality plays a crucial role in system management and the effectiveness of Key Performance Indicators (KPIs). Poor data quality can result in inaccurate analyses, flawed decision-making, and overall inefficiency within an organisation. The dimensions of data quality in the context of system management and KPIs encompass the following:

- **Accuracy:** This refers to how close the data values are to the true values they represent. Accurate data is free from errors or mistakes.
- **Completeness:** It signifies whether all the required data is available. Complete data contains all the necessary information without any missing values.
- **Consistency:** This refers to the absence of contradictions or discrepancies in the data. Consistent data shows uniformity and coherence across different sources or within the same dataset.
- **Timeliness:** This measures how up-to-date the data is. Timely data is relevant and reflects the current state of affairs without significant delays.
- **Validity:** This refers to information that is relevant and applicable for the intended purpose. Valid data adheres to the defined rules and standards.
- **Precision:** This refers to the level of detail in the data. Precise data is granular, providing specific and accurate information without ambiguity.
- **Reliability:** This refers to the data that is consistently relied upon to make decisions or draw conclusions. It is data that can be trusted for its accuracy and consistency over time.

- **Accessibility:** This dimension refers to how easily and readily data can be accessed and retrieved by authorised users. Accessibility ensures data is available when needed.
- **Security and privacy:** This refers to the data security that ensures that the information is protected from unauthorised access, breaches, or misuse. Privacy concerns safeguard sensitive information and ensure compliance with regulations regarding data privacy.
- **Interpretability:** It refers to the ease with which data can be understood and analysed by users. Interpretability ensures that data is presented in a way that users can comprehend and derive insights from it effectively.

## 11.3.2 BENEFITS OF DATA QUALITY

Ensuring high data quality in system management and Key Performance Indicators (KPIs) offers a multitude of advantages that positively impact decision-making, operational efficiency, and overall organisational success. Following are the key benefits:

- **Accurate decision-making:** High-quality data ensures accurate and reliable information for decision-making, empowering executives and managers to make well-informed choices aligned with organisational goals.
- **Enhanced customer satisfaction:** High data quality ensures accurate customer information, leading to more personalised interactions, improved customer service, and increased satisfaction.
- **Better planning and forecasting:** Improved data quality enhances the precision of forecasts, enabling organisations to better anticipate trends, demand, and market changes.
- **Compliance and regulatory adherence:** High data quality ensures compliance with regulations, reducing the risk of legal issues and penalties related to data privacy and accuracy.
- **Cost savings:** High-quality data reduces errors and inaccuracies, preventing costly mistakes and associated expenses, leading to significant cost savings.
- **Improved collaboration:** Data quality facilitates efficient collaboration across departments and teams, fostering productivity and consistency.

## 11.3.3 DATA QUALITY MANAGEMENT TOOLS

Data quality management tools play a critical role in ensuring the accuracy, reliability, and consistency of data used for system management and Key Performance Indicators (KPIs). These tools aid organisations in monitoring, cleansing, and enhancing their data to improve overall data quality. Following are the common types of data quality management tools used in system management and KPI monitoring:

- **Data profiling tools:** These tools help to analyse data structure, completeness, uniqueness, and distribution to assess overall data health before use in KPI calculations or system management.

Notes

- **Data cleansing (Data scrubbing) tools:** These tools help to identify and correct errors or inconsistencies in data, such as misspellings, duplicates, and formatting issues, ensuring accuracy for trustworthy KPIs.
- **Data quality monitoring tools:** These tools help to continuously track data quality, providing alerts or reports when deviations or issues are detected to maintain ongoing data quality for KPIs and system management.
- **Data standardisation tools:** These tools help to enforce consistency by transforming data into a common format, reducing the risk of inconsistencies in KPI calculations.
- **Master Data Management (MDM) tools:** These tools help centralise and manage master data to ensure consistency across systems, providing a single, authoritative source for critical data supporting accurate KPIs.
- **Data quality dashboards:** It helps to provide visual representations of data quality metrics for easy monitoring, allowing users to assess overall data quality related to KPIs and system management.
- **Data quality rule engines:** It helps to automate enforcement of data quality rules and policies, ensuring data conforms to predefined standards to maintain quality for accurate KPIs.
- **Data quality scorecards:** It helps to assign scores to different aspects of data quality, offering a quantifiable way to measure and track improvements over time in the context of KPIs.
- **Metadata management tools:** They help track and manage metadata, providing insights into the origin, usage, and quality of data to support transparency and accountability in data quality processes.
- **Data governance tools:** These tools help facilitate the establishment and enforcement of data governance policies to ensure data quality throughout its lifecycle, creating a structured framework for system management and KPIs.

### Self Assessment Questions

3. Which of the following is not a data quality dimension?
   a. Completeness
   b. Consistency
   c. Identity theft
   d. Conformity
4. _____ of the data means that the same set of standards have been followed for entering the data.
5. Which data quality management tool is designed to analyse data structure, completeness, uniqueness, and distribution before utilisation in KPI calculations or system management?
   a. Data cleansing tools
   b. Data quality monitoring tools
   c. Data profiling tools
   d. Metadata management tools

### Activity

Search and prepare a report on popular data quality software used in a business.

## 11.4 BUSINESS METRICS

The main purpose behind the measurement of business objectives is to track an organisation's financial expenses or investments made. Business metrics help an organisation in judging its progress towards the set short-term or long-term goals. Business metrics are very important for the key stakeholders of business. Key stakeholders are those people whose input plays a significant role in an organisation.

Business metrics are stated by an organisation in its mission statements which are nothing but organisational communication with its customers and general public. Sometimes, an organisation also includes business metrics in its workflows. Business metrics are important for the different departments of an organisation according to their interests. For example, the marketing department measures the success of its conducted campaigns; whereas, the sales department uses business metrics to track sales over a certain period of time or at any instant.

In other words, there must be some context attached to business metrics as they have no worth without it. An organisation generally considers metrics in terms of the existing benchmarks, approaches and goals. People often confuse business metrics with Key Performance Indicators (KPIs), but there exists a fine line between them.

Figure 3 shows some important business metrics:

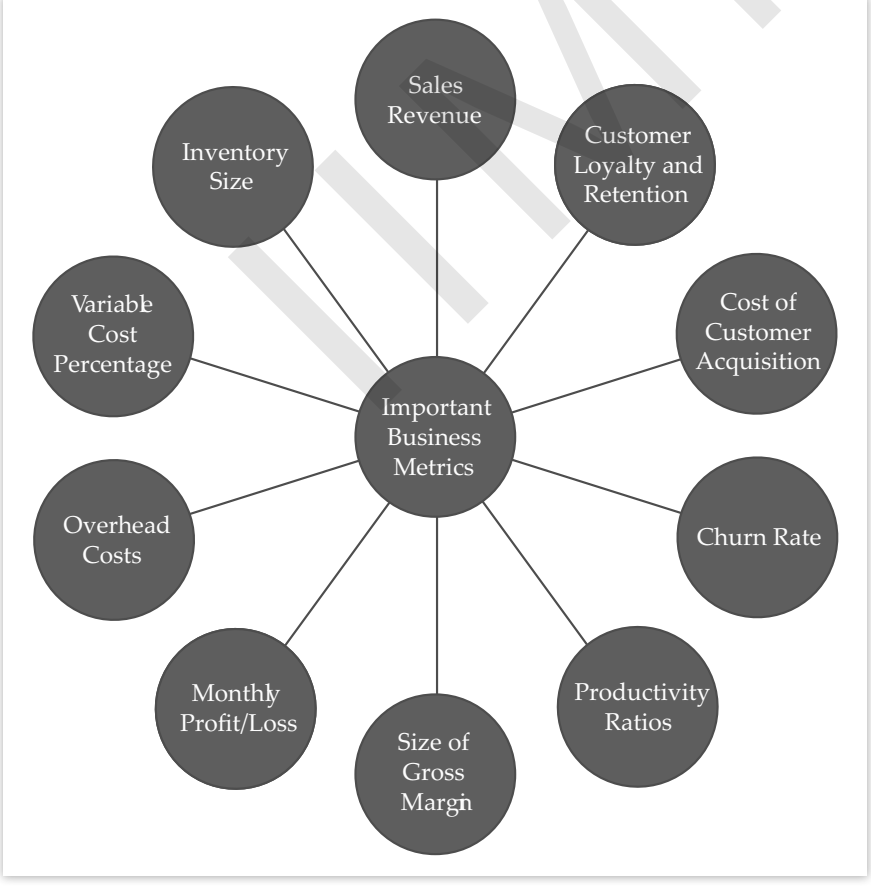



FIGURE 3: Displaying Important Business Metrics

Notes

Let us discuss each business metric in detail:

- **Sales revenue:** This business metric is used for assessing the sales made by an organisation. The sales revenue signifies the total profit earned after subtracting the invested money from the total sales made. The money is invested in different business processes, such as marketing campaigns, advertising, price evaluation of products and others for increasing or making sales.
- **Customer loyalty and retention:** This business metric measures how an organisation lures the customers in order to increase its sales and retain them for long-term business profit. The long-term relatioship with the customers helps in making long-term profits by an organisation. An organisation conducts surveys in order to get feedback from customers. Sometimes, an organisation also gets feedback from direct interaction with customers or by performing other kinds of analyses and uses the feedback to enhance more satisfaction while offering the products or services to the customers. The feedback implementation helps an organisation in generating more loyalty among the customers and retaining strong customer base.
- **Cost of customer acquisition:** This business metric helps in assessing the new customers acquired by investing money in different processes implemented to acquire them. This metric is calculated by dividing the total expense made by the organisation in acquiring customers by total number of new customers over a certain interval of time.
- **Churn rate:** This business metric is used to assess the cost of acquiring the lost customers of the organisation. This metric indicates an increase in the cost of acquiring customers and a decrease in the customers' values to an organisation in the long term.
- **Productivity ratios:** This business metric is used to assess the productivity of the employees working in an organisation. This business metric is calculated by the total revenue generated by the employees of a particular organisation and is then compared with the productivity of the employees of another organisation to gain deeper insight of the effectiveness of the employees. This metric finds its application in almost any area of business.
- **Size of gross margin:** This metric is calculated by subtracting the cost of the products sold from the total sales revenue and divided by the total sales revenue. The size of the gross margin is then converted into percentage. If the value obtained is higher, then the organisation can spend more money on other costs it has incurred and can generate more profit.
- **Monthly profit/loss:** This metric helps in measuring the fixed and variable operational costs paid on monthly basis. The costs might include office rent, insurance paid, payments made against mortgage, taxes or salaries paid, etc.
- **Overhead costs:** This business metric helps in assessing the fixed costs that do not rely on the production levels of the products or services. The fixed cost includes salaries paid to the employees and the rent paid against the usage of various services. The overhead costs do not get affected by the earning and growth of the business. Therefore, their tracking must be done separately.

- **Variable cost percentage:** This business metric is used to assess the cost of goods. The cost of goods is variable as it depends upon various factors, such as cost of raw materials, labor charges, shipping cost, and other costs related to the production or delivery of goods. Therefore, the cost of goods might increase from time to time with an increase in the charges of the different factors.
- **Inventory size:** This business metric is used to track the inventory that is ready to be sold at any instant. This metric is also used to assess how much inventory will be ready for sale after a certain period of time. An organisation always keeps a close eye over the inventory as it is the primary source of its income.

Business performance metrics play a significant role in conveying the information to organisational executives, investors, and clients to make them aware about the overall organisational performance. The simplest, easiest and the mot effective method to assess your company's performance is by keeping the key business metrics on a dashboard. Various departments of an organisation keep a close eye on different metrics. So, the dashboard varies from one department to another, and also from one organisation to another.

### 11.4.1 BENEFITS OF TRACKING BUSINESS METRICS

Tracking business metrics, also known as Key Performance Indicators (KPIs), provides numerous benefits for organisations across various industries. Here are some key advantages:

- **Performance monitoring:** Metrics offer a quantitative way to assess the performance of different aspects of the business, providing insights into what is working well and what needs improvement.
- **Informed decision-making:** By tracking metrics, organisations can make decisions based on data rather than intuition alone. This leads to more informed and strategic decision-making processes.
- **Goal alignment:** Metrics help align individual and team objectives with the overall goals of the organisation. This ensures that everyone is working towards common objectives, promoting a sense of unity.
- **Identifying areas for improvement:** Metrics highlight areas of the business that may require attention or optimisation. This allows organisations to proactively address weaknesses and enhance overall performance.
- **Resource optimisation:** Metrics assist in the efficient allocation of resources by identifying high-impact activities. This ensures that time, money, and manpower are used effectively to achieve business goals.
- **Customer satisfaction:** Metrics related to customer satisfaction and engagement provide insights into the effectiveness of products, services, and customer interactions. This information helps in enhancing customer experiences.
- **Risk management:** Metrics aid in the early identification of potential risks and challenges. This allows organisations to develop strategies to mitigate risks and navigate uncertainties.

NOTES

- **Competitive analysis:** By comparing metrics with industry benchmarks, organisations can assess their competitive position. This information helps in identifying opportunities for improvement and innovation.
- **Employee productivity and engagement:** Metrics related to employee productivity and engagement help organisations assess workforce effectiveness. This information can guide strategies to improve employee satisfaction and performance.
- **Adaptability and agility:** In a rapidly changing business environment, metrics enable organisations to adapt quickly to market conditions, customer preferences, and emerging trends.

EXHIBIT

**BUSINESS METRICS IN DECISION MAKING**

Business metrics share many similarities with KPIs, but the property which distinguishes them from each other is that business metrics focus on the overall development of the business rather than the KPIs, which provide information and details about a particular domain for which it is focused. Business metrics play a critical role in optimising business and discovering loopholes. These metrics prove to be a key element in diagnostics for a certain problem inside the organisation. Taking the example of a technology company which now also wants to start customer-centric division for itself, which includes direct communication with customers along with providing them with various services, such as insurance and loan financing, needs to get a thorough evaluation of its strategy and soft points to avoid any future problems. Business metrics can be a great resource for diagnostic evaluation as they provide correct insight of the company's past records in various domains. Assuming a case where the company wants to know if it has a breach-proof security system installed to protect the customers' data, metrics can come in handy to show how many times the company has faced a security threat and which were the weak areas. Those areas can be easily identified using metrics and diagnosis can be narrowed down, ultimately resulting in saving time and efforts. Also, using metrics, business managers can react to dynamic customer behavior in a flexible way.

SELF ASSESSMENT QUESTIONS

6. A _____ is defined as a quantifiable measure that is used by an organisation for tracking, monitoring and assessing the success or failure of its business process.
7. Business metrics are stated by an organisation in its mission statements which are nothing but organisational communication with its customers and general public. (True/False)
8. _____ provides insights into the effectiveness of products, services, and customer interactions, helping in enhancing customer experiences.

## 11.5 KEY PERFORMANCE INDICATORS (KPIs)

KPIs are very helpful for an organisation or its department, team and executives in order to take remedial steps before happening of events that may lead to huge financial loss to an organisation. KPIs are used to define targets throughout the business for meeting strategic objectives. KPIs also help an organisation in focusing over common objectives and ensure that they remain aligned in the organisation. Therefore, it is very important for an organisation to decide what it wants to measure exactly before using KPIs.

### 11.5.1 NEED OF KPIS IN BUSINESS

Key Performance Indicators (KPIs) play a crucial role in businesses for several reasons. They serve as measurable metrics that help organisations evaluate their performance, track progress towards goals, and make informed decisions. Following are some key reasons highlighting the need for KPIs in business:

1. **Goal Alignment:**
   - **Purpose:** KPIs help align individual and team efforts with the overall goals and objectives of the organisation.
   - **Benefit:** It ensures that everyone is working towards common objectives, fostering a sense of purpose and direction.
2. **Performance Monitoring:**
   - **Purpose:** KPIs provide a quantifiable way to measure and monitor the performance of various business processes and activities.
   - **Benefit:** It enables organisations to identify areas of success and areas that need improvement, facilitating proactive decision-making.
3. **Strategic Planning:**
   - **Purpose:** KPIs contribute to the development and refinement of strategic plans by providing data-driven insights.
   - **Benefit:** It helps organisations set realistic goals, prioritise initiatives, and allocate resources effectively.
4. **Data-Driven Decision-Making:**
   - **Purpose:** KPIs offer a basis for making decisions grounded in quantitative data rather than intuition alone.
   - **Benefit:** It enables informed decision-making, reducing the risk of relying on subjective judgement or incomplete information.
5. **Continuous Improvement:**
   - **Purpose:** KPIs support a culture of continuous improvement by identifying areas for optimisation.
   - **Benefit:** It encourages learning from successes and failures, fostering a mindset of ongoing enhancement in processes and strategies.

6. **Resource Allocation:**
   - **Purpose:** KPIs assist in the efficient allocation of resources by highlighting high-impact activities.
   - **Benefit:** It ensures that time, budget, and manpower are directed towards activities that contribute most to the organisation's success.
7. **Customer Focus:**
   - **Purpose:** KPIs related to customer satisfaction and engagement help organisations understand and meet customer needs.
   - **Benefit:** It enables businesses to enhance customer experiences, leading to improved loyalty and positive brand perception.
8. **Benchmarking and Competitive Analysis:**
   - **Purpose:** KPIs allow organisations to benchmark their performance against industry standards and competitors.
   - **Benefit:** It provides insights into competitive positioning and identifies areas where the business can outperform competitors.
9. **Risk Management:**
   - **Purpose:** KPIs contribute to the early identification and monitoring of potential risks and challenges.
   - **Benefit:** It allows organisations to develop strategies for risk mitigation, enhancing resilience and adaptability.
10. **Employee Engagement and Productivity:**
    - **Purpose:** KPIs related to employee performance and satisfaction help organisations manage and optimise their workforce.
    - **Benefit:** It promotes a positive work environment, improving employee satisfaction, productivity, and retention.
11. **Financial Health:**
    - **Purpose:** Financial KPIs provide insights into the overall financial health of the business.
    - **Benefit:** It helps in monitoring profitability, cash flow, and other financial indicators crucial for sustainability and growth.
12. **Transparency and Accountability:**
    - **Purpose:** KPIs promote transparency by providing clear performance indicators.
    - **Benefit:** It enhances accountability as teams and individuals are held responsible for their contributions to organisational objectives.

### 11.5.2 TYPES OF KPIs

As discussed earlier, different types of KPIs are used in the organisations to track business goals.

Some indicators of KPIs are shown in Figure 4:

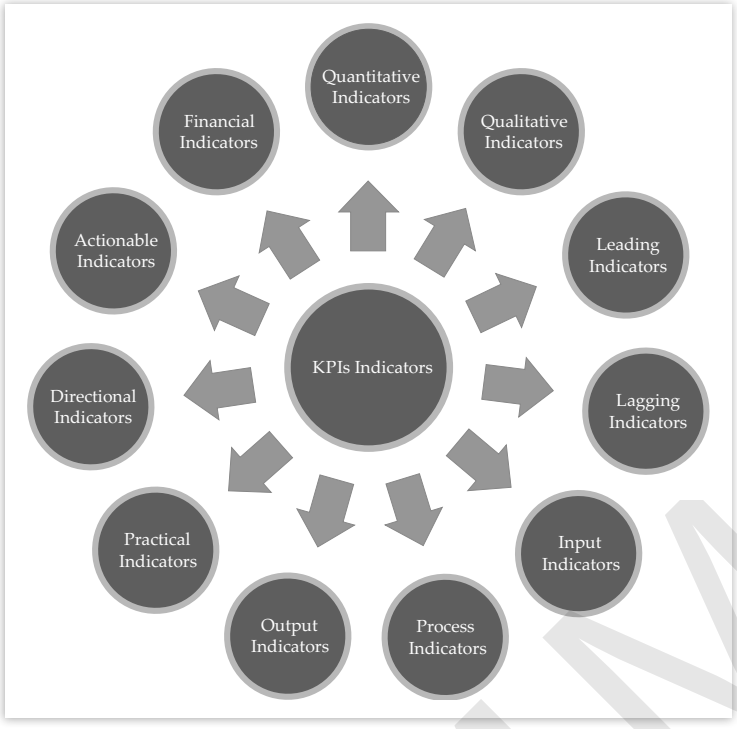



Figure 4: Displaying Broad Categories of KPIs

The description of each indicator in KPIs is as follows:

- **Quantitative indicators:** These can be stated using a number.
- **Qualitative indicators:** These cannot be stated using a number.
- **Leading indicators:** These can be used for predicting the result of a process.
- **Lagging indicators:** These can be used for representing the success or failure after the execution of the business processes.
- **Input indicators:** These are used for measuring the quantity of resources used for generating the desired result.
- **Process indicators:** These represent the efficiency or the productivity of the business processes.
- **Output indicators:** These denote the outcomes of the business process activities.
- **Practical indicators:** These indicators interface with existing processes in an organisation.
- **Directional indicators:** These denote whether an organisation's performance is getting better or not.
- **Actionable indicators:** These are used by an organisation to make changes.
- **Financial indicators:** These are used for the performance measurement of an organisation.

Notes

Some types of KPIs across different business functions in an organisation are shown in Figure 5:


Figure 5: Displaying the Types of KPIs

Let us discuss each of the KPIs in detail.

- **IT operations and project execution KPIs:** KPIs in this area are used for determining the execution of operations related to IT and projects. Some KPIs used for judging IT operations and project execution are as follows:
  - **Mean-time between failure (MTBF):** It is used for tracking the time between two failure situations.
  - **Estimate to complete:** The amount required to complete a project.
  - **Amount spent per month:** This KPI is used for determining the amount spent per month on the ongoing projects in a particular currency.
  - **Average initial response time:** This KPI is used to determine the average time that is generally taken by the service desk of an organisation in responding against the reported incident by the customer.
- **Project management KPIs:** These are the KPIs that help in managing the project. Some project management KPIs are as follows:
  - **Actual cost of work done:** It enables an organisation in determining the actual money in the activities that are performed till their completion.
  - **Percentage of milestones missed:** It allows the executives of an organisation in keeping track of the percentage of projects which have missed their objectives or goals.
  - **Estimate at completion:** It is used to determine the actual cost required in completing the project and the actual cost required to complete the remaining work of the project.
  - **Cost of managing processes:** This allows in estimating the periodic cost that is required for managing the processes.

- **Financial Performance KPIs:** These KPIs are used to measure the financial performance of an organisation. Some financial performance KPIs are as follows:
  - **EV/EBITDA:** EV stands for Enterprise value and EBITDA stands for Earnings before interest, taxes, depreciation and amortisation. The ratio of these two helps you in analysing the debt value of an organisation.
  - **Return on investment (ROI):** It is used to evaluate the performance of an organisation by dividing net profit by net worth.
  - **Debt-equity ratio:** It is used for measuring the proportion of shareholders' equity to the debt used for financing the assets of an organisation.
  - **Operating margin:** It is used to evaluate the strategy of an organisation's pricing and its operating efficiency.
  - **Return on assets/return on equity (ROA/ROE):** ROE is used to evaluate the money taken from the shareholders. On the other hand, ROA signifies the measure of an organisation's profitability to its assets.
- **Human resource performance KPIs:** These are used in an organisation to measure the different aspects of human resources. The different types of Human Resource KPIs are as follows:
  - **Revenue per employee:** It is used to evaluate the productivity of an organisation's workforce. It is used for determining the amount of sales made per employee and also how effectively the human resources of an organisation are getting utilised.
  - **Employee satisfaction index:** It allows an organisation to determine how much satisfied the employees are with the organisation.
  - **Salary competitiveness ratio:** It helps in gathering the data about how much the competitor organisations are paying to the employees. Thus, it helps in determining the salary levels of an organisation in comparison to the other organisations.
  - **Human capital ROI:** It helps in measuring the return on capital invested on an employee in terms of pay and benefits.
- **Supply chain and operational performance KPIs:** KPIs in this area are used for improving the experience of the customers of an organisation. Some KPIs used in this area are as follows:
  - **Order fulfillment cycle time:** This KPI is used to measure the total time taken from ordering a product till its delivery to the customer. This KPI helps in developing the customer responsiveness towards the organisation. Moreover, it also helps in determining the time required for completing a manufacturing order.
  - **Yield:** This metric is used for improving the quality of the products after measuring it. It denotes the percentage of correctly manufactured products which do not require rework or do not need to be scraped.
  - **Throughput:** It is used to evaluate the speed of the production process on the basis of inputs and outputs.

Notes

- **Consumer insights and marketing KPIs:** These KPIs are used for getting insights of a customer. Some of the KPIs used in this area are:
  - **Market growth rate:** It is used for analysing the change on the basis of the given number of consumers in a particular market after a certain interval of time.
  - **Customer satisfaction index:** This KPI is used for evaluating the performance of the products and services, whether they have met or surpassed the customer's expectations or not.
  - **Social networking footprint:** This KPI is used for evaluating the presence of an organisation on social media.
  - **Brand equity:** It is used for measuring the premium that a brand name may provide to a product.
  - **Customer lifetime value:** This KPI is used to evaluate the revenue that can be generated till the customer remained in relationship with the company.
  - **Customer acquisition cost:** This KPI is used for determining the cost incurred in marketing and campaigning to acquire new customers over a certain period of time.

### 11.5.3 KPI SOFTWARE

Different organisations provide different KPI solutions as per the need of an organisation. Different KPI solutions provide different benefits to organisations, some of which are as follows:

- Enable users, executives and organisations to measure and handle targets and objectives.
- Enable tracking of project performance and provide reporting to the shareholders of the project.
- Allow conveying the updated information to the right people in the organisation. Provide a snapshot of the performance against the targets set by an organisation.
- Integrate all KPI data at one place easily and there is no requirement of depending upon the creation of the spreadsheets for assessment.
- Display the performance of the departments residing at different locations at a single location, and thus provide the integrated view of the organisation irrespective of its geographically separated departments.
- Enable an organisation to view transparency in performance at any instant.
- Allow its online access to the executives of the organisation.
- Enable an organisation to assess the impact of the initiatives taken by it.
- Allow an organisation to evaluate the contributions made by an employee, department and office accurately and clearly.
- Allow an organisation to assess the impact of its vision and strategy.
- Allow an organisation to assess its weakness and set KPIs for the improvement in its performance.

- Enable an organisation to judge its operational performance.
- Help an organisation in gaining competitive advantage over others.

Some examples of popular KPI software are as follows:

1. **Scoro KPI dashboard:** The Scoro KPI dashboard software allows an organisation to see every aspect of its business on one or several dashboards. It also allows the organisation to track KPIs related to a project, work and finance at any instant.
2. **Datapine:** It allows an organisation to view and monitor most significant KPIs at a single location. Some more features of this software include advanced analytics, automated reporting, highly interactive dashboard and intelligent warning or alarming system.
3. **Inetsoft dashboard:** This is an analytical dashboard and reporting software. Some main features of this include data modeling, online data mashup, embedding of dashboards, highly secure infrastructure with high performance, etc.
4. **Tableau:** It is the best solution for those organisations that have clients with very few number of users and wish deployment of solution in multiple organisations. Some main featutres of Tableau include adding of the number of users as the needs grow, ability of refreshing the data automatically using Web applications, such as Google analytics, Salesforce, etc., allowing the site administrators to manage authentication and permissions to users and data.
5. **SimpleKPI:** It is a powerful and higlhly flexible KPI dashboard. It allows you to customise your dashboard and reports.

Besides the KPI software discussed above, the names of some more popular KPI software are: Smartsheet, Bilbeo, DATAZEN, Databox, InfoCaptor, KPI Fire and Dasheroo.

### 11.5.4 BARRIERS OR ISSUES IN IMPLEMENTING KPIS

The successful implementation of Key Performance Indicators (KPIs) is indeed crucial for effective business performance monitoring, but organisations often face various barriers and challenges. Some common issues encountered during KPI implementation are as follows:

- **Lack of clear strategy:** Implementing KPIs without a defined strategic direction can lead to confusion about which metrics are crucial for business success.
- **Unclear definition and measurement:** If KPIs are not well-defined or their measurement is ambiguous, it becomes challenging to track progress accurately.
- **Data quality and accuracy:** Inaccurate or poor-quality data can significantly impact KPI analysis, leading to misleading insights and flawed decision-making.
- **Resistance to change:** Employees or stakeholders might resist KPI implementation due to fear of change or uncertainty about its impact on their roles.
- **Insufficient resources:** Inadequate financial, human, or technological resources can hamper the effective implementation and monitoring of KPIs.
- **Complexity and overload:** A surplus of KPIs or overly complex metrics can overwhelm teams, causing them to lose focus on critical objectives.

Notes

- **Lack of automation:** Manual data collection and analysis processes can be time-consuming and prone to errors, affecting the timely utilisation of KPI insights.
- **Inadequate technology infrastructure:** Lack of appropriate tools or technological support can hinder the smooth collection, processing, and visualisation of KPI-related data.
- **Poor communication and training:** Inadequate communication about the importance of KPIs and insufficient training on their use can lead to misunderstanding and underutilisation.
- **Lack of ownership and accountability:** Without clear ownership of KPIs or accountability for their performance, it is challenging to drive necessary actions for improvement.
- **Inability to adapt:** Rigidity in KPI frameworks can prevent adaptation to changing business environments, rendering them less relevant over time.
- **Security and privacy concerns:** Concerns about data security and privacy breaches can impede the willingness to collect or share sensitive information for KPI analysis.
- **Overemphasis on short-term results:** Focusing excessively on short-term gains might undermine the significance of long-term strategic goals, impacting overall business growth and sustainability.

**Self Assessment Questions**

9. KPIs refer to _____.
   a. Key Performance Indicators
   b. Key Performance Indications
   c. Key Performer Indications
   d. None of these
10. Which of the following comes under financial performance KPIs?
   a. Actual cost of work done
   b. Debt-equity ratio
   c. Estimate at completion
   d. Operating margin
11. _____ indicators represent the efficiency or the productivity of the business processes.
12. What is a common challenge organisations face in implementing Key Performance Indicators (KPIs) related to data quality?
   a. Lack of clear strategy
   b. Resistance to change
   c. Data quality and accuracy
   d. Insufficient resources

## 11.6 SUMMARY

- System management signifies the management of Information Technology (IT) assets of an organisation in a centralised manner.
- The system management is an umbrella term which includes a lot of management solutions.
- The quality of data is extremely important for an organisation. The bad-quality data provides inaccurate results and makes the decision-making process slower in an organisation.

- Business intelligence solutions are used by the organisations to access the data from the data warehouse to gain better insight of their business at any point of time.
- If the data available in the data warehouse is of bad quality, then the same set of data might give inaccurate results when accessed at different intervals of time.
- A business metric is defined as a quantifiable measure that is used by an organisation for tracking, monitoring and assessing the success or failure of its business process.
- KPIs are very helpful for an organisation or its department, team and executives in order to take remedial steps before the happening of events that may lead to huge financial loss to an organisation.
- KPIs also help an organisation in focusing over common objectives and ensure that they remain aligned in the organisation. Therefore, it is very important for an organisation to decide what it wants to measure exactly before using KPIs.
- The dimensions of data quality, including accuracy, completeness, consistency, and timeliness, among others, collectively contribute to the reliability and usefulness of data.

## 11.7 KEY WORDS

- **Mean-time between failure (MTBF):** It is used for tracking the time between two failure situations.
- **Churn rate:** This business metric is used to assess the cost of acquiring the lost customers of the organisation.
- **Completeness:** Completeness of the data refers to the meeting of expectations of the executives of an organisation.
- **Integrity:** Integrity of data means that the data entered in the database must have relationships or are connected appropriately.
- **Consistency:** Consistency of the data means that it must reflect the same type of information across all the systems or units of an organisation.

## 11.8 CASE STUDY: VALUECODERS KRA AND KPI MANAGEMENT SOFTWARE

For any organisation, it is necessary to make its employees understand the various areas of work and to tell them which areas of work are important and need attention. In addition, areas to be focussed on and the areas that need measuring and monitoring also need to be mentioned. All this can be achieved by having in place a set of well-defined Key Responsibility Areas (KRAs) and Key Performance Indicators (KPIs). The process of defining, measuring, monitoring and evaluating the KRAs and KPIs can be done manually or by automating it using software resource. Using the KRA and KPI Management software eases up the work of the people responsible for managing this process.

PR HR Consultants (name changed) is a US-based HR consultancy organisation. PR was in the process of automating task and operation management system. Using

Notes

the automated system, they wanted to improve the employee rating process, KRA description and privacy of employee reviews. Over time, PR decided to move over to an automated solution to manage its KRAs and KPIs. There were two main reasons due to which PR took this decision – first, to reduce costs as the cost of managing KRA and KPI were high using manual processes; and second, they wanted a more user-friendly system having various customised functionalities which were not available with the manual system. For this purpose, they hired ValueCoders and wanted to address the following tasks:

- Developing software that could be personalised
- Developing a seamless, scalable and secure software
- Transferring the data from the old database into the new one
- Making the software as user-friendly as possible
- Developing an easy-to-manage software
- Properly defining the permissions and access levels for each level

While executing this project, various challenges faced by the ValueCoders team were as follows:

- Software had to be developed using MEAN stack instead of MySQL.
- Software had to be designed in such a manner that the end reports generated in Excel could be directly fetched from the database.
- Software had to be designed to ensure security, privacy and authenticity from a user's perspective.

To make the software as efficient as possible, ValueCoders implemented denormalisation method using which a user could generate customised reports on the day, month or quarter basis as requested by the user. In addition, different modules were used to retrieve data. The developers also ensured that different grades of employees got different permissions and access rights.

The software development project was divided into five major phases as shown in Figure (A):

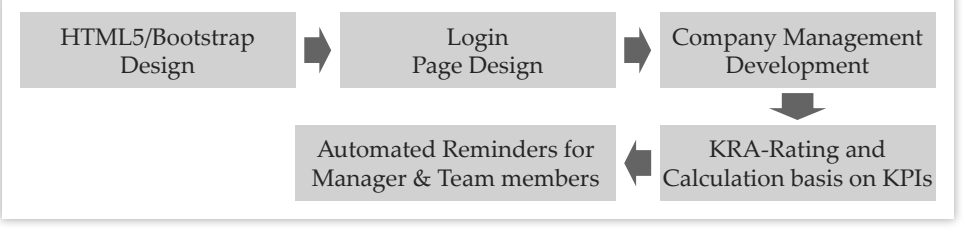



Figure (A): Five Phases of Software Development Project

A combination of technologies is used for the front end. AngularJS was used at the client's end. The front end was kept as a single-page application, and Angular UI router and RequireJS were used for loading specific pages. Any http call from Angular API is redirected to ExpressJS API. At last, the ExpressJS API directs the final response back to AngularJS API.

This process is shown in Figure (B):

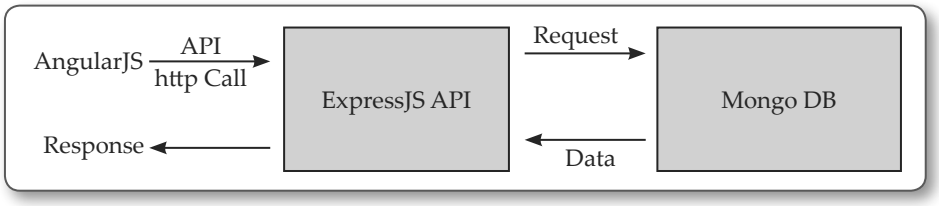



Figure (B): Process of Software Development Project

The login into this software was authenticated using three methods:

1. Navigation-based login
2. Direct URL login
3. API-level secure login

As a result of the development process, ValueCoders was able to develop a high-performance and easy-to-use, feature-rich and well-structured KRA and KPI management software. The major features of the management software include:

- Building the software using MEAN Stack instead of MySQL
- Generating end reports in Excel directly from the database
- Keeping the software as secure as possible
- Preparing the KRA and KPI management software well within the time

The product helped the organisation in gaining repeat business.

**Source:** *https://www.valuecoders.com/case-studies/online-kra-kpi-management-software*

## QUESTIONS

1. How are KRAs and KPIs defined?

   (**Hint:** The process of defining, measuring, monitoring and evaluating the KRAs and KPIs can be done manually or by automating it using software resource.)

2. How did PR HR consultants want to improve the employee rating process?

   (**Hint:** By using the automated system, they wanted to improve the employee rating process, KRA description and privacy of employee reviews.)

3. List the reasons which led PR to develop KRA and KPI software.

   (**Hint:** PR wanted to reduce costs and to have in place a more user-friendly system.)

4. Why did PR HR Consultants hire Value coders?

   (**Hint:** PR HR hired ValueCoders and wanted to address the following tasks:

   - Developing software that could be personalised
   - Developing a seamless, scalable and secure software)

5. List one feature that was specifically designed for managers.

   (**Hint:** Automated reminders.)

Notes

## 11.9 EXERCISE

1. What do you understand by system management in an organisation?
2. Discuss the need of a system management in an organisation.
3. Explain the different dimensions of data quality.
4. Discuss the importance of business metrics for an organisation.
5. Write a short note on KPI and its needs.

## 11.10 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Need for a System Management | 1. | d. All of these |
| | 2. | firmware |
| Data Quality | 3. | c. Identity theft |
| | 4. | Conformity |
| | 5. | c. Data profiling tools |
| Business Metrics | 6. | business metric |
| | 7. | True |
| | 8. | Customer satisfaction metrics |
| Key Performance Indicators (KPIs) | 9. | a. Key Performance Indicators |
| | 10. | b. Debt-equity ratio |
| | 11. | Process |
| | 12. | c. Data quality and accuracy |

## 11.11 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Albright. (2014). *Business Analytics: Data Analysis & Decision Making*. Cengage Learning.
- Bartlett, R. (2013). *A practitioner's Guide to Business Analytics: Using Data Analysis Tools to Improve your Organisation's Decision Making and Strategy*. New York: McGraw-Hill.


### E-REFERENCES


- What is Systems Management, and Why Should You Care? (1970, August 21). Retrieved November 16, 2018, from https://www.smallbusinesscomputing.com/news/article.php/3928971/What-is-Systems-Management-and-Why-Should-You-Care.htm
- What are Data Quality Dimensions? | Experian. (2015, April 15). Retrieved November 16, 2018, from https://www.edq.com/uk/glossary/data-quality-dimensions/
- Dashboard Insight - Dashboard Design and Development, Defining Performance Indicators and Key Performance Indicators (KPIs), and Business Intelligence News. (n.d.). Retrieved November 16, 2018, from http://dashboardinsight.com/articles/digital-dashboards/fundamentals/the-benefits-of-keyperformance-+indicators-to-businesses.aspx

CHAPTER 12

# Business Analytics in Practice

## Table of Contents

## LEARNING OBJECTIVES

*After studying this chapter, you will be able to:*

- Describe the concept of financial and fraud analytics
- Explain the importance of HR analytics
- Discuss the marketing analytics
- Define the healthcare analytics
- State the significance of supply chain analytics
- Describe the functions of Web analytics
- Explain the functions of sports analytics
- Discuss how analytics is used by the government and NGOs

## 12.1 INTRODUCTION

In the previous chapter, you have learned about the need for a system management and the data quality. Next, it has explained about business metrics and their importance. Further, the chapter has discussed about the KPI solution and the types of KPIs. Towards the end, the chapter has discussed about the barriers and issues in implementing KPI.

Business analytics has emerged as a growth driver for most new era organisations. Gone are those occasions when managers used to settle on choices on the premise of their own guts or use large-scale financial indicators and their imaginable effect on individual organisations. Choices made without data and information have turned out to be unfortunate for many associations. With the advent of data innovation and increased data handling ability of PCs, supervisors are utilising numerous methods to anticipate the fate of business and enhance gainfulness of the venture. The application of analytics, client relationship management tools and different process improvement devices brings the benefit to the organisation. The entire business world is taking a look at huge information as an open door and source of a competitive advantage.

Business analytics has expanded consistently over the previous decade as confirmed by the constantly developing business analytics software market. It is targeting more organisations and reaches out to more number of users, from administrators and line of business supervisors to examiners and other information specialists, inside organisations.

This chapter first discusses financial and fraud analytics. Next, the chapter explains HR analytics, marketing analytics and healthcare analytics. The chapter also explains supply chain analytics and Web analytics. Towards the end, the chapter discusses stock marketing analytics and how analytics is used by the government and NGOs for providing various beneficial services to people.

## 12.2 FINANCIAL AND FRAUD ANALYTICS

Fraud impacts organisations in several ways. It might be related to financial, operational or psychological processes. While the money-related misfortune owing to fraud is huge, the full effect of fraud on an organisation can be more shocking. As fraud can be executed by any worker inside an organisation or by an external source, it is essential for an organisation to have successful fraud management or a fraud analytics programme to defend its reputation against fraud and prevent financial loss. Numerous organisations stay helpless against extortion and money related crime since they are not exploiting new abilities to battle today's dangers. These abilities depend intensely on huge information and analytic innovations that are currently accessible.

With these advancements, organisations can oversee and examine terabytes of recorded and outsider information. The capacity to break down enormous information volumes empowers organisations to make exact and precise models for perceiving and forestalling future fraud.

By utilising the most recent advancements in robust analytics, organisations can unhesitatingly ensure themselves and their clients regarding privacy and security of data while doing business with them or offering them various services which require their personal data to be utilised.

Advanced analytics can also be connected to all key fraud information to foresee whether an activity is possibly fraudulent before losses happen. Taking a look at just little arrangements of security information, for example, occasion logs, decreases a bank's capacity to anticipate or identify sophisticated crime.

Intelligent investigation of suspicious movement requires performing and managing requests that are bolstered by careful investigation and data availability. With these tools, organisations can rapidly confirm fraud and then the further activities such as prosecution and recuperation can be taken.

Organisations can use the already recorded information and analyse it to detect and prevent frauds in future. This information also helps in detecting the past and future impressions of the fraud. The recorded information related to fraud can help organisations to prevent huge losses of money and data related to it or clients.

Data management software empowers auditors and fraud analysts to break down an organisation's business information to gain knowledge into how well internal controls are working and distinguish transactions that appear to be fraudulent. Generally, data analysis can be done at places in an organisation where electronic transactions are recorded and stored.

The companies also use whistleblower hotlines which help individuals for reporting speculated fake conduct or unsafe conduct and violations of its law and policy. However, using hotlines alone are insufficient. Why be just receptive and wait for a whistleblower to come forward at the last approach? Why not search out indicators of fraud in the information? To successfully test for fraud, every

Notes

important transaction must be analysed over all pertinent business frameworks and applications. Breaking down business exchanges at the source level provide auditors with better knowledge and a more entire view with regards to the probability of fraud happening. Analysis involves the investigation of those activities that are suspicious and help control weaknesses that could be misused by fraudsters.

**Self Assessment Questions**

1. Companies also use _____ hotlines to help individuals report speculated fake conduct or unsafe conduct and violations of its law and policy.
2. _____ can also be connected to all key fraud information to foresee whether an activity is possibly fraudulent before losses happen.
3. It is essential for an organisation to have successful fraud management or a fraud analytics programme to defend its reputation against fraud. (True/False)

**Activity**

Collect information from a nearby local bank related to the impact of fraud in the financial system and all the measures taken by the banking institution to reduce the fraud. Prepare a report on this topic.

## 12.3 HR ANALYTICS

HR analytics is a zone in the field of analysis that alludes to applying analytic processes to the human resource department of a company in the expectation of enhancing worker execution along with improving the degree of profitability. Organisations generally move to HR analytics and data led solutions when there exists problems that cannot be resolved with current management practices.

HR analytics does not simply manage the gathering of information on employee performance and efficiency; instead, they also provide deeper details of each process by accumulating data and then use it for making important decisions about improving these processes.

HR analytics establishes a relationship between business data and individual's data, which further help in building important connections between them. The main aspect of HR analytics is to show people the impact of HR department on the whole organisation. HR analytics also help in building a cause-and-effect relationship between the tasks of HR and business outcomes and then making strategies on the basis of that information.

The core functionalities of HR can be improved by applying various processes in analytics which include acquisition, optimisation, paying and creating the employees workforce for the organisation. HR analytics can also help in digging problems and challenges using analytical workflow and guide managers in answering questions. It also help managers in gaining deeper details from information at hand, then make important decisions and take proper actions.

The field of HR analytics can be further divided into the following segments:

- **Capability analytics:** It is a talent management process that enables you to identify capabilities or core competencies that you require in your business. It helps in identifying the capabilities of your workforce which includes their skill, level and expertise.
- **Competency acquisition analytics:** It refers to the process of assessment how well or otherwise your business can attain the required competencies. Acquiring and managing the talent is very critical for the growth of business.
- **Capacity analytics:** It helps in identifying how many operationally efficient people are in business. For example, it identifies whether people are spending time in profitable work or not.
- **Employee churn analytics:** Employee churn analytics refers to the process of estimating the staff turnover rates for predicting the future and reducing employee churn.
- **Corporate culture analytics:** It refers to the process in which assessment and understanding about the corporate culture or the different cultures that are followed across an organisation is done.
- **Recruitment channel analytics:** It refers to the process of finding out the source of getting or recruiting best employees and most efficient recruitment channels.
- **Employee performance analytics:** Every organisation requires employees that are capable and perform well to survive and thrive. Employee performance analytics is used in assessing the performance of an individual employee. The resulting information can be used to determine which employee is performing efficiently and which employee may require some extra support or training for improving its performance.

**Self Assessment Questions**

4. HR analytics is also known as _____ analytics.
5. HR analytics help managers in gaining deeper details from information at hand, then make important decisions and take proper actions. (True/False)
6. _____ analytics helps in identifying how many are operationally efficient people are in business.

**Activity**

Visit an organisation and meet its HR executives to know how HR analytics help them to motivate their employees and reduce employee turnover in the last five years.

## 12.4 MARKETING ANALYTICS

Every organisation strives to gain an edge over its competitors. This can be possible if an organisation develops an effective industry level strategy. For this, an organisation needs to analyse various forces, such as level of competition in the

Notes

market, entry of new organisations, availability of substitute products, etc. For this purpose, marketing analytics are used by organisations.

Marketing analytics helps in providing deeper insight into customer preferences and trends. Despite various benefits, a majority of organisations failed to realise the benefits of marketing analytics. With the advancement of search engines, paid search marketing, search engine optimisation (SEO) and efficient new software solutions, marketing analytics has become more effective and easier to get implemented than ever.

You need to follow these three steps to get the benefits from marketing analytics:

1. **Practice a balanced collection of analytic methods:** In order to get the best benefits from marketing analytics, you need an analytic evaluation that is balanced, i.e., one that merges methods for:
   - **Covering the past:** Utilising marketing analytics to research on the past. You can answer a few queries such as which campaign component was used to make most income from last quarter?
   - **Exploring the present:** Marketing analytics enables you to decide how your marketing activities are acting at this moment by asking questions such as: How are clients doing? Which channels do clients use to gain maximum benefits? What is the reaction of different networking media personnel on the company's image?
   - **Predicting influencing what is to come**: Marketing analytics can be used to deliver data-driven expectations to change the future by putting few inquiries such as: How would we be able to transform here and now win into dedication and continuous engagement? In what capacity, should we include more sales representatives to meet expectations? Which urban communities would be a good idea for us to focus next by utilising our present situation?
2. **Evaluate your analytical capabilities and fill in the gaps:** Marketing organisations have an access to a lot of analytic abilities for supporting different marketing goals. Estimating your present analytic capabilities is necessary to attain these goals. It is significant to know about your present situation along with an analytic spectrum, so that you can determine gaps and take steps to create a strategy for filling those gaps.

   Consider an example in which a marketing organisation is already gathering data from sources like the Internet and POS transactions, but is not providing importance to the unstructured information coming from social media platforms. Such unstructured sources are very useful, and the technology for transforming unstructured data into actual insights is available today that can be used by marketers. A marketing organisation can plan and allocate budget for adding these analytic capabilities that can be used to fill that particular gap.
3. **Take action as per analytical findings:** In the continuous process of testing and learning, marketing analytics allows you to enhance the performance of your marketing programme as a whole by performing the following tasks:
   - Determining deficiencies in the channel

- Doing adjustment in strategies and tactics as and when required
- Optimising processes

Due to a lack of capability to test and evaluate the performance of your marketing programmes you would not be able to know what had worked and what had not. Moreover, you would not be able to know whether things needed to be changed or in what manner. In other words, if you are using marketing analytics for evaluating success and doing nothing with the details gained, then what is the point of using analytics? Marketing analytics enables better, more successful marketing for your efforts and investments. It can lead to better management which helps in generating more revenue and greater profitability.

**Self Assessment Questions**

7. SEO stands for
   a. Search engine optimisation
   b. Searching engine optimisation
   c. Search engine operation
   d. None of these
8. Marketing analytics enables you to decide how your marketing activities are acting at this moment. (True/False)
9. The information collected after performing marketing analytics remain useful whether you act or not on that information. (True/False)

**Activity**

Prepare a report on the total sales and revenue generated by a store at your nearby location by using marketing analytics.

## 12.5 HEALTHCARE ANALYTICS

Healthcare analytics is a term used to describe the analysis of healthcare activities using the data generated and collected from different areas in healthcare such as pharmaceutical data, research and development (R&D) data, clinical data, patient behavior and sentiment data, etc. In addition to this, data also gets generated from patients buying behavior in stores, claims made by patients, preference of patients in selecting activities and more. The analytics are applied on this data to get insight of data for providing healthcare services in a better way.

Organisations in the field of healthcare are quickly receiving data frameworks to enhance both business operations and clinical care. Many classes of data frameworks develop in the human services area, extending from electronic medical records (EMRs), specialty care management, to supply chain system.

Notes

Healthcare organisations are also implementing approaches, for example, lean and six Sigma to take a more patient-driven concentration, lessen errors and waste and increase the number of flow of patients with the objective of enhancing quality. The healthcare analytics industry is a growing industry and it is estimated that it will cross $18.7 billion by 2020 alone in the United States (US). The industry also emphasises on various areas such as financial analysis, clinical analysis, fraud analysis, supply chain analysis and HR analysis.

In addition to reveal data about present and past organisational performance, analytical tools are also used to study large informational collections by using statistical analysis procedures to uncover and comprehend recorded information with an eye to foresee and enhance operational execution later on.

Healthcare analytics is used as a measurable instrument in getting deeper details of medicinal services related information keeping in mind the end goal to determine past performances (i.e., operational execution or clinical results) to enhance the quality and proficiency of clinical and business procedures and its execution in future.

As the volume and accessibility of healthcare information keeps on increasing, healthcare organisations progressively need to depend on analytics as a key competency to comprehend and enhance their operations.

Healthcare data is not easily available in a unified and informative way and therefore, restricting the industry's endeavour to enhance quality and effectiveness in healthcare. Real-time analytics tools are used in healthcare for addressing these issues by bringing data from various sources at a single location with the purpose of presenting it in a unified manner so that fruitful information can be derived from it.

Moreover, the data picked up from breaking down gigantic measures of collected health information can give noteworthy knowledge to enhance operational quality and effectiveness for providers, insurers and others. The healthcare industry is quickly transitioning from volume-to value based healthcare. Presently like never before, the analytics is crucial for clinicians and health service providers so that they can distinguish and address gaps in care, quality and hazards and use it to bolster changes in clinical and quality results and financial performance.

Real-time analytics is capable of continuous reporting that illustrates the status of the patients and how to enhance the current quality of the services.

**Self Assessment Questions**

10. EMRs stands for
    a. Electrical Medical Records
    b. Electronic Medical Records
    c. Electronic Mediclaim Records
    d. None of these
11. Healthcare analytics is based on the verification of patterns in healthcare data for determining how clinical care can be enhanced while minimising excessive cost. (True/False)
12. _____ analytics is capable of continuous reporting that illustrates where a patient stands and how to enhance the quality of services.

## 12.6 SUPPLY CHAIN ANALYTICS

Generally, a supply chain comprises suppliers, manufactures, Wholesalers, retailers and customers. Intense competition and compulsion to reduce cost have impelled organisations to maintain an effective supply chain network. Therefore, organisations came up with various tools and techniques of effectively managing a supply chain.

Globalisation gave a major push to supply chain management. Organisations that operate in a highly competitive global environment needs to have a highly effective supply chain management system in place. For example, Apple faces huge demand for their products as soon as the products are announced in the market. Most Apple products are manufactured in China; therefore, Apple needs to have a highly efficient supply chain to ship items from China to different countries in the world.

It can be clearly concluded from the above discussion that supply chain is a dynamic process in which various parties, such as suppliers and distributors, are involved in delivering products and services to fulfil customer requirements. Thus, in the absence of a supply chain, there would be disruptions in the flow of products and information.

It can be said that a supply chain plays an important role in an organisation. Thus, it is of utmost importance for an organisation to manage activities involved in a supply chain. The activities in a supply chain converts the raw material into a final product which further can be delivered to the customer.

Almost every economy is getting globalised today, and the companies are competing to increase their presence in the global market. The operations performed by global manufacturing and logistic teams are getting more intrinsic and challenging. Delay in shipments, ineffective planning and inconsistent supplies can lead to an increase in the supply chain cost of the company. Some issues faced by supply chain organisations are as follows:

- Visibility of global supply chain and various processes in logistics
- Management of demand volatility
- Fluctuations of cost in a supply chain

To overcome such challenges in the supply chain, the supply chain analytics is used by most organisations or supply chain executives. Organisations are planning to increase their investment in analytics to perform better in the market in comparison to their competitors. With improvement in the supply chain analytics in the past few years, it helps in making decisions for critical tactical and strategic supply chain activities.

Various solutions are provided by the supply chain analytics to supply chain organisations are as follows:

- **Use of smarter logistics:** The use of smarter logistics helps supply chain organisations in providing more visibility in the global market. With the growth of businesses, opportunities are developed Worldwide to lure customers by satisfying their needs related to the product irrespective of the geographical

Notes

location. As customers are present Worldwide, a complex Web of supply chain has been created that must be monitored closely to remain competitive in business.

The use of advanced analytics-driven 'control metrics' allows the monitoring of real-time critical events and key performance indicators (KPIs) with the help of various touch points. The intergation of these metrics with predictive analytics provide a high amount of savings in areas such as freight optimisation. Organisations that are making investments in supply chain visibility can take decisions to increase supply chain responsiveness, optimise cost and minimise customer impact.

- **Managing customer's demand through inventory management:** Due to globalisation and variations in products to fulfill the requirements of globally available customers, demand volatility has enhanced to a significant level. Industries in various sectors such as retail, consumer goods, automotive need daily or real time prediction to perform better in the market. Advanced supply chain analytics can be implemented to these sectors or related industries more precisely to forecast demand and describe and monitor policies related to supply and replenishment. It is also used for planning inventory flow of goods and services.
- **Reducing cost by optimising sourcing and logistic activities:** The cost involved in supply chain is a major portion of company's overall cost. The supply chain costs significantly impact various financial metrics such as the cost of goods sold, working capital and cash flow. There is a constant requirement to improve organisations' financial performance which can manage huge amounts of inventories. The main areas where costs can be handled by using analytics-driven intelligence include materials, logistics and sourcing. Analytical tools help in providing better visibility to the actual total component cost of products. It is necessary to make decision regarding the buying and selling of products. With the availability of complete information at the fingertips, organisations can decline the material purchases through improved practices in supply chain and better price negotiation. The fluctuation in patterns of demand of customers and an increased base of suppliers and logistics partners make organisations redesign their logistics network planning. With the growth of business, suppliers also increase, the companies can use these suppliers against each other by applying analytics to get the lowest price from them. Supply chain managers can use sophisticated analytics programmes which can provide them real-time supplier performance management data to improve their strategies.

**Self Assessment Questions**

13. Intense competition and the compulsion to reduce cost have impelled organisations to maintain an effective supply chain network. (True/False)
14. _____ supply chain analytics can be implemented to various sectors such as retail, consumer goods, automotive or related industries more precisely to forecast demand and describe and monitor policies related to supply and replenishment.

**Activity**

Prepare a report on how the use of business analytics tools in supply chain has helped in improving the production of the manufacturing industry.

## 12.7 WEB ANALYTICS

Web analytics refers to measuring, collecting, analysing and reporting of Web data to understand and optimise the usage of Web. Web analytics also help companies in measuring the outcomes of traditional print or broadcast advertising campaigns, in estimating how traffic to a website alters after launching of a new campaign of advertising, in providing accurate figures of visitors on a website and page views, and in gauging Web traffic and popularity patterns which are useful in market research. The four basic steps of Web analytics are as follows:

- **Collection of information:** This stage involves gathering of basic or elementary data. This data involves counting of things.
- **Processing of data into information:** The purpose of this stage is to process the collected data and derive information from it.
- **Developing KPI:** This stage focuses on using the derived information with business methodologies, referred to as KPIs.
- **Formulating online strategy:** This stage emphasises on setting online goals, objectives and standards for the organisation or business. It also lays emphasis on making and saving money and increasing marketshare.

There are two categories of Web analytics: off-site Web analytics and on-site Web analytics. Off-site Web analytics allows Web measurement and analysis irrespective of whether you own or maintain a website. It includes the measurement of a website's potential audience, visibility and comments that are going on the Internet. On the other hand, On-site Web analytics is used to measure the behaviour of a visitor who had once visited the website. The On-site Web analytics is used to measure the effectiveness and performance of your website in a commercial context.

This data generated is further compared against KPIs for performance and is used for improvement of a website. Google Analytics and Adobe Analytics are popular on-site Web analytics services.

There are mainly two methods of gathering data technically. The first method lays emphasis on server log file analysis in which the log files are read and used by the Web server for recording file requests sent by browsers. The second method, known as page tagging, uses JavaScript embedded in the Web page for tracking it. Both the methods can gather data which can be processed for generating reports of Web traffic. This second method provides more accurate result as compared to the first method.

Web analytics is helpful to any business for deciding the division of market, determining target market, analysing market trends and deciding the conduct of site

Notes

visitors. It is additionally helpful to comprehend visitor's advantages and priorities. Some important uses of Web analytics for business growth are as follows:

- **Measure Web traffic:** Web analytics can track the number of users visiting the site and identify the source from where they are coming. It also focuses on the keywords that the visitors utilise to query items on the website. It also demonstrates the quantity of visitors on the Web page by means of the diverse sources like Web search tools, through messages, online networking and promotions.
- **Estimate visitors count:** Frequent or large number of visits from visitors shows the activity the site is getting. The Web analytics tool helps in deciding how frequently a visitor came back to a site and which pages of a site were given more preference by visitors. It additionally tells various traits about visitors such as their nation, language, etc. Web analytics also provide report about the time that was spent by a particular visitor on the website or total time by visitors as a whole. Such reports help to enhance pages and reduce their bounce rate (or low engagement). It additionally demonstrates high engagement time of pages and tells in which item or service visitor may be interested.
- **Track bounce rate:** A bounce describes a situation in which a visitor visits a page on the site and leaves that page without making any move or clicking on any links on that page. A high bounce rate could mean visitors were unable to find what they were searching for in the site.
- **Identify exit pages:** A few pages on a site may have a high leave rate, similar to the thank you page on an online e-commerce website after purchasing is done successfully. A high exit rate on a particular page demonstrates that the page has some issue and should be investigated quickly. Examination of such pages should be done to determine whether visitors are not getting the intended information for which they have visited the website. Web analytics tools help in finding such pages quickly and rectifying the problems with those pages.
- **Identify target market:** It is essential for advertisers to understand their visitors and deliver information according to their requirements. The discoveries of analytics services uncover the present market requests which generally change with a geographic area. By utilising Web analytics, marketers can track the volume and geographical information of visitors and can offer things according to the interest of visitors.

**Self Assessment Questions**

15. _____ analytics helps in gauging Web traffic and popularity patterns which are useful in market research.
16. There are two categories of Web analytics, which are _____ Web analytics and _____ Web analytics.

## 12.8 STOCK MARKET ANALYTICS

Stock market analytics involves the use of various tools, techniques, and data analysis methods to understand and interpret the movements and trends in financial markets. It plays a crucial role in helping businesses make informed decisions related to

investments, risk management, and overall financial strategy. The following points explain the key aspects of stock market analytics:

- **Data collection and cleaning:** Stock market analytics begins with the collection of relevant financial data. This includes historical stock prices, trading volumes, company financial statements, economic indicators, and other relevant information. Data cleaning and preprocessing are essential to ensure the accuracy and reliability of the data. This involves handling missing values, correcting errors, and transforming data into a usable format.
- **Descriptive analytics:** Descriptive analytics involves summarising and presenting historical data to gain insights into past market trends. This can include the calculation of key financial metrics, charts, and graphs to visually represent stock performance. Descriptive analytics helps in understanding the overall market conditions, identifying patterns, and assessing historical performance.
- **Predictive analytics:** Predictive analytics uses statistical models and machine learning algorithms to forecast future stock prices and market trends. Time series analysis, regression models, and machine learning techniques are commonly employed in this phase. Predictive analytics helps businesses anticipate potential market movements, enabling them to make proactive decisions in response to changing market conditions.
- **Risk management:** Analysing stock market data is crucial for assessing and managing investment risks. Businesses use various risk metrics and models to understand the potential downside of their investment portfolios.
- **Sentiment analysis:** Sentiment analysis involves assessing market sentiment based on news articles, social media, and other textual data sources. Natural Language Processing (NLP) and machine learning algorithms help in extracting insights from unstructured data.
- **Portfolio optimisation:** Business analytics in the stock market also includes optimising investment portfolios. This involves selecting the right mix of assets to achieve a balance between risk and return. Modern portfolio theory and optimisation algorithms help businesses construct portfolios that aim to maximise returns given a level of risk or minimise risk given a target level of return.

**Self Assessment Questions**

17. Sentiment analysis in stock market analytics involves assessing market sentiment based on numerical data sources such as stock prices and financial ratios. (True/False)
18. Which phase of stock market analytics involves summarising historical data to gain insights into past market trends?
    a. Predictive analytics
    b. Descriptive analytics
    c. Prescriptive analytics
    d. Quantitative analytics

## 12.9 ANALYTICS FOR GOVERNMENT AND NGOS

Data analytics is also playing its role in the government sector. Not only it is important for government, it is also equally beneficial for non-governmental organisations (NGOS). Data analytics is used by these organisations to get deeper details of data. These details are used by the organisations for modernising their services, progress and determining the solutions faster.

Big Data analytics is used in almost every part of the world for deriving useful information from huge sets of data. Not only private organisations and industries are employing data analytics but also many government enterprises are adopting data analytics for taking smart decisions for the benefit of its citizens. Lot of data gets generated in the government sector and processing and analysing this data helps the government in improving its policies and services for citizens. Some benefits of data analytics in government sector are as follows:

- With the rise of national threats and criminal activities these days, it is important for any government to ensure safety and security of its citizens. With the help of data analytics, intelligence organisations can detect crime prone areas and be prepared to prevent or stop any kind of criminal activity.
- The analytics also help in detecting the possibility of the cyber attacks and identifying criminals. It also helps in detecting their patterns of attacks. The government can therefore, takes appropriate action in advance to prevent people from any kind of financial loss.
- Government can use analytics to track and monitor health of its citizens. It can also be used for tracking disease patterns. The government can launch proper healthcare facilities in advance in the areas prone to diseases. It also helps in arranging and managing free medicines, vaccinations, etc., in order to save life of people.
- Real-time analysis and sensors help government departments in water management in the city. The officials can detect the issues in the flow of water, pollution level in water, predict scarcity of water on the basis of usage, detect areas of leakage, etc. Government departments can take proper action to avoid these issues to ensure supply of clean water in city.
- Government organisations also use analytics to detect tax frauds and predict the revenue. Government can take necessary steps to prevent tax frauds and increase the revenue.

You can say that data analytics is helping government in building smart cities having the capability of fast detection and rectification of problems. For example, in India, the government led by Prime Minister Narendra Modi has been encouraging people to adopt Digital India initiative. This will lead to ease in collection and quicker availability of data for analytics to detect flaws in money transactions and prevent people from becoming the victim of fake currency.

Data analytics also helps NGOs in improving their services to needy or poor people. Mainly, NGOs help people in several ways such as by providing free education,

books, medicines, clothes, etc. NGOs use data analytics to become more efficient while raising and allocation of funds, predicting trends and planning campaigns, identifying prospective donors and encouraging donors who have made contributions earlier, etc.

Consider the case of a non-profit organisation, Akshaya Patra foundation, which supplies food in government schools in Bangalore. The foundation was finding it difficult to supply food to government schools due to high cost involved with it. Therefore, they looked for a cost-effective solution to deliver food in schools without any interruption.

According to Chanchalapathi Dasa, Vice-chairman of Akshaya Patra, the foundation use 34 routes for delivering food to government schools in Bangalore and expenditure on each route is ₹60,000 per month approximately.

Therefore, the organisation has decided to use the data analytics to find a cost effective solution to this problem.

While analysing various parameters required in food delivery such as the number of vehicles utilised, the time and fuel used on each route, they analysed that ₹ 3 lakh can be saved by reducing the number of routes by five.

Besides Akashya Patra, several other large NGOs such as Bill and Melinda Gates Foundation India, Save the Children India and Child Rights and You (CRY) are also utilising data to raise their efficiency in getting and allocating funds, predicting trends and planning campaigns.

These NGOs often face difficulties with data collection because they use traditional ways of data collection. In order to overcome these challenges, NGOs have adopted mobile phones equipped with apps so that real-time collection and recording of data can take place. The data recorded in this manner would be accurate and will give more precise information on the basis of which further decisions or action plans can be made.

**SELF ASSESSMENT QUESTIONS**

19. NGO stands for:
    a. Non-governmental organisation
    b. Non-governer organisation
    c. Non-governing organisation
    d. None of these
20. Analytics is helpful for government in building smart cities. (True/False)

## 12.10 SUMMARY

- Business analytics has expanded consistently over the previous decade as confirmed by the constantly developing business analytics software market.
- Fraud impacts organisations in several ways which might be related to financial, operational and psychological processes.

Notes

- Numerous organisations stay helpless against extortion and money related crime since they are not exploiting new abilities to battle today's dangers.
- Organisations generally move to HR analytics and data led solutions when there exists problems that cannot be resolved with the current management practices.
- Marketing analytics helps in providing deeper insights of customer preferences and trends. Despite various benefits, a majority of organisations failed to realise the benefits of marketing analytics.
- Healthcare organisations are also implementing approaches, for example lean and Six Sigma to take a more patient-driven concentration, lessen errors and waste, and increase the number of flow of patients with the objective of enhancing quality.
- Organisations that operate in a highly competitive global environment need to have a highly effective supply chain management system in place.
- The use of smarter logistics helps supply chain organisations in providing more visibility in the global market.
- Web analytics can provide accurate figures of visitors on a website and page views. It helps in gauging Web traffic and popularity patterns.
- Stock market analytics involves the use of various tools, techniques, and data analysis methods to understand and interpret the movements and trends in financial markets.

## 12.11 KEY WORDS

- **Capacity analytics:** It helps in tracking the number of people who are operationally efficient and currently in business
- **Employee churn analytics:** It refers to the process of estimating your staff turnover rates for predicting the future and reducing employee churn
- **Employee performance analytics:** It is used in assessing the performance of an individual employee
- **Fraud analytics:** It is used to detect whether a financial activity is fraudulent or not to prevent any kind of financial loss
- **Marketing analytics:** It helps in providing deep insight of customer preferences and trends

## 12.12 CASE STUDY : US POSTAL SERVICE OPTIMISES ITS FRAUD, WASTE AND ABUSE DETECTION

US Postal Service (USPS) is a federal agency of USA. The USPS Office of Inspector General (OIG) maintains the integrity and accountability of the US Postal Service. The work of USPS OIG relates to detection of fraud, waste, and abuse.

In 2016, the USPS delivered mails to 146 million delivery points from its 37,000 postal facilities. It was estimated that the Postal Service also managed about USD 33 billion of postal contracts in 2016.

During this period, the USPS's OIG received some serious information, according to which there were instances of fraud, waste and abuse. There were also some cases of questionable postal contracts. As a matter of process, the USPS OIG initiated investigations in cases where they received any information/tip. The USPS OIG felt that they needed a Business Intelligence system to help analysts in getting data-driven leads related to fraud, waste and abuse. They wanted to be proactive in their approach so that the questionable postal contracts could be prioritised. The major reasons which led USPS to look for a BI solution included:

- Previous processes to validate the tips were time intensive
- Incomplete and inconsistent data
- Scattered financial data
- Non-availability of solicitation

In order to develop the required BI solution, USPS OIG partnered with Elder Research to develop a customised solution which could generate leads on the basis of risk indicators and may also help in detecting anomalies. It was difficult to develop this software as the USPS office could cite only a few cases of fraud that served as a guide to the analytics process.

Elder Research used approximately 30 custom fraud indicators to develop a predictive model that could detect suspicious postal contracts. The model ranked and scored the contracts on the basis of a weighted combination of the risk indicators and produced an output accordingly. This model displayed results in a visualisation tool called Risk Assessment Data Repository (RADR). Using RADR, the analysts could easily determine the high-risk contracts.

RADR is a browser-based model. While developing the model, the developers investigated each metric's usefulness and contribution to the overall model. Using RADR, the developers could drill-down into the data to determine the drivers behind each risk score. Using the risk score, the analysts could develop actionable cases. A thorough analysis of the results of the RADR model revealed that about 74% of the leads generated by the system were actionable. It was also proved that out of 31 contracts, 23 contracts involved fraud, waste and abuse.

As regards the RADR model, Director of the USPS stated, "Using RADR, we are able to assign risk scores to whatever we are measuring. We are able to model every single contract or every single transaction, whatever is being investigated. In the past, you would have to do statistical sampling or you may have to wait until someone calls to look for something. It puts a lot of information in front of the investigator."

**Source:** *https://www.elderresearch.com/hubfs/Elder_Research_Analytics_Case_Study_Fraud_Detection_Reducing_Fraud_Waste_and_Abuse_USPS_OIG.pdf?t=1472148785724*

## QUESTIONS

1. What is the role of USPS Office of Inspector General (OIG)?

   (**Hint:** The USPS Office of Inspector General (OIG) maintains the integrity and accountability of the US Postal Service.)

Notes

2. What were the challenges due to which USPS wanted to develop its BI solution?

   (**Hint:** The previous processes to validate the fraud, waste or abuse tips were time-intensive; data was incomplete and inconsistent.)

3. Why USPS OIG felt the need of a Business Intelligence system?

   (**Hint:** The USPS OIG felt that they needed a Business Intelligence system to help analysts in getting data-driven leads related to fraud, waste and abuse.)

4. Why USPS OIG partnered with Elder Research?

   (**Hint:** USPS OIG partnered with Elder Research to develop a customised solution which could generate leads on the basis of risk indicators and may also help in detecting anomalies.)

5. Comment on the utility of the Risk Assessment Data Repository (RADR).

   (**Hint:** Elder Research used approximately 30 custom fraud indicators to develop a predictive model that could detect suspicious postal contracts. The model ranked and scored the contracts on the basis of a weighted combination of the risk indicators.)

## 12.13 EXERCISE

1. Discuss the importance of financial and fraud analytics for an organisation.
2. Describe the role of HR analytics in an organisation.
3. What do you understand by marketing analytics? Discuss the steps in getting the best assistance from marketing analytics.
4. How is healthcare analytics useful in the medical field? Explain with suitable examples.
5. Write a short note on the following concepts:
   a. Supply chain analytics
   b. Web analytics

## 12.14 ANSWERS FOR SELF ASSESSMENT QUESTIONS

| Topic | Q. No. | Answer |
|---|---|---|
| Financial and Fraud Analytics | 1. | whistleblower |
| | 2. | Advanced analytics |
| | 3. | True |
| HR Analytics | 4. | Talent |
| | 5. | True |
| | 6. | Capacity |
| Marketing Analytics | 7. | a. Search engine optimisation |
| | 8. | True |

| Topic | Q. No. | Answer |
|---|---|---|
| | 9. | False |
| Healthcare Analytics | 10. | b. Electronic Medical Records |
| | 11. | True |
| | 12. | Real-time |
| Supply Chain Analytics | 13. | True |
| | 14. | Advanced |
| Web Analytics | 15. | Web |
| | 16. | Off-site, On-site |
| Stock Market Analytics | 17. | False |
| | 18. | b. Descriptive analytics |
| Analytics for Government and NGOs | 19. | a. Non-governmental organisation |
| | 20. | True |

## 12.15 SUGGESTED BOOKS AND E-REFERENCES

### SUGGESTED BOOKS


- Yang, H., & Lee, E. K. (2016). Healthcare analytics: from data to knowledge to healthcare improvement. Hoboken, NJ: John Wiley & Sons, Inc.
- Marketing Analytics: Data-Driven Techniques with Microsoft Excel. (n.d.). Retrieved May 03, 2017, from http://www.wiley.com/WileyCDA/WileyTitle/productCd-111837343X.html


### E-REFERENCES


- Data analysis techniques for fraud detection. (2017, April 26). Retrieved May 03, 2017, from https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection
- HR Analytics. (2017, March 17). Retrieved May 03, 2017, from https://www-01.ibm.com/software/analytics/solutions/operational-analytics/hr-analytics/
- Health care analytics. (2017, March 26). Retrieved May 03, 2017, from https://en.wikipedia.org/wiki/Health_care_analytics