# SMACS (Social, Mobile, Analytics, Cloud, and Security) Technologies for Business

**Block**

# 3

## BUSINESS ANALYTICS

**Editorial Team**

Prof. R. Prasad
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sindhuja
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sanjay Fuloria
IFHE (Deemed-to-be-University), Hyderabad

Dr. Nasina Jigeesh
IFHE (Deemed-to-be-University), Hyderabad

**Content Development Team**

Dr. Y. V. Subrahmanyam
IFHE (Deemed-to-be-University), Hyderabad

Prof. G. S. B. Subramanya Choudary
IFHE(Deemed-to-be-University), Hyderabad

Prof. V. Srinivasa Murthy
IFHE (Deemed-to-be-University), Hyderabad

Prof. Venkata Dharma Kumar
IFHE(Deemed-to-be-University), Hyderabad

Prof. R. Muthukumar
IFHE(Deemed-to-be-University), Hyderabad

**Proofreading, Language Editing and Layout Team**

Ms. M. Manorama
IFHE (Deemed-to-be-University), Hyderabad

Mr. K. Venkateswarlu
IFHE(Deemed-to-be-University), Hyderabad

Ms. C. Sridevi
IFHE (Deemed-to-be-University), Hyderabad



**Ref. No. SMACS-IFHE – 092022B3**

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

*Our E-mail id: cwfeedback@icfaiuniversity.in*

**Centre for Distance and Online Education (CDOE)**
**The ICFAI Foundation for Higher Education**
(Deemed-to-be-University Under Section 3 of UGC Act, 1956)
Donthanapally, Shankarapalli Road, Hyderabad- 501203

# BLOCK 3: BUSINESS ANALYTICS

Block 3 - *Business Analytics* discusses the use of available big data generated in various business transactions. There is a need to handle unstructured data as part of the business. Analyzing big data, known as data analytics, can help the top management in decision-making areas. The best advantage of big data analytics is using this for necessary market intelligence application areas like research, extrapolation, innovation needs, etc.

As organizations are able to collect voluminous data, business analytics approaches help the business executives to analyse the data for various operational, decision making and predictive business uses. Hence they need to learn these concepts.

There are four units in this block.

*Unit 9:* Business executives need to be playing a role in crucial decision making activities using evident data, and hence need to learn about use of big data, cloud and analytics. *Decision Making using Big Data* begins by explaining big data, analyzing this voluminous data through the latest technologies and tools – 5G, discusses the challenges, storage considerations and analyzing and managing big data. It also discusses the evidence-based decision-making and the role of cloud computing in big data management. A case study is added to let the reader get a complete practical view of the concepts detailed.

*Unit 10:* The present day data from various sources has heterogeneous unstructured data formats, and the business executives should learn about the storage, use, and analysis of this data. *Handling Unstructured Data* focuses on formats, web pages, video, audio, images and office software. It discusses issues in the storage of unstructured data; usage patterns; ways of analyzing and managing unstructured data; and opportunities and problems with unstructured data.

*Unit 11:* Established mathematical and structured technical methods need to be understood by the business executive for scientific data analysis. *Data Analytics for Top Management Decision Making* is a quantitative databased unit, which discusses business intelligence, business analytics, correlation, regression, multiple linear regression, factor analysis, exploratory factor analysis, principal factor analysis, confirmatory factor analysis, classification, Recency Frequency Monetary (RFM) analysis and market basket analysis.

*Unit 12:* Business executives should get confident on linking analysed business intelligence to the day to day decision making and thus need to learn these approaches. *Business and Marketing Intelligence Using Analytics* begins by defining the business intelligence need, components, architecture and methodologies. It discusses data warehousing and data mining techniques. It relates market intelligence and decision-making. After discussing these, it explains Google BigQuery, Apache Spark, and Google Dataflow.

# Unit 9

# Decision Making Using Big Data

## Structure



*"Information is the oil of the 21st century, and analytics is the combustion engine."*

- Peter Sondergaard, Senior Vice President, Gartner

## 9.1 Introduction

"Oil is valuable only when changed into gas, plastic, chemicals, etc. to create a valuable entity. So must data. It needs to be analysed for it to have value."

In the previous unit, we discussed mobile Business Process Management (BPM) covering business process management through SMACS.

Big data has become a critical challenge of the present time. It is created by the accumulation of years of data sets, as well as generation of voluminous data because of e-commerce, retail growth etc. It is normally defined in terms of volume, velocity and variety. Therefore, handling big data needs special technology like Hadoop, SAP HANA, etc.

In business, the related data is generated from customers, competitors and the companies themselves. It is either transactional or decision-making-based and the nature of data is both quantitative and qualitative. At present, it is a data-driven society because technological evolution has made "data capturing and its analysis" easy and convenient. Any volume of data can be captured and warehoused; it is up to the decision makers either to use data or to lose it.

The data available is structured, semi-structured and unstructured and is available in many forms such as numeric text tables, graphs, etc.

In this unit, complete working methods with big data, moving from operational dashboards to evidence-based decision making and the role of cloud computing in big data management are discussed.

## 9.2 Objectives

After going through this unit, you will be able to:

- Explain the meaning of big data and its elements.
- Describe the challenges of big data implementation.
- Explain analyzing and managing big data.
- Discuss evidence-based decision making using big data.
- Define the role of cloud computing in big data management.

## 9.3 What is Big Data?

Big data is a concept which defines the large volume of data which can be both structured and unstructured that floods in a business on a day-to-day basis. How huge is the data is not important but the most important thing is what the organizations do with the data. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

The data can be taken from any source and analyze it to get the solutions that help in 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making and 5) Increased customer satisfaction. The business related goals can be achieved when big data is combined with high-powered analytics which can be as follows:

- Determining root causes of failures, issues and defects in near-real time.
- Generating coupons at the point of sale based on the customer's buying habits.
- Recalculating entire risk portfolios in minutes.
- Detecting fraudulent behavior before it affects your organization.

How "Big" is the Big Data?

Business houses and science-based institutions have been using data for a very long time. They have been using very small to huge databases, but the present day technologies like cloud storage and cloud computing facilitate storage of huge volumes of data. The data dealt is in petabytes, exabytes, zetta and yottabytes. The data in petabytes and exabytes is usually referred to as big data.

Doug Laney of Gartner group defined big data as "Big data is high volume, high velocity and/or high variety information that requires new forms of processing to enable enhanced decision making, insight discovery and process optimization".

Velocity in the above definition means the speed at which the data is handled.

Internet of Things (IoT) and big data are closely intertwined and although they are not the same thing, it is very hard to talk about one without the other. The Internet of Things (or Industrial Internet) operates at machine-scale, by dealing with machine-to-machine generated data. This machine-generated data creates discrete observations (e.g., temperature, vibration, pressure, humidity) at very high signal rates (1,000s of messages/sec). Added to this, the complexity that the sensor data values rarely changes (e.g., temperature operates within an acceptably small range). However, when the values do change the ramifications, the changes will likely be important.

Consequently, to support real-time edge analytics, we need to provide detailed data that can flag observations of concern. We also need to ensure that it will not overwhelm the ability to get meaningful data back to the core (Data Lake), which is used for more broad-based, strategic analysis.

**Charateristics of Big data**

The characteristics of big data are as follows:

Volume – The name Big Data itself states that the size of the data is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data.

Velocity – The term 'velocity' refers to the speed at which the data is generated. It determines the speed at which the data is generated and processed to meet the demands and regulates actual potential in the data.

Velocity in Big Data deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is enormous and continuous.

Variety – The next feature of Big Data is its variety. Variety denotes to assorted sources and the nature of data, which can be both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc.which are considered for analysis. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

Variability – This refers to the inconsistency which is shown by the data at times, which hampers the process of being able to handling and managing the data effectively.

Veracity – Veracity refers to the quality of data. Since the data comes from various sources it’s difficult to link, match, cleanse and transform data across systems. Businesses need to connect and correlate relationships, hierarchies and multiple data linkages. Otherwise, their data can quickly spiral out of control.

Complexity – Since the data is in volumes and is coming from different sources, managing data is challenging job. Inspite of its complexity, it is necessary for the businesses to know when something is trending in social media, and how to manage daily, seasonal and event-triggered peak data loads.

In fact, the data sets are so big and complex that it becomes very difficult and challenging to process them using the traditional data processing applications. It is estimated that about 2.5 quintillion bytes of data is created every day.

This implies that about 90% of the world's total data was created in the last two years. It should also be obseverved that about 80% of the total data is unstructured data which is collected from sensors used to gather weather information, social media posts, digital photos and videos, purchase transaction records, to mobile phone's GPS and many more.

Both government and private sectors are using Big Data to increase their productivity. The United States Federal government owns six of the ten most powerful supercomputers of the world.

In the private sector, Facebook uses Big Data to handle 50 billion photos from its user's base. Amazon.com used Linux based technology to handle millions of back end operations every day. eBay.com uses two data warehouses of 7.5 PB and 40 PB as well as a 40 PB Hadoop cluster for search. FICO Falcon Credit Card Fraud Detection system secures 2.1 billion active accounts across the globe.

Walmart handles 1 million+ customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes of data. According to estimates, the volume of data worldwide doubles every 1.2 years.

The growth of Big Data databases has empowered enterprises to know the importance of data in their growth and success. These databases have helped enterprises to save money, increase revenue and achieve many other business objectives. The real challenge faced by the enterprises is finding that critical piece of information that offers the competitive edge. Hadoop helps in managing and handling massive amount of data. It also helps in transforming the data into a more usable structure and format, and extract valuable analytics from it.

**Brisk Insights into Handling Big Data**

Prominent among the technologies that handle big data is Hadoop.

Hadoop is a distributed file system and mapReduce processing technology that stores and processes data by dividing workloads across several thousands of servers.

Precisely, it is a unique "divide and conquer" approach.

Another approach is SAP – HANA (High-Performance Analytic Appliance) which is a very powerful server with terabytes of memory and which is used to compress the data as one unit. It uses a brute force method. The dataset is compressed in memory and analytics is performed on it.

Further, to bring meaning out of big data, one needs to analyze this voluminous data through the latest technologies and tools.

Consider, 5th generation mobile network which is popularly used in the massive Internet of Things (IoT) in Industry.

**What is 5G?**[1]

5G is the 5th generation mobile network. It is designed to connect people, machines, and devices and operate in a virtual environment. 5G Technology is developed to deliver peak data speeds of multiple-Gbps *(Giga bits per second)*. It has very low latency *[or, network delay - network delay is the overall amount of time that it takes for information transmission from the source to the destination in a data network]*. 5G network offers more reliability with massive network capacity and increased availability. Moreover, 5G network provides a uniform rich user experience.

**Why 5G?**

5G is estimated to yield $13.1 trillion worth of goods and services by various types of industries across the globe by the year 2035. This impact is much greater than previous network generations like 4G and 3G. 5G networks are also rapidly expanding to industries such as the automotive industry when compared to the traditional mobile networks such as 2G, 3G and 4G. 5G Network value chain includes several stakeholders such as Original Equipment Manufacturers (OEMs), Telecom operators, content creators, application developers etc. that could alone create more than 228 lac jobs. Presently, 5G is popularly used in increased mobile broadband, mission-critical communications in defence, and the massive Internet of Things (IoT) in Industry. 5G Networks will deliver cutting-edge and creative user experiences. The experiences include new enterprise applications, boundless extreme reality (XR), seamless Internet of Things (IoT) capabilities. Users can easily access local interactive content and content available from Cloud instantly from SmartPhone or any other device connected to internet. Modern factories could use 5G to run industrial Ethernet for increasing operational productivity and precision manufacturing. In the same way, Smart cities could use 5G in multiple applications to provide better amenities for the people. The 5G applications are meant primarily providing greater efficiencies like increased connectivity between people and things. Using 5G, information exchange with higher data speeds, and ultra low latency is possible than ever before in amenities areas like vehicle safety, water supply, waste disposal management, virtual reality arcades and interactive gaming. Many new applications are still emerging that will be defined in the near future.

[1] https://www.coai.com/5g_india_forum
https://www.smartcitiescouncil.com/article/bhopal-be-first-5g-enabled-smart-city-india
https://prsindia.org/policy/report-summaries/india-s-preparedness-for-5g
https://www.qualcomm.com/5g/what-is-5g
https://www.business-standard.com/article/economy-policy/india-to-account-for-15-of-global-market
https://indianexpress.com/article/technology/tech-news-technology/airtel-huawei-conduct-indias-fi

## India - 5G Availability

Global Telecom operators have launched 5G networks in early 2019. 5G has already been deployed in more than 60 countries and many more countries are still counting. It is a fact that more than 3.5 billion consumers are already using 5G compatible smartphones powered by Snapdragon. These phones are already built to deliver information or data delivery with high speeds and low latencies. In India, network operators like Airtel, Vodafone Idea, Reliance Jio, BSNL etc., have already partnered with vendors like Ericsson, Huawei, Ciena and Samsung for planned trials and commercial launch of services by the end of the year 2022. For instance, Huawei and Bharti Airtel have conducted India's first 5G network trial successfully and found that they could achieve a user throughput of more than 3 Gbps. Samsung is one of Reliance Jio's partners for its 5G field trials. It is expected that India will account for about 15 percent of worldwide market with an assumption that the telcos will be able to provide 5G coverage in metros, major cities and towns in the next two years i.e., 2023 and 2024.

5G's commercial launch in India is around the corner as the spectrum auction has been successfully concluded in July 2022. Jio (₹ 88,078 crore) emerged as the largest spender in the first-ever 5G spectrum auction in the country by forking out almost twice as much as Airtel (₹ 43,084 crore) and acquiring half of all the airwaves, including the 700MHz. Airtel was much more circumspect than Jio, while Vi (Vodafone Idea) only spent ₹ 18,799 crore for 6,228 MHz of spectrum. Adani Data Networks Limited has made bids worth ₹ 212 crore to acquire 400 MHz spectrum in 26 GHz frequency band. There is a potential for telcos to charge a premium for 5G vs 4G. The 5G Services will take time to mature and rollouts would likely be granular, starting with metros and larger cities.

## India - 5G Opportunities

**Indigenous Technology:** There will be a mani-fold increase in demand for telecom network equipment to provide 5G Network connectivity. It provides an opportunity for the promotion of local manufacturing as well as development of indigenous technology due to change in network components as compared to 3G and 4G. Further, the need for software development on OTS (Off-The-Shelf) hardware required in 5G provides an immense opportunity to leverage on vast availability of software skilled resources in India.

**Fibre as a national asset**: Deployment of Fibre Cable Network is an important requirement for the rollout of 5G Network Connectivity. But in India, we have about 30% of the towers fiberised with less than 7% of households connected as of Jan 2022. It is more likely that fibre will be accorded the essential national infrastructure status to avoid delays in government permissions and minimize costs associated.

**Setting up of 5G use case labs**: 5G Network adoption will require phenomenal investment in downstream innovation when compared to investments made in previous generations of communications systems such as 4G and 3G. It is expected that new business ecosystems are likely to emerge where multiple players will meet, compete and work together. 5G networks will open enormous opportunities to start-ups and smaller ecosystem players, who will eventually benefit from the innovation capabilities. Developers provide open interfaces to network "apps" and services. Government is promoting the development of use case labs. It is proposed by the Government that the digital readiness of various sectors shall be monitored by a cross-sectoral entity like NITI Aayog.

For instance, 5G Technology is widely adapted in Smart City Projects for the development and deployment of new applications which include air quality monitoring, energy consumption, vehicle movement on roads, traffic patterns, street lighting, smart vehicle parking management, people gathering management, and police/fire emergency response.

**Rich and Uniform Consumer Experience**: 5G deployment will certainly boost adoption of advanced technologies such as the Internet of Things, industrial IoT, edge computing, and robotics across industries like agriculture, e-commerce, healthcare, education, and pharma. The convergence of Artificial Intelligence (AI) with the Internet of Things, Industry 4.0 will reduce manufacturing costs, improve quality and yields. Data Analytics, AI and Machine Learning (ML) will penetrate new market segments and thus enable better, data-driven decisions for businesses resulting in more precise, real-time, and predictive analytics. Indian micro, small and medium enterprises (MSMEs), which are focused towards driving revenue growth and are reluctant to invest in IT infrastructure solutions initially, will require a resilient journey to Cloud as 5G Network is set to become a reality soon. MSMEs require flexibility to manage workloads by upgrading existing IT infrastructure to gain edge over competition. They can offer optimal services to customers and thereby strengthen their relationships with partners in the age of digital transformation.

India’s transition into a digital economy will be led by telecom operators. 5G Networks will present a game-changing opportunity to drive the digital transformation of industries, enterprises and contribute to the socio-economic development of India. In conclusion, 5G will bound to create a powerful impact on India’s consumers and businesses by taking mobile experiences to a new level, thus introducing a gamut of applications and enhanced information exchange capabilities.

## Challenges of Big Data

Usually, the saying is that big data is for computers and small data is for people. To understand the challenges of big data one has to get into the characteristics of big data.

The definition given by John Naisbitt makes us aware of the challenges of big data.

"We have, for the first time, our economy based on a key resource (information) that is not only renewable but self-generating. Running out of it is not a problem, but drowning in it is."

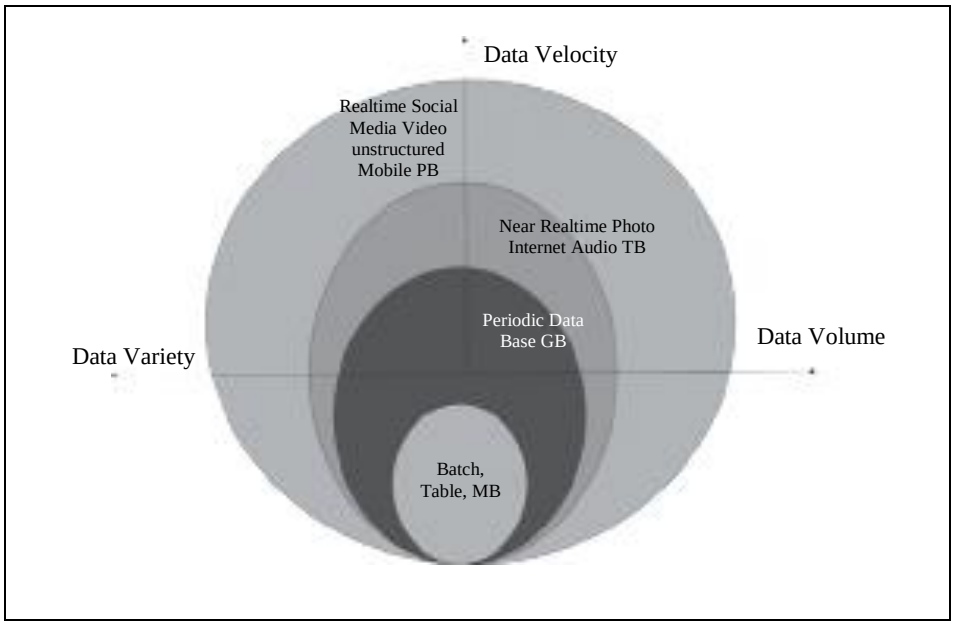
### Characteristics of Big Data

Let us spend some time on understanding the definition of big data given by Gartner business consulting.

Part of the definition is as follows (see Figure 9.1).

"Big data is high volume, high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization".

**Figure 9.1: 'Big Data' Characteristics**




*Source: ICFAI Research Center*

### Looking further deep into the definition, especially the 3 Vs:

**Volume** in the definition is the size of big data. In earlier days, data was generated by the internal employees and was in kilobytes. In the present day situation, the data is generated by employees, customers, competitors, etc., and thus, the data is in gigabytes and petabytes. Its volume is ever increasing with the audio and video contents coming into the picture along with the regular text files. The Parkinson's Law indicates that "Data expands to the extent to fill the space available for storage". Because of cloud storage and other hardware and software

technologies, storage is not an issue as of now but considering the rate at which data is getting generated, its storage, usage, and analysis will certainly become a concern in the near future.

**Velocity** in the definition is the speed at which the data is used or processed. In earlier days, the data was analyzed or processed by batches and the scheme worked efficiently when the incoming rate was slower than the process time. In the present era, data is streaming into the server in real-time in a contiguous fashion and results are useful, if they are ready in a rapid turnaround time. This requires advanced hardware and software tools and techniques, which is a challenge.

**Variety** is the form and format of the data that is available; conventionally, data is in MS-Excel tables and other databases, warehouses. The data structures are losing their form due to the availability of data in different formats. Today's mobile data includes Text which is in qualitative form, Photo, Audio, Video, Web, GPS data, Sensor data, Data from relational databases, Short Messaging Services, Pdf, Flash, and other unconventional/informal formats.

With the advent of mobile apps, the data getting generated is in a variety of formats and even the conventional *modes* and *methods* are challenged very frequently. Research is on in a big way to present and analyze the data.

Another term to add to the above definition is **"**veracity". For the analysis results to be accurate, the quality of the data needs to be purely reliable and accurate, which is a huge challenge. Another challenge is regarding managing the reliability and predictability of informal, unstructured or semi-structured data.

**Value** is a definite addition to big data activities. Many retail business managers use big data analytics for many decision-making activities. Similar is the case in scientific applications too.

**Example: Apollo Hospitals uses Big Data Analytics to Improve its Emergency Care**

Apollo Hospitals, the major hospital chain recognizes emergency care is very critical to its operations. The hospital management works continuously to improve the emergency care to its patients. The hospital used big data analytics to vast amount of historical data on its servers and found out that the healthcare givers are spending more time away from the patients by way of accessing the lab reports, doing data entry etc.This was corrected by implementing Emergency 2.0 where by the cubicles are redesigned so that they were made independent units enabling the doctors and nurses to access all relevant data within the cubicle and eliminating time to be away from critical parents.

*Source: How Apollo Hospital's digitization has uplifted its emergency treatments, CIO News, ET CIO (indiatimes.com), 14th April, 2022, Accessed on 03/08/2022.*

## 9.4 The Challenges of Big Data

Due to digitalization, there has been a sudden surge in data consumption and more and more organizations are continuing to invest in their digital infrastructure. This will raise issues of technological concerns — both software and hardware— to handle data for storage representation and analysis.

Let us discuss the challenges of big data management. Data offers huge insights and advantages for decision makers. But terabytes and petabytes of data will be pouring in day-in-day-out. Conventional methods of data analysis infrastructure proved to be of no use in bringing out the information from these data mountains. IT teams are burdened with ever-growing requests for data and ad-hoc reports. Data visualization is gaining popularity since it gives a quick view and understanding of the data.

Organizations which are handling big data are relying on statistics, choosing an appropriate sample of the population and then the sample is analyzed. The results were attributed to the population based on the behavior of the selected sample, which is not a tested (conventional) method for drawing inferences on the population.

Companies are using SaaS (software as a service) visual analytics to represent the data in a graphical and pictorial form which gives the decision makers a better understanding of the voluminous data.

The following are the challenges of big data in brief:

- **Matching the Speed of Results/Inferences**

  Analysis needs to be performed on big data with a very quick turnaround time which was not possible by conventional methods. Some companies are using hardware capabilities like huge memory and powerful parallel processing to crunch large volumes of data fairly and quickly. Another method used is to apply grid computing where several machines are used for computing; both the methods are used to analyze and present data with a quick turnaround time almost in real-time.

- **Clear and Comprehensive Understanding of the Data**

  In order to get meaning out of data, one needs to have excellent domain expertise and understanding of multiple methods of analyzing available data. People who are handling analysis of big data must have a fair understanding of the incoming data. They also need to be cognizant of the sources of the data, who the end users are, and what will they look forward to infer from the data, etc. This is, as well, a requirement for any data analysis situation, specifically when the data is large.

- **Quality of Data**

  One is aware of the fact that garbage-in will result in garbage-out. A good quality data will only ensure good analysis and proper inferences. We need to use proper statistical tools, techniques, data governance, and information management system. It is always better to have proper data quality checking mechanisms so that good quality data is always in.

- **Discrepancies in the Data**

  As part of data quality, one usually encounters missing values, outliers, etc. Missing values might occur in the collection or at the entry point, etc. Since data is huge, 1% missing value is also a big issue in data analysis. One may use the method of averages or interpolations, etc. Outliers are the numbers which fall outside the normal range if the data is numeric. If the data is qualitative, the outliers need to be treated properly. Even if 1% deviation exists in big data activities, it means quite a good volume of data, and availability of this may be critical for analysis and inferences. One may need to segregate outliers and analyze them separately.

- **Inferences about the Data**

  Representing and drawing inferences about the data is the key to decision making. Use of visual representations with the latest software and using advanced statistical tools and techniques, like multivariate data analysis techniques, help the decision maker to understand and gather actionable intelligence.

  To conclude on the challenges of big data, the three Vs, apart from veracity and other data-related issues, are best addressed by a combination of advanced software and improved hardware.

**Example: HDFC Bank Meets the Challenge of Big Data Talent Shortage by Starting a Specially Designed Training Programme**

HDFC Bank, the banking major in India has been harnessing the power of big data analytics to enhance its operational efficiency and hence customer satisfaction and customer retention. The company has realized that availability of the right talent for big data analytics in the BFSI segment is a major challenge. It is finding it difficult to recruit outside talent because there is storage and even the sourced talent do not have domain knowledge of healthcare operations. So the bank has designed a training programme which will provide the talent for itself and other companies in the BFSI segment.

*Source: https://www.hrkatha.com/news/learning-development/hdfc-bank-to-train-and-hire-100-data-scientists-this-year/, 14th July 2022, Accessed on 03/08/2022.*

## 9.5 Analyzing and Managing Big Data

Let us start this section with some definitions given by experts on data, statistics, and knowledge.

*"Data is just like crude. It's valuable, but if unrefined it cannot really be used".*

Michael Palmer (2006)

*(Source: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html)*

The above definition, an expanded version of Humby's quote (*"Data is the new oil"*), emphasizes the need for analyzing data.

*"The need for three Rs in reading, writing and arithmetic is well understood. These do not take us far unless we acquire the fourth R, reasoning under uncertainty for taking decisions in real life".*

*Prof.C.R.Rao (2014)*

*(Source:www.crraoaimscs.org/Stat_Day_2014_TJR*)

"'Big data' is a high volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

Douglas Laney, of META Group (Gartner now), 2020

*(Source: https://www.google.co.in/books/edition/Knowledge_Graphs_and_Big_Data_Processing/qWTxDwAAQBAJ?hl=en&gbpv=1&dq=definition+of+%22data%22+in+big+data+analytics&printsec=frontcover)*

This definition given by Prof. C.R.Rao reiterates the need for data analysis, particularly large volumes of data.

Walmart collects 2.5 pegabytes of unstructured data from 1 million customers every hour[2]. Decoding the human genome is becoming difficult now.

Big data is often managed by hybrid cloud technology because of the need to analyze huge volumes of data. It has three characteristics — Volume, Variety, and Velocity — as discussed earlier.

The following Table 9.1 gives an idea about terabytes and petabytes.

**Table 9.1: Big Data Numbers**

| Name | Byte | Gigabyte | Terabytes | Petabytes |
|---|---|---|---|---|
| Value | $10^0$ Bytes | $10^9$ Bytes | $10^{12}$ Bytes | $10^{15}$ Bytes |

[2] https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109, June 29th 2022, (accessed on July 11th 2022)

Big data comes in different shapes and sizes:

- **Structured data:** Most of the analysts used to deal with structured data which can be included in the database, for example, number of items sold, number of customers visited the outlet, etc.
- **Semi-structured data:** It has some structure but not in the form of tables as in a database; it includes EDI (Electronic Data Interchange) formats and XML (Extensible Markup Language) formats, etc.
- **Unstructured data:** It includes text, images, audio, documents, e-mails, tweets, blogs, etc. Unstructured data amounts to 80% of the data in use in the current time.

While analyzing big data, the following terms are heard very commonly: text analytics, sentiment analytics, voice analytics, moment analytics, and machine to machine analytics.

### 9.5.1 Analysis of Big Data

The concept of big data has been around for years. Many organizations are understanding that the significant value of the data captured into their businesses can infer many business issues, using analytics. Even in the olden days before the term "big data," was expressed, businesses were using basic analytics (essentially numbers in a spreadsheet that were manually examined) to reveal understandings and trends

The new benefits that big data analytics brings are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and extracted information that could be used for future decisions, whereas the businesses are able to identify insights for instant decisions in todays scenario. The ability to work faster and stay agile gives organizations to have competitive edge which was lacking earlier.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously unexploited data sources independently or together with existing enterprise data.

Big data is analyzed for a better understanding of the customer and in the process also identifying target customers, etc. Retailers can predict what products will have maximum sales, understand and optimize the business by synthesizing the data of customers from social media and make decisions on whether to maintain an inventory of certain goods.

One of the problems in dealing with large amounts of data is that it is not always easy to identify patterns. Thus we need data visualization software, which helps in converting streams of text and numbers into graphical representations that reveal a greater picture of the data.

Association rule mining is a methodology that is used to discover unknown relationships hidden in big data. Rules refer to a set of identified frequent item sets that represent the uncovered relationships in the dataset. The underlying idea is to identify rules that will predict the occurrence of one or more items based on the occurrences of other items in the dataset. It is an unsupervised machine learning method. This means that no direct guiding output data is given to find the patterns.

### 9.5.2 Management of Big Data

In many commercial environments, large quantities of data is accumulated in databases from day-to-day operations. This lays the foundation for mining association rules. In retail, for example, "customer purchase data" is collected on a daily basis at the checkout counters of city stores or when shopping at online stores. The accumulated data items are often market basket transactions. Managers of stores are interested in analyzing the collected data in order to learn the purchasing behavior of customers. This enables a large variety of business-related applications based on the identified rules in the data.

Sequential pattern mining is a data mining task specialized for analyzing sequential data, to discover sequential patterns. More precisely, it consists of discovering interesting subsequences in a set of sequences, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has numerous real-life applications due to the fact that data is naturally encoded as sequences of symbols in many fields such as bioinformatics, e-learning, market basket analysis, texts, and webpage click-stream analysis.

**Big data use**

When probed with powerful visualization tools, big data can produce a wide variety of insights allowing to test relationships between multiple variables, detect anomalies, compute the statistics from basics to advanced level for prediction, etc.

So far we have discussed big data, issues of big data analytics, apart from the data analysis tools and techniques, from the basic to an advanced level for a better understanding of the business.

Big data analytics makes use of advanced tools like data mining, statistics, predictive analytics, and natural language processing and they are used against large volumes of data. It allows analysts, researchers and business users to make better and faster decisions.

### 9.5.3 Insights into Tools that are used for Analysis of Big Data

**Hadoop** is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop – native data processing and analysis options include:

- Apache Hive (provides data accessing and data warehousing similar to Structured Query Language -SQL),
- Apache Mahout (supports machine learning on top of Hadoop– for finding patterns in data),
- Apache MapReduce (for searching, filtering and sorting–ways to generate useful nuggets from big data), and
- Apache Pig (a language for writing MapReduce jobs).

**Alternative SQL access/analysis options:** Hive is slow by relational database standards and it does not support all SQL-analysis capabilities. The following will give an advantage of working on big data and SQL operations: Teradata, Query Grid, Oracle, Big data, SQL, IBM, and Big SQL.

**Analytics and BI (Business Intelligence) options designed to run on Hadoop:** These tools are a blend of SQL and BI. This has the facility of querying with big data and data-oriented advanced analytics capabilities. Examples include Apache Spark, Apache Strom, SAS Visual Analytics, Data Meer; many of these analytic engines run on Hadoop 2.0's YARN resource management system.

The value in big data analysis is often in finding correlations among disparate datasets or insights hidden in semi-structured or highly variable data sources. The tools are used to summarize, compare, represent, and predict/forecast for customers or customer-related queries.

**Example: Netflix Leverages Big Data Analytics to Understand and Customize Customer Experience**

The success of Netflix is largely due to analysing and managing big data to offer customers exactly what they want to watch at any point of time. By using the technologies like machine learning, deep learning, predictive data analytics, the company is leveraging its continuously growing data from various sources to figure out the customer preferences at an individual level. Today it has such insights as what kind of content a customer will watch on which week day, where did the viewer pause, what is the probability of she continuing to watch etc.

*Source: https://towardsdatascience.com/how-data-science-is-boosting-netflix-785a1cba7e45, 19-April 2020, Accessed on 3/08/2022.*

**Activity 9.1**

**Big Data on Cloud Platform**

A private hospital has a chain of orthopedic clinics across Mumbai. It manages its surgical equipment inventory using an existing IT solution which is well-connected over the data network. But to handle complex surgeries, it requires certain imported surgical equipment. It approaches IT solution provider for developing larger software to handle its imported equipment operations electronically, and to analyze cost and transport requirements from a pool of suppliers.

Mr. Ram Sastry, who is the project manager, gave Cloud Platform's "Big Data" application as a suitable alternative. If you were the technical lead, how would you support his choice to convince the client?

**Answer:**

## Check Your Progress - 1

1. Which of the following is not a characteristic of Big Data?
    a. Volume
    b. Velocity
    c. Variety
    d. Totally Structured Data
    e. Veracity
2. Which of the following is an example of unstructured data?
    a. Database
    b. Word File
    c. Doctor's Prescription
    d. File Management System
    e. Note Pad Text Document

3. Which of the following is the right set of challenges for big data?
   a. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data, Size about the Data
   b. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Volume of the Data, Inferences about the Data
   c. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data
   d. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data, Inferences about the Data
   e. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Discrepancies in the Data, Inferences about the Data
4. The term visualization refers to which of the following large volumes of data?
   a. Archiving
   b. Pooling
   c. Organizing
   d. Viewing graphically
   e. Retrieving
5. Which of the following bytes equals to size of Petabytes?
   a. $10^{12}$
   b. $10^{13}$
   c. $10^{15}$
   d. $10^{24}$
   e. $10^{18}$

## 9.6 Big Data Storage Considerations

Every company/organization generates some data in some format or other and many others in the organization use/analyze the data as per their requirements. Proper data storage is a huge challenge. The following will take you through the data storage evolution.

### 9.6.1 Data Storage Evolution

The evolution in data storage has been marked from magnetic tapes of the 1920s to floppies, CDs, DVDs, pen drives, hard disks to holographic data storages.

Present day organizations are using cloud-based storage in which businesses are assured of security for their data, disaster recovery and faster access due to high bandwidth internet being available.

The issues with respect to data storage of big data have been initiated in the earlier section. Moving further, the currently available internet speed and infrastructure raise storage and transport issues as it takes a long time for data transfer. This is addressed by transferring only the required data and running the code, subsequently transmitting only the results.

### 9.6.2 Data Storage Issues

Data storage and management issues like access, utilization, updating, governance, and reference are the major stumbling blocks. The sources of the data are varied in terms of size and format, and the method of collection, unlike the collection of data using manual methods, where rigorous protocols are often followed. In order to ensure accuracy and validity, digital data collection has become much more flexible and methodologies are being developed. Other issues in big data storage are 'privacy' and 'security'. The most important thing is, it includes conceptual, technical as well as legal significance. The personal information of an individual is combined with external data sets. The person may not prefer to share this secretive information with others.

### 9.6.3 Data Accessing and Sharing

Another important issue is data accessing and sharing information. In the present day open society, anybody, including the government and regulatory agencies, can access information from any corner. Thus, the agencies which manage and maintain data should be very careful in giving access to the databases. In addition, one needs to be cognizant of data ownership issues. From the point of human resource issues, big data is an upcoming technology and people need to be trained in this area to handle the storage and manage it.

**Technical Challenges:** Following are the challenges that accompany big data:

i) *Fault tolerance:* With the latest technologies like cloud computing, there could be a failure of servers. The failure must be within acceptable levels and it should not lead to starting from scratch.

ii) *Scalability:* The scalability issue of big data led to cloud computing, which now aggregates disparate workloads with varying performance goals into very large clusters, and the respective storage requires latest hardware technology like solid state drives.

iii) *Quality of Data:* Quality of the data is another issue. The data in big data is used for predictive analytics and other advanced analysis and it requires good quality data. Quality checking is a big challenge and needs a robust methodology.

**Example: Agoda selects VSAT's Universal Storage Solution for its Big Data Storage to Optimize Storage Costs and Serve Its Customers at Affordable Prices**

Agoda is a digital travel platform enabling the customers to book tickets online. The company chose VAST's Universal Storage platform for its big data foray. Efficient infrastructure was the main criteria in the selection process of Agoda. The solution allowed the company to deploy all the Spark workloads on a simple scalable big data platform. The high-performance infrastructure enabled thousands of customers to derive the benefit of Real time data analytics in selecting the best deals on flights, hotels, etc.

*Source: VAST Data selected by Agoda for big data applications | Security Info Watch, 10-Dec-2021, Accessed on 6th September, 2022*

## 9.7 Moving from Operational Dashboards to Evidence-Based Decision Making

The present era is of big data. People believe that Big data = Big opportunity and Big money.

*"Concepts without percepts are empty; percepts without concepts are blind".*

- German Philosopher
Immanuel Kant (~1780)

This quote gives us a hint that neither intuition/theory (concepts) nor data (percepts) can exist on their own. They need to be combined together scientifically for getting meaning out of the data given.

Operational Dashboards are informative templates that give quick information on the subject to the concerned decision makers. Dashboards have become front-end and the first line of access for business intelligence. The best way to get insights into an organization and performance is through dashboards.

For example, looking at sales operations, looking into the performance of salespeople, individual product categories, understanding customer behavior, and demographics.

These can be monitored together or separately as part of broader initiatives. The value this brings to the business is significant. Once companies gain regular insights into these performances, they dig deeper into the data. For instance, sales of products/services can identify (1) what products or services are most successful, (2) how to drive sales based on the leads, (3) how demographics is related to sales, etc. An integrated approach building solutions for the decision making using the data from multiple sources will help the organizations better.

Areas where organization's benefit from the dashboards include:

- Time saving
- Money saving

- Insight into customer behavior
- Aligning strategy with tactics
- Ensuring a widespread, goal-driven approach and
- Performance culture

Let us spend some time on operational and analytical dashboards.

***Operational dashboards*** manage internal and intra-daily business processes, frequently changing current performance metrics or key performance indicators. The areas span from sales, marketing, helpdesk, supply-chain, etc. Overall operational dashboards are best suited to departments that require low latency data feeds and a continual view into happenings within the business unit. In short, operational dashboards are meant to help an organization understand if its performance is "on or off" the target and by how much it is on/off the target in real-time.

***Analytical dashboards*** focus more on gaining insights from a volume of data collected over time – last month or last quarter – and use this to understand what happened. This helps in deciding, why and what changes should be made for the future. For instance, organizations may want to compare trends over time or products' performance across regions. Companies want to measure the success of marketing campaigns by combining the sales data with product placements and marketing campaign strategy. Analytical dashboards use sophisticated advanced statistical models with what-if analysis, etc. These dashboards are regularly used by business analysts and experts, who are responsible for preparing the reports for general consumption. In short, analytical dashboards are meant to help an organization visualize the real-time key metrics and performance indicators. Dashboards give insights into a situation.

### 9.7.1 Evidence-based Decision-making

Evidence-based decision-making will facilitate the decision maker to make decisions based on data/information.

Business success depends on gaining new insights faster than the competition and turning these insights into good decision making which in turn depends on good data. Many organizations have huge volumes of data. But the dilemma is which data/information to use. This is where evidence-based management comes in. It appears simple and straight that organizations have problems in collecting reliable and relevant information. However, the technique of finding relevant data among the huge amounts of datasets is available through analytical tools. Hence, the data can be used to turn it into information and knowledge for the benefit of the organization. Top organizations use this approach to ensure that they collect the most relevant information to support key decisions. The big data management techniques come into picture when right data is picked.

Relevant analytical tools can be operated on the big data for evidence–based decision-making.

The following steps are in use for evidence-based decision making:

- Define the objective - research problem - formulate the hypothesis.
- Collect the data from data warehouses of big data/cloud spaces.
- Analyze the data using big data analysis techniques.
- Validate the hypothesis and present the findings.
- Take decisions based on this information.

Big data management and analysis will give us information based on conclusions which can be used for validating the hypothesis/the business statement proposed.

The data analysis for decision making from big data requires real-time, stream computing capabilities. These enable one to combine multiple information sources to facilitate decision and initiate action. The stream computing helps business move from predefined decision paths to advanced analytical techniques making use of software like R, etc. Most of the organizations see a decrease in storage and other infrastructure costs after implementing a real-time analytics platform.

One needs to emphasize that big data power does not erase the need for human insight. The companies that are moving from operational dashboards to evidence-based decision-making need to take caution. There is either too much evidence present or the evidence does not apply. A combination of operational dashboard inputs, evidence-based inputs and decision makers' expertise leads to a better decision.

**Example: Department of Veteran Affairs Implements a Solution which Empowers to Take Evidence-based Decision Making**

The Department of Veteran Affairs has set up an organization wide data analytics platform Rockies. The platform allows information-based decision making leading to better health for the families of the veterans. The costs are also reduced for the veterans and the general tax paying public. The platform has a data lake which is divided into zones and contains some 3 million data rows. One of the new benefits is the ability to track whether the veterans are taking the prescribed medications.

*Source: USDA stands up 500 data dashboards from executive brainstorming | Federal News Network, 20th October, 2020, Accessed on 30/04/2022.*

## 9.8 Role of Cloud Computing in Big Data Management

Till a decade back, few functionaries of the business used to generate data and the entire organization used the data. Now, many business entities/organizations

are generating data. Data is growing exponentially in different formats: unstructured, structured, semi-structured, etc.

What is Cloud Computing?

Pay - use - store and be secure

Distributed computing on the internet was introduced way back in 1950 by IBM as RJE (Remote Job Entry process). In 2006, Amazon provided the first public cloud, Amazon Web Service. Usually, cloud computing has three components — client computers, distributed servers and data centers. Clients are the people who make use of the cloud facility; they usually use mobiles, thick and thin client systems. Data center is the collection of servers, where the application is placed and accessed via the internet. Distributed servers are in geographically different locations. Those act as if they are next to each other. The central server administers the data traffic and attends to the client demands, etc. It functions through special software called middleware.

**SaaS**: Software as a service (SaaS; pronounced / sæs /) is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. It is sometimes referred to as "on-demand software", and was formerly referred to as "software plus services" by Microsoft.

**PaaS:** Platform as a service (PaaS) or application platform as a service (aPaaS) is a category of cloud computing services that provides a platform allowing customers to develop, run and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app.

**IaaS:** Infrastructure as a Service (IaaS) is a form of cloud computing that provides virtualized computing resources over the Internet. IaaS is one of the three main categories of cloud computing services, alongside Software as a Service (SaaS) and Platform as a Service (PaaS).

Types of cloud deployment models in use are:

- Public Cloud
- Private Cloud
- Community Cloud and
- Hybrid Cloud

Cloud services are popular since they can be customized. One need not buy a software; it reduces the complexity of networks, enhances affordability, and its services are reliable, efficient and scalable.

**Activity 9.2**

**Cloud for Archrivals**

An apparel manufacturing unit in Mumbai has a number of outlets across the country. They want to develop a system to monitor CCTV footages of all their stores and store them on a cloud platform so that, in case of emergency, the footage can be viewed at their headquarters to manage all products, missing issues and the related complaints with law enforcing agencies. Suggest the cloud deployment approach that is used, the data that needs to be archived on cloud, and the allied services to be used in future from cloud.

**Answer:**

### 9.8.1 Cloud Computing and Big Data Management

Adaptability of Big Data is increasing in the recent years due to competitive price, safety and facility to hold big data. Researchers are of the opinion that this data can be analyzed and used for better business and for a better society. For example, analysis of which person purchases which item on what day, analysis of the density of vehicular activity on coming week in different routes with maximum levels of accuracy, etc., help many decisions in business and governance.

**Big data on servers:** Big data processing on clouds might involve hundreds of entities such as application servers accessing data and this will generate massive read/write requests. Hence, large numbers of servers are interconnected such that the failure of a few of the servers does not stop the entire service.

**Data read write:** One of the techniques is distributed data store, specifically, distributed Key Value Store (KVS). With this technique, a data structure composed of keys and values is distributed and is stored in a number of servers. Read and write take place on one server according to the request specified by a key to return its response. It offers the best performance because the load can be prevented from getting concentrated on one server even if many requests are pouring in. The chances of failures are less since data is copied and stored on many servers at a time and processed. Thus, failure of some servers does not lead to loss of data.

**Complex Event Processing:** Another methodology is parallelization of Complex Event Processing (CEP). CEP refers to the technology that processes and analyzes in real-time, the complicated and massive event series those are constantly generated in real-world activities and operations. This model allows the queries to be dynamically distributed, system-wide according to the event stream characteristics and load data, thereby optimizing the processes.

*Dynamic load balancing CEP has the following technical challenges:*

Applying real-time, cost-effective CEP in the cloud has been an important goal in recent years. Distributed Stream Processing Systems- DSPS have been widely adopted by major computing companies such as Facebook and Twitter for performing scalable event processing in streaming data. However, dynamically balancing the load of the DSPS' components can be particularly challenging due to:

- The high volume of data
- The components' state management needs and
- The low latency processing requirements

Systems should be able to cope with these challenges and adapt to dynamic and unpredictable load changes in real-time.

One approach makes the following contributions:

(i) Formulate the load balancing problem in distributed CEP systems as an instance of the job-shop scheduling problem, and

(ii) Present a novel framework that dynamically balances the load of CEP engines in real-time and adapts to sudden changes in the volume of streaming data by exploiting two balancing policies. The detailed experimental evaluation using data from the Twitter social network indicates the benefits of this approach in the system's throughput.

In the present day, business or real-life situations and innovations are the key differentiators. Big data and cloud computing will facilitate these innovations and the data is analyzed in multiple ways to bring out the innovations much beyond data mining.

**Example: Cognizant Technology Solutions helps a Financial Services Customer Rein in its Costs in the Microsoft Corp. Azure**

CTS (Cognizant Technology Solutions Corp), the IT major was selected by a financial services company to save costs while deploying Microsoft Azure Cloud. CTS found many unused copies of data bases and some of them very huge. Round the clock running of some of the virtual machines (even when they are not needed) also was observed to lead to unnecessary rental costs. The company was only concerned about the customer service deadlines. It never focussed on unnecessary costs. CTS made the company realize this and this led to huge savings.

*Source: Why your cloud computing costs are so high - and what you can do about them – SiliconANGLE, 28th November, 2021, Accessed on 30/04/2022.*

**Check Your Progress - 2**

6. What does CEP stand for?
    a. Communication Engine Processing
    b. Complex Element Processing
    c. Complex Event Processing
    d. Computerized Event Processing
    e. Customized Event Processing
7. In a grid-based cloud implementation, the time required for data migration is:
    a. Hours
    b. Days
    c. Weeks
    d. Months
    e. Minutes
8. Which of the following is not a cloud deployment model in use?
    a. Corporate Cloud
    b. Public Cloud
    c. Private Cloud
    d. Community Cloud
    e. Personal Cloud
9. In which of the following context, the Key Value Store is used?
    a. Multiple databases
    b. Multiple applications
    c. Multiple sources of data
    d. Multiple servers
    e. Cloud-enabled environment
10. Which of the following is not services and delivery model in Cloud?
    a. SaaS
    b. PaaS
    c. IaaS
    d. gPaaS
    e. MaS

## 9.9 Summary

- Big Data is a collection of structured, semi-structured and unstructured data, which has volume, variety and velocity as its characteristics.
- Big data is a set of tools and techniques that require innovative/latest methods of integration to find hidden patterns from large data sets that are complex, diverse and massive.
- Qualities of data, understanding its nature, the inconsistencies of big data are major concerns while consolidating and integrating it on a single platform for analysis.
- But the ability of a number of BI (Business Intelligence) and analytical tools like Apache Spark, Apache Storm, SAS Visual Analytics, and Data Meer to manage large volumes of data efficiently is restricted.
- A big data case study to highlight the benefits of implementing a prediction mechanism for forecasting product demand of bakery product manufacturer in Europe.

## 9.10 Glossary

**Business Intelligence:** It is a technology-enabled data analyzing and presentation method for displaying information to help business executives and managers in making decisions.

**Dashboard:** It is a user interface which organizes and presents information in an understandable format such as a graph. Usually, dashboards are helpful in drawing inferences from the graphics, useful in the decision-making process.

**Hadoop:** It is an open-source distributed data processing framework to handle big data used in the cloud environment. It is sponsored by the Apache Software Foundation used for executing applications on systems with a large number of nodes without interruption.

**Predictive Analytics:** Predictive analytics is a method for extracting information from existing data sets in order to identify patterns and predict future outcomes and trends.

**Unstructured Data:** Any information which does not have a pre-defined structure to organize itself is called unstructured data. The unstructured content is generally text-based documents like building designs and medical machine prescriptions.

## 9.11 Self-Assessment Test

1. Briefly discuss the different characteristics of Big Data with suitable examples.
2. Discuss how Big Data will be helpful in managing the retail sector.

3. Give a few examples of structured and unstructured data found at a hospital.
4. Explain the challenges of managing Big Data.
5. List a few BI and Analytical tools used on a cloud platform.

## 9.12 Suggested Readings / Reference Material


1. Rodney Heisterberg and Alakh Verma (April 2022). "Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video and Big Data Enables Competitive Advantage," Narrated by Stephen Graybill.
2. Jonathan S Walker (2021). Social Media Marketing For Beginners - How To Make Money Online: Guaranteed Strategies To Monetizing, Mastering, & Dominating Any Platform For Your Brand, JW Choices.
3. Barry Connolly (2020). Digital Trust: Social Media Strategies to Increase Trust and Engage Customers, Bloomsbury Business.
4. Seema Gupta (6 August 2020). Digital Marketing McGraw Hill; Second edition.
5. Tracy L. Tuten, Michael R (15 June 2020). Solomon et al, Social Media Marketing, SAGE Publications Pvt. Ltd; Third edition.
6. Paul Martin Thomas Erickson (2019). Social Media: Usage and Impact, Global Vision Publishing House, 2 edition.
7. Steve Randazzo (2019). Brand Experiences: Building Connections in a Digitally Cluttered World, Paipen publishing.


## 9.13 Answers to Check Your Progress Questions

**1. (d) Totally Structured Data**

Big data is distributed and is a mix of both structured and unstructured data.

**2. (c) Doctor's Prescription**

A medical prescription is a classic example of unstructured data as it is generally handwritten.

**3. (d)** Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding

of the Data, Quality of Data, Discrepancies in the Data, Inferences about the Data

**4. (d) View Graphically**

Visualization refers to the viewing of large volumes of data in graphics mode for better understanding.

**5. (c) $10^{15}$ bytes**

Petabyte is a memory measuring metric which accommodates $10^{15}$ bytes of content.

**6. (c) Complex Event Processing**

Complex Event Processing (CEP) refers to technology that processes and analyzes in real-time, complicated and massive event series that are constantly generated in real-world activities and operations.

**7. (a) Corpoarte Cloud**

Types of cloud deployment models in use include: Public Cloud, Private Cloud, Community Cloud and Hybrid Cloud. Corporate cloud is not a cloud deployment model in use.

**8. (e) Personal Cloud**

All ther options are existing cloud options.

**9. (e) Cloud-enabled environment**

In a cloud-enabled environment, one of the techniques is distributed data store, specifically, distributed Key Value Store (KVS). With this technique, a data structure composed of keys and values is distributive and is stored in a number of servers, and "read and write" take place in one server according to the request specified by a key to return its response.

**10. (e) MaS**

All other options are service models in Cloud.

# Unit 10

# Handling Unstructured Data

## Structure



*"The ability to capture end to end customer information is the biggest challenge for a retailer. The ability to merge structured and unstructured data into an Analytics Platform and be able to generate cohorts effectively is a challenge in today's landscape. There is sudden surge of data and ability to manage the same and generate meaning out of the same is a challenge."*

- Piyush Kumar Chowhan, VP & CIO, Arvind Lifestyle Brands

## 10.1 Introduction

Social media posts, email, and sensor data from IOT devices are generating huge volumes of unstructured data which cannot be analyzed using traditional methods. Special tools are needed for this.

In the previous unit, we have briefly looked at decision-making using big data. However, handling big data is itself a rigorous task that needs a careful categorization of data to understand how it must be handled. In this unit, we will visit the handling procedure of unstructured data.

Big Data is made of both machine and human-generated unstructured, semi-structured and structured data. Unstructured data means everything from social media posts to images of data like fingerprints, from email to weblogs on the

internet and many other forms of content. It is growing at an unprecedented pace, contributing to huge volumes of content.

Small volumes of unstructured data can be handled easily. But as the unstructured data becomes large, its analysis becomes quite complicated because of the heterogeneity of the data. The solution for this is to adopt available IT solutions. Basically, unstructured data is divided into two broad categories, textual and non-textual data. Textual data includes all the documents, e-mails, text files, and chat files. The non-textual data covers graphics, sound, movies, etc. Analytics are applied to the data for analysis purpose. Such data will support the decision-making process in areas like retail business, e-commerce, education, healthcare, marketing domain, etc. In addition, the IoT (Internet of Things) is growing and generating much larger volumes of data. To manage, store and provide security to such huge volumes of data is a complex task to handle. This unit concentrates on issues relating to unstructured data management practices. It describes various forms of data generated on social media and mobile to be analyzed and stored using cloud technology. The cost of storage, backup and recovery operations incurred by organizations are also discussed. This unit also gives an insight into different file archiving solutions along with the problems that exist for analyzing textual and non-textual content.

## 10.2 Objectives

After going through this unit, you will be able to:

- Explain the different formats of unstructured data that is generated in organizations, with an example.
- Discuss the mechanisms used for storing unstructured data.
- Explain different ways of analyzing and managing unstructured data.
- Discuss file archiving solutions possible to handle unstructured data.
- Explain the opportunities and problems with unstructured data.

## 10.3 Different Formats of Unstructured Data

Data is the heart of any business activity and it is maintained by users in different formats based on the requirement. For example, it may range from a person's name you want to remember, or an address entry in a diary or a diagnostic lab report. When this data is organized in the form of a report, it is known as information (i.e. data with a meaningful purpose).

There are two basic types of information based on the structural standard that is being followed. The first category is structured data, which is generally stored in a database based on a predefined format, known as metadata. The second category data is the unstructured and semi-structured one, which does not have a defined structure of its own.

**Unstructured Data**

Unstructured data is information that which does not have a well-defined data model and it is not systematized. Unstructured information is generally text based but may also contain data related to images, numbers and evidences. Because of this nature of information, leads to anomalies and uncertainties and becomes difficult to understand using traditional programs compared to data stored in structured databases. Common examples of unstructured data include audio, video files or No-SQL databases.

Due to advancements in technologies and availability of many tools in the market which have increased greatly in the recent years, the capability to store and process specialized unstructured data has significantly has gone up. For example, MongoDB is augmented to store documents. Apache Giraph, as an opposite example, is optimised for storing relationships between nodes.

As most of the data in organisations is unstructured, the capability to analyse unstructured data is particularly appropriate in the situation of Big Data. For example, extracting value of unstructured data such as pictures, videos or PDF documents is one of main drivers behind the fast growth of Big Data.

**Semi-structured Data**

Semi-structured data is a method of structured data that does not follow the formal structure of data models associated with relational databases or other forms of data tables, but nevertheless contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure. Examples of semi-structured data include JSON and XML are forms of semi-structured data.

The motive of this third category which is between structured and unstructured data is because semi-structured data is substantially easier to investigate than unstructured data. Many Big Data solutions and tools have the ability to 'read' and process either JSON or XML. This reduces the complication to analyse structured data, compared to unstructured data.

**Metadata – Data about Data**

A last category of data type is metadata. From a technical point of view, this is not a separate data structure, but it is one of the most important elements for Big Data analysis and big data solutions. Metadata is data about data. It provides additional information about a specific set of data.

For example, in a set of photographs, metadata could define when and where the photos were taken. The metadata then provides fields for dates and locations which, can be considered as structured data. Because of this reason, metadata is frequently used by Big Data solutions for preliminary analysis.

**Example: Toyota Mapmaster chooses Oracle Database Cloud Services to Automate its Map Production Process for its Integrated Approach to Processing Various Types of Unstructured Data**

Oracle Database for cloud was chosen by Toyota Mapmaster after comparing with competing products. The chosen Oracle database is able to manage all types of data required for map processing spatial data (road and landmark locations), graph data (road topology) structured data for attributes of roads and unstructured data about landmarks. This made the map production process more productive and efficient. The solution also avoided unnecessary data going between data bases.

*Source: Oracle Cloud Infrastructure helps Toyota Mapmaster to Further Accelerate Digital Transformation in Map Production, 6th August, 2021, Accessed on 02/05/2022.*

## 10.4 Key Issues in Storing Unstructured Data

Unstructured data contains data of any format, so it doesn't fit and cannot be organized precisely in the form of rows and columns of a database. It can be a file server that has become a dumping place for all of the Word documents, Excel spreadsheets and PowerPoint presentations which are used by organizations. It can also include photos, videos, email messages and a whole lot more.

For years, IT professionals struggled to find ways to deal with unstructured data. The three key unstructured data storage challenges they face mostly are the following:

The complete volume of unstructured data continues its unchecked growth.The unstructured data in new forms is generated all the time. For example, the log files created by IoT devices must be managed and stored at a place. This has to managed efficiently otherwise management and mining unstructured data, consumes storage capacity without adding value. Generally, storage systems don't make it easy to find unstructured data after it's been stored. In recent years, IT vendors have tried to help businesses make sense of unstructured data and make it more usable and valuable.

New technologies along with the best business processes are being adopted by organizations in different fields like education, manufacturing and healthcare due to growing competition. This has resulted in an increase in the amount of operational data at the workplace, such as spreadsheets, graphic designs, videos, and audio files. All this data is unstructured in nature and consumes expensive storage infrastructure to store and process. There are software solutions which help in processing this unstructured data. Some of the important factors to be studied in this regard are as follows:

- **Cost:** Cost of data storage ranks third in the total project cost when compared to total IT infrastructure expenditure by the companies. This made companies

to rework their storage cost strategy to minimize expenses by adopting low-cost alternatives. It is a challenging task for the organizations to handle this because of the increasing volume of data generated with time.

- **Impact on backup operations:** As the volume of data rises over time, requiring more and more storage capacities, it becomes difficult to handle backup operations by an organization. Failure to achieve total data protection can be risky while honoring Service Level Agreements (SLAs) with clients. This causes loss of data, customer service delays, and also the loss of goodwill. Unless huge investments are made by the companies to upgrade their backup system, it may lead to disruption of services.
- **Disaster recovery:** Unless the value of information is accessed and its importance is identified, it would be difficult for the organization to distinguish between very important data and less important data. This is an important factor as it allows deciding which data is critical and which is not. As a result, the companies can prioritize what data to be stored on primary disks. This helps in having a better data recovery strategy by avoiding non-mission critical data during recovery operations. It improves the speed and reduces storage requirements for data recovery.
- **Unstructured data complicates regulatory compliance and legal discovery:** Corporate and government regulations often require IT firms to store data for many years and to provide particular data on request. Unstructured data, in particular, is very difficult to categorize and locate for the purpose of archive/retrieval. When terabytes of unstructured data is involved, manual methods of searching can cripple an organization.

**Example: Alfa Romeo F1 Team ORLEN is tackling its Digital Storage (of Unstructured Data) Challenges head on with Seagate Lyve Cloud**

The Formula one team Alfa Romeo generates gigabytes of data collected by sensors on the racing cars, at the factory. Data storage is a very big challenge as the cloud rentals are going up and the complexity of the data management has increased many folds. The Alfa Romeo management chose the solution offered by Seagate (solution based on ORLEN) which offers high level of privacy, no lock-in, no billing for APIs (application interfaces). The Alfa Romeo team retains 100% control of data at lower costs. Scalability is also enhanced.

*Source: Seagate Lyve Cloud Makes Mass Data Storage Easy for Alfa Romeo F1 Team ORLEN | Business Wire, 26th April, 2022, Accessed on 02/05/2022.*

## 10.5 Unstructured Data Usage Patterns

The main problem with unstructured data is that the users do not have an idea of how to prioritize data by deciding which data is important and which is not to the

company. As a result, the data accumulated may consist of duplicate files, files which are not accessed for years, obsolete data which is stored on costly disks for which backup is to be generated. Storing, protecting and archiving of such data is very costly and time-consuming operation.

Kryder's law refers to rapid increases in magnetic drive storage density over decades, during which the hard drives have transformed from storing a few thousand individual bits of information in the 1950s to the latest innovation of small, cheaper and high-volume drives storing gigabytes and terabytes of information. This phenomenon led to the development of different deployable technologies for both commercial and research needs. Volume of storage is the pivot of measurement as per Kryder's law. Generally, users do not delete unwanted files, due to which volume of data grows enormously. These result in capacities getting exhausted, irrespective of the type of storage being used such as direct, NAS (Network Attached Storage) or attached SAN (Storage Area Network). Companies have to forklift their storage capacities by adding additional servers to their existing ones, leading to more maintenance overheads.

Without having a well-defined strategy, this problem of unstructured data management will be there to stay forever, leading to additional cost for handling storage, backup and recovery operations. To address these issues, IT companies need solutions to handle the visibility of unstructured data with the ability to automatically sort itself. This may include data quality issues like automatically identifying the usage frequency of the data, duplicate data and unused data. Its file archiving technologies should be intelligent to automatically detect these issues so that a substantial reduction of storage management cost is possible. Some other techniques include:

1. Implementation of multi-tiered automated storage.
2. Making use of metadata for data categorization and applying storage policies regarding unstructured data.
3. Developing new optimized backup strategies.

**Example: JP Morgan Chase & Co. Completes 360,000 Hours of Finance Work (Involving Unstructured Data) in Just Seconds**

JP Morgan Chase & Co. is the largest bank in the USA employing the largest number of employees. The bank's lawyers and loan officers spend a total of 360,000 hours annually going through and vetting loan agreements. Using machine learning, it is brought down to seconds. The machine learning solution also reduced errors in loan approvals. Before this solution, there were errors in processing some 12000 documents annually.

*Source: Improved Storage Allocations Prevented Wasteful Spending | Auto Parts Retailer - VSI (visualstorageintelligence.com), 2022, Accessed on 02/05/2022.*

**Activity 10.1**

**Healthcare Service Planning**

A healthcare company provides services to underprivileged children and the destitute, who are housed in various state-run rehabilitation centers managed by various NGOs and self-help groups. As the network of NGOs is very large and widespread, the healthcare service provider wants to make use of a cloud-based solution to predict monthly usage of drugs and the number of lab visits needed by the people it serves. These monthly trends will help the company to effectively plan its quarterly fund requirements. Suggest a technology that suits and efficiently handles the above given scenario.

**Answer:**

## Check Your Progress - 1

1. Unstructured data is a combination of human and which of the following generated data?
    a. Machine
    b. Nuclear
    c. Research
    d. Social media
    e. Fingerprint
2. Which of the following is an example of structured data?
    a. Database
    b. PDF
    c. Biometric
    d. Voice
    e. Medical prescription
3. Which of the following is an example of human-generated unstructured data?
    a. Text data
    b. Audio files
    c. Videos

d. Website content

e. Photos

4. What does Data recovery mean?

a. Deleting a record

b. Appending a record

c. Viewing a record

d. Reconstructing data

e. Querying data

5. What is Metadata?

a. Processed data

b. Client data

c. Big data

d. A software

e. Description of structure, administrating the data

## 10.6 Different Ways of Analyzing and Managing Unstructured Data

Predictive analytics is one of the most popular methods used to predict the occurrence of an event, such as to know the trends and to find the possibilities of future occurrence of a certain event out of the existing data universe. This is applicable in the field of digital commerce (e-Commerce), location awareness services, cloud computing, mobility, and digital security.

But, most analytics-based technologies are tailor-made to suit structured data rather than unstructured data and there is a need for attention to address the contentious issue such as, how to handle unstructured data. As unstructured data includes social media tweets, blogs, posts, and other forms, it also includes emails, images and open-ended surveys. There are proven IT Solutions for addressing the processing of this type of unstructured data

To analyze unstructured data, there are some basic steps or phases to be followed to prepare, organize and manage it.

These basic steps for preparing data for analysis are:

- Identify important data sources
    - The first step for a good analysis of unstructured data is to identify all sources of data which are essential and important. Use data which is absolutely relevant, avoid duplicate records and obsolete content for analysis.

- Set analytics method and goals
    - Without a purpose, any analysis is of no use; so, having a predefined, focused goal is very much necessary. Therefore, make sure of what output is required, for example, quantity-based trends or comparison between products, to name a few.
- Technology availability
    - Study technology options available on hand and assess their capabilities to meet the final requirements of analysis. Set up the information architecture along with factors for choosing data storage and retrieval based on scalability, volume and variety of data for the purpose of analysis.
- Real-time content
    - For online e-commerce companies, real-time data access is important to provide real-time quotes. It requires tracking of real-time events and provides an output based on predictive analytic engine's output. The data considered for analysis should include content from social media engagement and machine-generated content. The technology platform should ensure no data is lost from the real-time stream of multiple sources.
- Save the original before sending to data warehouses
    - With the advent of big data, storing information in its native format is very useful. It preserves metadata and other information that might assist analysis when required. So keep a copy of the same.
- Prepare data for storage
    - While keeping the original file, clean up a copy from the noise in the content. For example, remove noise like white spaces and symbols while converting informal text to formal language.
- Ontology evaluation
    - Analysis allows building of relationships among sources and the extracted entities for designing a structured database as per specifications. Though time-consuming, it is worth doing for quality reasons.

**Analyzing the unstructured data**

- Unstructured data is analyzed by mining. For example, while matching fingerprint, the actual fingerprint image is completely unstructured. To analyze a fingerprint, crucial facts are identified and then mapped.

The mapping is done with the structured data. Unstructured prints remain unanalyzed;

- In general, most unstructured data uses extraction, text analysis and text abstraction with a relational database to create an unified view of the data, enabling the organization to make smarter business decisions.
- Retailers such as Chico's FAS are able to integrate social media communications with its customer data to offer targeted promotions to customers, while healthcare providers such as Seton Healthcare Family are able to save money from readmitting patients by identifying readmission causes.

### 10.6.1 Big Data: 9 Steps to Extract Insights from Unstructured Data

According to many experts, about 80% to 90% of data held by an organization is unstructured data. This is in addition to the voluminous diagnostic information logged by embedded and user devices. It is difficult to make meaning out of unstructured data. Organizations need to study both structured and unstructured data to assess meaningful business decisions, including determining customer sentiment, cooperating with e-discovery requirements and personalizing product for their customers. This must be done to ensure that the organization is on top of any network security threats, and proper functioning of embedded devices.

However, by carefully studying the disparate sets of unstructured data, one can identify connections from unrelated data sources and find patterns.

Along with the increase in the digitization of information the gathering of multi-channel transactions has resulted in a data overflow. The ever- increasing leap of digital information has led the world's collective data to double in even shorter intervals than ever before. According to Gartner, about 80% of data held by an organization is unstructured data, which consists of information from customer calls, emails and social media feeds. This is in addition to the huge analytical information recorded by fixed and user devices. It is very discouraging that making proper analysis from organized data, but it is more difficult to extract intelligence of unstructured data.

As a result, organizations have to study structured, semi- structured and unstructured data sets to arrive at meaningful business decisions, including determining customer sentiment, cooperating with e- discovery requirements and personalizing the offerings for their customers.

Going through huge amounts of information can look like a lot of work, but there would be encouraging results. By reading large, disparate sets of unstructured data, one can identify connections from unrelated data sources and find patterns. The most important aspect which makes this method effective is that it helps in discovering trends.

There are nine steps to analyze unstructured data so that one can see more than meets the eye. They are:

1. **Make sense of the disparate data sources**

   Before one can begin, one needs to know what sources of data are important for the analysis. One information channel is log files from devices, but that source won't be of much help when searching for user trends. If the information being analyzed is only tangentially related to the topic at hand, it should be set aside. Instead, only use information sources that are absolutely relevant.

2. **Sign off on the method of analytics and find a clear way to present the results**

   The analysis is useless if it is not clear what the end result should be. One must understand what sort of answer is needed - is it a quantity, a trend or something else? In addition, one must provide a roadmap for what to do with the results so that they can be used in a predictive analytics engine before undergoing segmentation and integration into the business's information store.

3. **Decide the technology stack for data ingestion and storage**

   Even though the raw data can come from a wide variety of sources, the results of the analysis must be placed in a technology stack or cloud-connected information store so that the results can be easily utilized. Factors that are important for choosing the data storage and data retrieval depend often on the scalability, volume, variety and velocity requirements. A potential technology stack should be well evaluated against the final requirements, after which the information architecture of the project is set.

   A few likely influential requirements are that the results of the analysis must be available in real-time, have high availability for access while still functioning in a real-time multi-tenant environment. Real-time access is crucial, as it has become important for e-commerce companies to provide real-time quotes. This requires tracking real-time activities, and providing offerings based on the results of a predictive analytics engine. Technologies that can provide this include Storm, Flume and Lambda. High availability is crucial for ingesting information from social media. The technology platform used must ensure that no loss of data occurs in a real-time stream. It is a good idea to use a messaging queue to hold incoming information as part of a data redundancy plan, such as Apache Kafka. The ability to function in real-time multi-tenancy environments is required if the results are required to avoid state changes and continue to be mutable data.

4. **Keep information in a data lake until it has to be stored in a data warehouse**

   Traditionally, an organization obtains or generates information, sanitizes it and stores it away. For example, if the information source is an HTML file, the text may have to be stripped and the rest discarded, such that information is not lost during storage in a data warehouse.

   Anything useful that was discarded in the initial data load was lost forever, and as a result, the only thing one could do with the data was with what was available after extraneous information was stripped away. The appeal of this prior strategy was that the data before sanitization was in a pristine, mutable format that could be used whenever needed. However, with the advent of Big Data, it has come into common practice to do the opposite, i.e., store first and sanitize later. With a data lake, information is stored in its native format until it is actually deemed useful and needed for a specific purpose, preserving metadata or anything else that might assist in the analysis.

5. **Prepare the data for storage**

   While keeping the original file, if one needs to make use of that data, it is best to clean up a copy. In a text file, there can be a lot of noise or shorthand that can obscure valuable information. It is a good practice to cleanse noise like whitespaces and symbols, while converting informal text in strings to formal language. If it is possible to detect the spoken language, it should be categorized as such. Duplicate results should be removed, the dataset treated for missing values, and off-topic information removed from the dataset.

6. **Retrieve useful information**

   Through the use of natural language processing and semantic analysis, one can make use of Parts-of-Speech tagging to extract common named entities, such as "person," "organization," "location" and their relationships. From this, one can create a term frequency matrix to understand the word pattern and flow in the text.

7. **Ontology evaluation**

   Through analysis, one can then create the relationships among the sources and the extracted entities so that a structured database can be designed to specifications. This can take time, but the insights provided can be worthy for an organization.

8. **Statistical modeling and execution**

   Once the database has been created, the data must be classified and segmented. It can save time to make use of supervised and unsupervised machine learning, such as the K-means, Logistic Regression, Naïve Bayes,

and Support Vector Machine algorithms. These tools can be used to find similarities in customer behavior, targeting for a campaign and overall document classification. The disposition of customers can be determined with sentiment analysis of reviews and feedback, which helps to understand future product recommendations, overall trends and guide introductions of new products and services.

The most relevant topics discussed by customers can be analyzed through temporal modeling techniques, which can extract the topics or events that customers are sharing via social media, feedback forms or any other platform.

9. **Obtain insight from the analysis and visualize it**

From all the above steps, it all comes down to the end result, whatever it might be. It is crucial that the answers to the analysis are provided in a tabular and graphical format, providing actionable insights for the end-user of the resultant information. To ensure that the information can be used and accessed by the intended parties, it should be rendered in a way that it can be reviewed through a handheld device or web-based tool, so that the recipient can take the recommended actions on a real-time or near-real-time basis.

The above steps are used for planning analysis of unstructured data and there are many business intelligence (BI) platforms that are already available from different companies like SAS, Tibco's Spotfire, SAP/NetWeaver and IBM/Cognos, to name a few.

## 10.7 Comprehensive File Archiving Solution

'Big Data' includes machine-generated data, IoT data; for example, web search logs, satellite images and other data like audio, video and health records. Analytics are performed on such data to support the decision and support applications. On the other hand, human-generated unstructured data is also a constituent of big data, which is to be organized to suit data mining activity. Human-generated data includes contracts, tenders, CADs (Computer Aided Design), and content from mobile phones and social media whose summarization is important for the organization. Organizing unstructured data is complex and time-consuming as it needs a strategy and integration method to club with structured data.

For considering the implementation of a big data-based archiving system in an organization, the following aspects are important:

- Reducing primary storage consumption
- Supporting data growth
- Identifying non-active data
- Multi-tier architecture for data storage
- Ability to retain, categorize and in future mine the data

- Compliance standards followed to suit both structured and unstructured data
- Density scaling
- High throughput
- Fast retrieval
- Protection and disaster recovery

Let us study the Strongbox technology to manage and handle unstructured data.

**Strongbox:** The usage of IT has been widespread across all walks of life. Data associated with these applications is also exponentially growing over time. Data from these IT-enabled applications need to be stored indefinitely. Its security and retrieval are also very important to the organization to access it when it is required.

Strongbox technology addresses these issues, within the reach of the organization's budget. Strongbox is a Network-Attached Storage (NAS)-based technology to simplify data access. It provides cost-effective, Linear Tape File System (LTFS)-based storage for structured/unstructured data. Strongbox ensures complete data protection for all online and archive storage needs.

Advantages of using Strongbox are:

- Controls storage costs
- Reduces TCO (Total Cost of Ownership) by more than 50%
- Provides online and archive storage with an easy-to-use solution
- Ensures long-term protection of files
- Eliminates backup requirements of fixed content
- Delivers performance to meet application workflow
- Reduces the gap between data growth and budget
- Provides an LTFS-based solution.

**Example: The Safety and Compliance firm J. J. Keller & Associates is Working with Hewlett Packard Enterprise (HPE) to Power its Cloud Transformation for its Data Analytics and Archival Needs**

Hewlett Packard Enterprise (HPE) was chosen by J.J. Keller & Associates (the safety and compliance firm) in its cloud migration journey. The cloud migration is planned to cater to growing demands for data storage and data processing for various applications running on premises and on remote data centres. The migration was able to consolidate workflows across various platforms. The number of platforms came down from 6 to 2.

*Source: J. J. Keller Taps HPE for Cloud Infrastructure | Datamation Date, 14th February, 2022, Accessed 02/05/2022.*

## 10.8 Opportunities and Problems with Unstructured Data

The data is continiuosly available which is mostly unstructured whch comes in the form of messages, mails. It's a challenge for an organozatin to handle and mangae this data due to the following:

- Lack of availability of tools that can easily manage unstructured data. Tools need to offer effective text describing and analytics, taxonomy and metadata management.
- Difficulty in assimilating unstructured data with existing information systems. The two are often seen as apples and oranges when it comes to analytics and decision making.
- Shortage of skills in existing staff
- Missing sense of perseverance for managing unstructured data

Regardless of the best hard work to enclose the unstructured data, it continues to grow larger and presents a real problem for organizations that want to automate and improve their ability to understand their business, anticipate what's coming and act quickly on risk and opportunity. There are certainly tools that are growing and providing the solution. Nevertheless, the challenge, will be in finding the urgency and getting the organizations understand the value of getting data out of its various hiding places and into a place that it can be used and valued.

It is estimated that about 80 to 85 percent of the useful data is in unstructured format. These are stored and processed on IT-enabled applications or technologies. Most of such data in smaller quantities can be analyzed by users in an optimized manner using their IQ and natural intelligence. But the same is not possible in the case of large amounts of unstructured data. The only alternative is to take the assistance of IT solutions available. Due to the existence of different kinds of unstructured data, the analysis is quite complicated. Basically, unstructured data is divided into two broad categories, namely, textual and non-textual data. Textual data includes all the documents, e-mails, text files, and chat files. Similarly, the non-textual data like graphics, sound, movies, etc., are also stored and processed.

There are a few sticky issues in this categorization. For example, in the case of a PowerPoint presentation, it is a mix of both textual and non-textual data. So, to which category does it belong to?

**i) Textual data**

There are many things that human beings can do which are not possible by machines like computers, especially when dealing with unstructured textual data originating from a number of avenues of communication like SMS,

email or a social media post. Let us look into the characteristics of text to understand it better.

a) **Context:** In the textual content, these problems may arise because of context. For instance, the difference between the following statements can be clearly marked "John rides in a Mustang" and "John rides on a mustang". The first statement mustang means a car (Mustang is a popular car model by Ford); in the second case, it may be considered a horse (an American feral horse which is typically small and lightly built is called mustang). The human brain picks up all of these options almost instantaneously and understands it implicitly. Computers cannot do the same as they have to be exclusively instructed on how to understand this statement?

b) **Style, structure and vocabulary:** Sometimes writing style, sentence structure and vocabulary used in the textual documents are different, while comparing e-mails and SMS messaging.

c) **Addressed to whom:** While speaking to friends, children, or boss, we use different styles with each of the groups or individuals. Human brain generally handles these changes easily based on the situation, but for computers it is a difficult task to handle.

The tools used to analyze textual data are search based on keyword/key phrases. Analysts used to prepare a large list of relevant keywords needed by computers to search large volumes of data. This method had its own problems like: if the search criteria are too narrow then there are chances of missing vital information, if it is too broad then the output may contain irrelevant data. Some advances are made by applying concepts like artificial intelligence, fuzzy approach and neural networks to address this type of problems. Neural networks are used to recognize simple tacit linkages between words and expressions.

**ii) Non-textual data**

Some of the characteristics and formats of non-textual data include:

a) **Common format:** In the case of non-textual data, the major problem for the computers is how to convert it into common searchable form. To handle large amounts of this data by converting multiple forms of data into a common single format, this can be used for performing searches. For instance, Optical Character Recognition (OCR) systems are used to perform this conversion activity reliably.

b) **Audio:** Modern voice and language recognition systems are becoming more and more capable of converting voice recordings to text as

technology advances. Many of the language recognition systems are based on neural network and artificial intelligence principles.

c) **Graphics/Video:** Dealing with graphics or movie files is a very complicated problem as it requires a significant amount of processing power. Due to the presence of possible illusions, searching of video content is possible by using a specialized predefined set of keywords. For example, identifying adult content and racist comments, etc.

**Activity 10.2**

**Data Storage Solution**

An automobile parts and service provider serves auto workshops by running a network of retail centers at various villages. It is a noble initiative, to reach and serve rural customers in their village to reduce their transport costs. Currently, the company does not have the proper hardware/software to securely store and manage its unstructured data, especially, the sales records of various parts are present at various centers. The company maintains many nodal centers at various district headquarters to consolidate and manage sales records procured from various rural retail centers in the surrounding villages in that district. They want to implement a cloud-based storage and data management technology connecting all nodal centers.

Suggest a cost-effective data storage technology suitable to address the situation. Also, list some of the major features of the technology suggested. They are planning to increase their operations if found beneficial.

**Answer:**

**Check Your Progress - 2**

6. Which of the following is an example of a non-textual data?
   a. Documents
   b. SMS messages
   c. databases
   d. Datawarehouses
   e. Videos

7. VideosWhich of the following is not an advantage with StrongBox solution?
    a. Controls storage costs
    b. Reduces TCO by more than 50%
    c. Provides online and archive storage with an easy-to-use solution
    d. Ensures the long-term protection of files
    e. Focusing on a particular brand
8. What does NAS stand for?
    a. Network Attached Storage
    b. Network Access System
    c. Network Access Storage
    d. Network Attached System
    e. Network Accessing Software
9. What does SAN Stand for?
    a. Social Area Network
    b. Storage Area Network
    c. Service Area Network
    d. Server Area Network
    e. Structured Area Network
10. What is full form of LTFS?
    a. Linear Type Flow System
    b. Linear Tape File Service
    c. Linear Tape File System
    d. Linear Technology for Files and Services
    e. Large type File system

## 10.9 Summary

- Big Data is a collection of structured and unstructured data, which includes unorganized data generated by humans and machines.
- Most of the existing data in organizations is unstructured in nature, which is nearly 80% of the total volume. Information which is not based on a well-defined structure is called unstructured data.

- Handling such unstructured data to manage, maintain and analyze is very difficult due to the lack of uniformity in its content. Its volume is on the rise as more and more technology is being used by companies. Proper analysis is the key to effectively using Big Data for decision-making. Though analyzing unstructured data is complex and time-consuming, a few companies like SAP and IBM have developed many technology platforms to address these issues.
- As Big Data comprises data from various sources, operations to store, backup and recovery of data in a secure manner are very crucial for all business activities. Analyzing unstructured data has many unresolved problems which are better handled by humans when compared to computers. Some of the issues in this regard are the context, style of writing and the receiver. Artificial intelligence and neural networks are being used to overcome these gaps.

## 10.10 Glossary

**Data lake:** Storage repository that holds a vast amount of raw data in its native format until it is needed.

**Kryder's Law:** Kryder's law represents an analysis of hard drive's density and capability to store data over time. It correlates Moore's law concept which says that the number of transistors placed on an integrated circuit should double every two years, resulting in predictable progress in microprocessor speed.

**Linear Tape File System:** It is a tape file system by IBM, which enables tape media to be read by the operating system when the magnetic tape is inserted into the tape drive.

**Mutable data:** A mutable object can be changed after it is created. Custom classes are generally mutable.

**NAS:** Network-Attached Storage (NAS) is a file-based computer data storage server which facilitates data access to a heterogeneous group of clients connected over a data network.

**Predictive Analytics:** Predictive analytics is a method of extracting information from existing datasets in order to identify patterns and predict future outcomes and trends.

**Real-time multi-tenant environment:** A tenant is a group of users who share a common access with specific privileges to the software instance. Multi tenant refers to more than one and real-time addresses the immediate nature.

**SAN:** A Storage Area Network (SAN) is a dedicated high-speed network that interconnects shared pools of storage devices with multiple servers.

**State changes:** State refers collectively to the data stored in the object that determines the current properties of the object, and a change refers to its alteration.

**TCO:** Total Cost of Ownership (TCO) is a cost estimate of the total cost incurred for purchasing the product along with the operational costs involved to maintain it. This helps buyers and owners in decision-making while purchasing a product or a system.

## 10.11 Self-Assessment Test

1. Briefly discuss the different forms of unstructured data known to you with suitable examples.
2. Discuss the relevance of big data in the banking industry.
3. Describe the key elements involved while preparing unstructured data for analysis.
4. Explain the following terms:
   a. Non-textual data.
   b. Backup and Recovery.
5. Does 'Big Data' technology support decision-making in an organization? If yes, give a good example in the areas of marketing, education and insurance domains.

## 10.12 Suggested Readings / Reference Material

1. Rodney Heisterberg and Alakh Verma (April 2022). "Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video and Big Data Enables Competitive Advantage," Narrated by Stephen Graybill.
2. Jonathan S Walker (2021). Social Media Marketing For Beginners - How To Make Money Online: Guaranteed Strategies To Monetizing, Mastering, & Dominating Any Platform For Your Brand, JW Choices.
3. Barry Connolly (2020). Digital Trust: Social Media Strategies to Increase Trust and Engage Customers, Bloomsbury Business.
4. Seema Gupta (6 August 2020). Digital Marketing McGraw Hill; Second edition.
5. Tracy L. Tuten, Michael R (15 June 2020). Solomon et al, Social Media Marketing, SAGE Publications Pvt. Ltd; Third edition.
6. Paul Martin Thomas Erickson (2019). Social Media: Usage and Impact, Global Vision Publishing House, 2 edition.
7. Steve Randazzo (2019). Brand Experiences: Building Connections in a Digitally Cluttered World, Paipen publishing.

## 10.13 Answers to Check Your Progress Questions

**1. (a) Machine**

Unstructured data is made of both machine and human-generated data. Such data can support the decision-making process in organizations in different domains such as education, healthcare and marketing.

**2. (a) Database**

Database is an example of structured data as it is based on a pre-defined structure called metadata.

**3. (d) Website content**

This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

**4. (d) Reconstructing data**

Data recovery is a method to rebuild data which is generally done when there is a loss of data due to technical reason.

**5. (e) Description of the structure of data**

Metadata is also called the catalog which describes the structure of the data.

**6. (e) Videos**

Due to the presence of possible illusions, searching of video content is possible by using a specialized predefined set of keywords.

**7. (e) Focusing on a particular brand**

In strong box solution, all the other factors are advantageous except Focusing on a particular brand

**8. (a) Network-Attached Storage**

NAS is a file-based computer data storage server which facilitates data access to a heterogeneous group of clients connected over a data network.

**9. (b) Storage Area Network**

Storage Area Network (SAN) is a dedicated high-speed network that interconnects shared pools of storage devices to multiple servers.

**10. (c) Linear Tape File System**

LTFS stands for Linear Tape File System. It is a tape file system, which enables tape media to be read by the operating system when the magnetic tape is inserted into the tape drive.

## Unit 11

# Data Analytics for Top Management Decision Making

**Structure**



*"Big data will spell the death of customer segmentation and force the marketer to understand each customer as an individual within 18 months or risk being left in the dust."*

– Virginia M. (Ginni) Rometty, chairperson, president, and CEO IBM.

## 11.1 Introduction

Only those companies which use data analytical tools on Big Data to get insights into customer preferences at the individual customer level and take appropriate

decisions to delight the customer will remain in the market. These decisions at strategic, tactical, and operational level driven by data decide the future growth of companies.

In the previous unit, we have discussed unstructured data handling methodologies in detail. We have also studied the conversion strategy of unstructured data to structured data which facilitates the environment to apply data analytics. In the present unit, we will discuss data analytics at large to understand the inference procedure out of the big/enormous data collection.

Focus on business analytics has increased in the past decade, reaching many organizations and a wider range of users like executives, business managers, analysts and knowledge workers within the organizations. With fast growing data volumes due to increasing use of applications in companies, business analytics allows to optimize operations and at the same time to maintain flexibility.

There are many statistical models developed and extended to execute and perform Business Intelligence and Business Analytics. Various tools like Correlation Analysis, Regression Analysis, Multiple Linear Regression and Logistic Regression are discussed in this unit. Data Mining Techniques like Exploratory Factor Analysis (EFA), Principle Factor Analysis (PFA), Confirmatory Factor Analysis (CFA), Classification, Predictive Analysis, Cluster Analysis, Association Analysis, RFM (Recency Frequency Monetary) Analysis and Market Basket Analysis (MBA) are also explained in this unit.

## 11.2 Objectives

After going through this unit, you will be able to:

- Explain Business Intelligence and how it is useful for organizations
- Describe Correlation Analysis
- Discuss Regression Analysis
- Describe Multiple Linear Regressions
- Define Logistic Regression
- Explain RFM (Recency Frequency Monetary) Analysis
- Define Market Basket Analysis (MBA)

## 11.3 Business Intelligence

Business Intelligence (BI) is a broad category of application programs and technologies used for gathering, storing, analyzing, and accessing data to help users make better business decisions. BI applications support the activities of query and reporting, decision support, Online Analytical Processing (OLAP) (computer processing that enables a user to easily and selectively extract and view data from different points of view), forecasting, statistical analysis, and data

mining (examining large pre-existing databases in order to generate new information). Business Intelligence includes concepts/methods and by using fact-based support systems, it improves business decision-making.

Business Intelligence is an art of gaining business advantage from data by answering various fundamental questions. These may include, how various customers rank products, organizations, etc., how business is doing at the present stage, and if continued on the current path, which clinical trials should be taken further and which should be stopped.

The concept of Business Intelligence (BI) discusses the technologies, applications and practices for the collection, integration, analysis, and presentation of business information. The purpose of Business Intelligence is to take care of better business decision making. Basically, Business Intelligence systems are data-driven Decision Support Systems (DSS). Business Intelligence is sometimes used interchangeably with briefing books, report and query tools and executive information systems.

**Importance of Business Intelligence tools or software solutions**

Business Intelligence systems offer historical, current, and predictive views of business operations. It mostly uses the data that is collected into a data warehouse or a data mart and rarely works from operational data. Software elements support reporting, interactive "slice-and-dice" pivot-table analyses, visualization, and statistical data mining. Applications tackle sales, production, financial, and many other sources of business data for purposes that include business performance management. Information is frequently collected about other companies in the same industry which is known as benchmarking.

**Business Intelligence Trends**

Now-a-days organizations are considering that data and content should not be considered as distinct features of information management. But, they should be managed in an integrated enterprise approach. Enterprise information management brings Business Intelligence and Enterprise Content Management together. Presently organizations are moving towards Operational Business Intelligence which is currently under served and not accepted by vendors. Traditionally, Business Intelligence vendors are targeting only top of the pyramid but now there is a standard shift toward taking Business Intelligence to the bottom of the pyramid with a focus of self-service business intelligence.

### 11.3.1 Business Intelligence: Components

In most cases, Business Intelligence involves multidimensional analysis and reporting, often based on the company data warehouse to organize the needed data.

Business Intelligence includes various key components that are explained in more detail in the following sections:

i) Multidimensional analysis

ii) Reporting

iii) Data mining

iv) Financial consolidation and budgeting

v) Key Performance Indicators

**i) Multidimensional analysis**

This area covers the possibility to slice-and-dice the data (that is, the factual information) in many dimensions. This is known as pivoting data. A pivot table is a tool to build and summarize data using spreadsheets. In Excel Sheet, one can summarize data in a pivot table mode on many levels on each dimension.

**ii) Reporting**

Companies need different types of reports. In many cases, hundreds of different types of reports, and often more, are needed. Business Intelligence software often has comprehensive reporting tools that can be applied to real-time data available from internal web pages, internet and Excel and PDF format. In many cases, these reporting facilities will be controlled by parameters that can be chosen in real-time

**iii) Data mining**

Data mining, a branch of computer science, is the process of extracting patterns from large data sets using a combination of statistics and artificial intelligence approaches to study the given data to provide actionable intelligence.

**iv) Financial consolidation and budgeting**

Business Intelligence methodology also covers systems and functionalities for groups to perform financial group consolidation and budgeting. BI tools help in

Reporting, Budgeting and dashboard modules:

- It creates complete financial statements and operational reports without having to learn proprietary report writers or complex formulas.
- Users can refresh reports on-demand and drill down to detail, eliminating the need to export and reformat reports or the use of multiple report writers across transaction systems. It quickly creates and deploys input templates based on the existing budget model, to completely redesign and modernize the model.

- It can consolidate many data sources into the tool for integrated reporting and analysis.
- Build quick dashboards, highly summarized and Key Performance Indicator -based goals for executives, and detailed operational dashboards for line managers and end users across the organization.

**v) Key Performance Indicators**

The key performance indicators are metrics that are measured periodically to keep track of historical trends as well as the goal or target (as per the suitability of conditions).

The Major Components of Business Intelligence (BI) can be as follows:

**OLAP (Online Analytical Processing)**

This component of BI allows managers to categorize and select collections of data for strategic monitoring. With the help of specific software products, a certification in business intelligence helps business owners to use data to make adjustments to overall business processes.

**Advanced Analytics or Corporate Performance Management (CPM)**

This allows business leaders to look at the statistics of certain products or services. For example a fast food chain may analyze the sale of certain items and make local, regional and national alterations on menu board offerings as a result. The data could also be used to predict in which markets a new product may have the best success.

**Real-time BI**

In a current internet culture, this component of BI is becoming very prevalent. Using software applications, a business can respond to real-time trends in email, messaging systems or even digital displays. Because it's all in real-time, an entrepreneur can announce special offers that take advantage of what's going on instantly. Marketing professionals can use data to expertise inventive limited-time specials such as a coupon for hot soup on a cold day. CEO's may be interested in following the time of day and location of customers when they enter the website so marketing department can offer special promotions in real-time while the client is involved on the website.

**Data Warehousing**

Data warehousing allows business leaders scrutinize the subsets of data and examine interrelated components that can help drive business. Looking at sales data over several years will help in improving product or provide seasonal gifts. Data warehousing can also be used to look at the statistics of business processes including how they relate to one another. For example, business owners can

compare shipping times in different facilities to look at which processes and teams work most efficiently. Data warehousing also involves storing huge amounts of data in ways that are useful to different departments within the company.

**Data Sources**

This component of BI involves numerous arrangements of stored data. It's about taking the raw data and using software applications to create meaningful data sources that each division can use to positively impact business. BI analysts using this strategy may create data tools that allow data to be put into a large supply of spreadsheets, pie charts, tables or graphs that can be used for a variety of business purposes. For example, data can be used to create presentations that help to structure achievable team goals. Looking at the strategic aspect of data sources can also help organizations make fact-driven decisions that take into account a more holistic view of the needs of the company.

**Example: Airtel leverages "Business Intelligence" to Obtain more Meaningful Customer Insights**

Bharti Airtel is one of the top three global telecommunications companies. The company understands the need to offer a consistent and valuable customer experience to achieve success. The company has planned to optimize its network performance with a view to offer customers a seamless experience.

The company took up the network analytics project. The idea was to collect data from the network, store it, and run advanced analytics to obtain deep customer insights on network performance. These insights would facilitate optimization of the network to cater to individual's need.

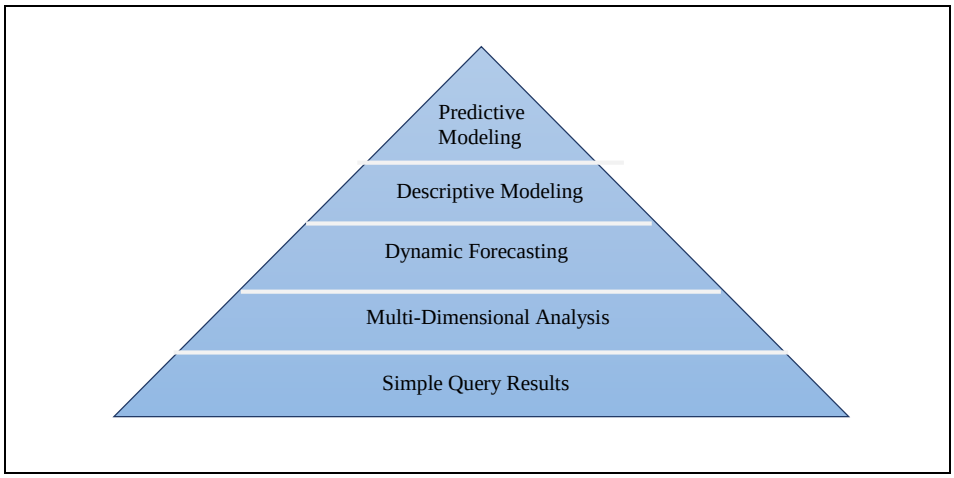
*Source: Bharti Airtel: Intelligent Network Optimization, March 2022, Accessed on 03/05/2022.*

## 11.4 Business Analytics

Business Analytics (BA) is a great tool and as per literature, it helped in decision-making. Improvements have been shown in increased profitability, reduced cost, faster decision-making, and critical performance in a business.

*Business Analytics* mostly focuses on creating different insights and understanding of the business performance based on statistical methods, quantitative analysis, predictive modeling and fact-based management to drive decision-making (see Figure 11.1). For example, the questions answered by Business Analytics would be: "why is this happening?", "what if these trends continue?", "what will happen next?", and "what will be the optimal outcome?" etc. The latest being that we are moving from diagnostic to prognostic to prescriptive, covering "Where are the recommendations made for optimal approaches to meet the goals?"

**Figure 11.1: Business Analytics Pyramid**




*Source: ICFAI Research Center*

The most important characteristics of business analytics for organizations are the use of analytics to get an inside view of data and the facts behind, to implement it in strategic planning and potential decision-making by senior management. This helps them to make better decisions by accessing real-time data which was earlier accessed and used only by IT-aware knowledge workers. Some applications of Business Analytics for businesses to optimize are critical product analysis, up-selling opportunities, improved customer services, better inventory management, and competitive price insights.

**Example: Walmart uses business analytics Data Café (private cloud-based analytics hub) to Better Strategize the Products Sold Both Online and in Stores**

Walmart (the largest retailer in the world) has developed the world's largest private cloud Data Café. It is a state- of -the- art analytics hub, that can handle 2.5 petabytes of data hourly. Internal and external resources feed the Data Café. The company is in a position to analyse the sales of every store every day. As a result, Walmart is in a position to strategize sales both online and offline.

*Source: Walmart taking a giant leap towards data analytics and supply chain analytics | by Shivanee Saini | DataDrivenInvestor, January 31, 2020, Accessed on 04/05/2022.*

## 11.5 Correlation Analysis

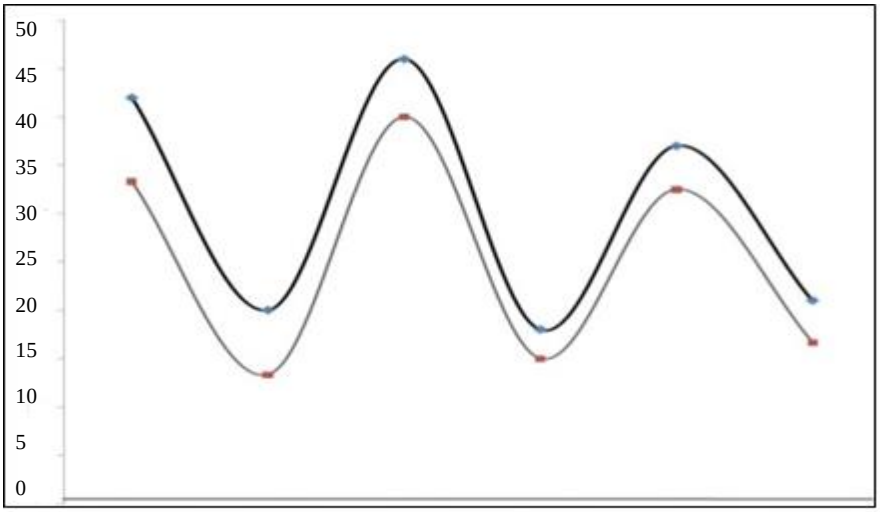
After voluminous data is gathered and stored to be analyzed for making better decisions, it is very much essential to identify if any relationship exists between the variables present in the data. This not only facilitates organizations to take better decisions, but also assists them to build an association between the variables.

Statistics gives an analytical approach in similar situations. Statistics helps in predictive analysis from the existing data. Thus it is quite useful in business intelligence.

Correlation Analysis is used as a statistical tool to discover the association between bivariate variables. It may be noted that correlation analysis is one of the most widely used statistical techniques adopted by the statisticians to find the relationship between the variables.

Many times we come across problems or situations where two variables seem to move in the same direction, either both increasing or both decreasing. At times, an increase in one variable is accompanied by a decline in another. Such changes in variables suggest that there is a certain relationship between them (see Figure 11.2).

**Figure 11.2: An Example of Graphic Correlation**




*Source: ICFAI Research Center*.

When we use correlation analysis and establish a relationship between two variables, then we confront a major question: Does this relationship indicate the existence of cause and effect relationship? A multidimensional analysis is done to know which variable is having more influence over the other.

The correlation may exist by chance, particularly when a small sample of data is involved. Also for a small sample series, no relationship may exist. It is possible that both the variables are influenced by one or more other variables or there may be another situation where both the variables may be influencing each other. This makes it difficult to establish the cause and the effect. The foregoing discussion clearly shows that correlation does not indicate causal and functional relationships. Even when there is no cause-and-effect relationship in bivariate series, if one interprets the relationship as causal, such a correlation is a spurious correlation.

Based on the data available in the organization, correlation analysis can be performed on various variables. Correlation can be positive or negative, linear or non-linear or simple, partial or multiple. The correlation coefficient closer to 1.0 indicates a strong relationship and less than 0.5 indicates a weak relationship between the two variables. The range is [-1,1]. Technically closer to -1 is also a strong correlation, but a negative correlation.

In business organizations, correlation analysis can be extremely helpful. It has been used extensively in agriculture, economics, business, and several other fields.

It can be enabled to estimate costs, sale prices and other variables on the basis of some other variables inferring closeness of the relationship with variables concerned. When a specific and reliable relationship has been established between any two given variables, we can find the value of a variable given the value of another. In fact, this is done with the help of regression analysis, which is discussed in the next section (11.6); however, regression analysis is a predictive tool that predicts the future state of the relationship between the given variables whereas correlation analysis is the indicator of the current state of the relationship between the variables taken. This also shows that the two concepts, correlation and regression analyses, are closely related with a thin line of segregation between them.Regression Analysis

In business, sometimes it is necessary to forecast in order to take a decision regarding a product or a particular course of action. To forecast, some relationship between a pair or group of variables relevant to a particular situation are to be ascertained. For example, a company wants to know how the sales will increase in the next five years, along with the growth of population and increase in demand of the product. Here, it is assumed that the increase in population will lead to an increase in sales. Thus, it is important for the company to determine the nature and extent of the relationship between these two variables.

**Example: Levi's Deploys Business Analytics (Based on Correlation Analysis) to Reinvent Denim Era**

China Sum is a partner of Levi's. China Sum used data analytics to find association between brands, people and the goods. The idea is to create more scenarios of wearing the brand`s jeans. The partner used data from CRM to build a model making use of the historical data and used predictive analytics to establish a strong correlation between brand, consumers and the goods. The model predicts purchasing potential of each target in a new campaign. They also used tag clustering to group customers with similar profiles to generate focussed personalized communication in each target group. The conversion rates went up.

*Source: Customer Story - Gridsum - Enterprise-class big data and ARTIFICIAL intelligence solution provider, 2021, Accessed on 04/05/2022*

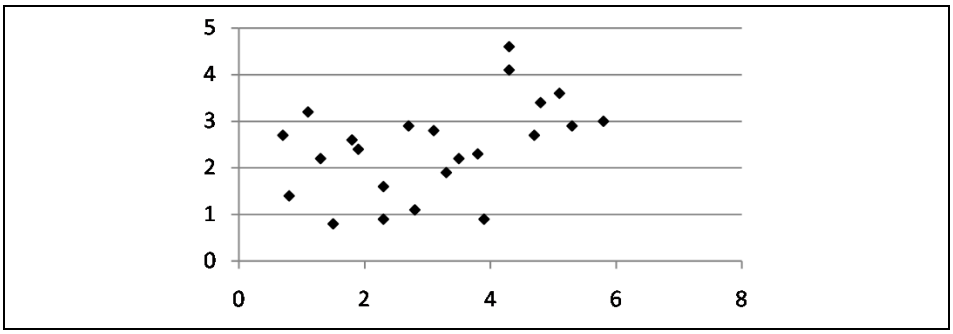
## 11.6 Regression Analysis

Regression model is defined as "a statistical model with a set of mathematical formulae and assumptions which describe a real-world situation". A statistical model tries to capture the systematic behavior of the given data, leaving out those factors that cannot be foreseen or predicted. Despite our best efforts, it is highly unlikely that a model may reveal a perfect real-world situation.

A good statistical model is one which provides as large a systematic component as possible, minimizing errors. These errors are on account of a number of factors that we are unable to identify. In case we are able to construct a good model, then the average of observed errors will be zero. These errors should also be independent of one another.

A common example in the business world is the relationship between advertising and sales. When a linear regression model involving these two variables is appropriate for prediction, we may use it for predicting sales for a given level of advertising expenditure. It may be noted that the level of advertising should be within the range of expenditure on the advertising covered. A scatter diagram can give us a broad idea of the type of relationship (or even absence of any relationship) between the two variables. When plotted on the scatter diagram, it visually gives a feeling of the type of relation between the two variables on the graph as shown in Figure 11.3.

**Figure 11.3: An Example of Scatter Diagram**




*Source: ICFAI Research Center.*

While regression analysis is an extremely useful technique for making predictions and as such frequently used, one should be careful in avoiding errors that may arise on account of the wrong application of regression analysis.

**Example: Walmart uses Regression Analysis to know which Products Customers Purchased before a Storm**

Walmart leadership team wanted to figure out what are the products customers procure before a storm. A regression analysis was done on the huge sales database of the company. The analysis showed a surprise.

*Contd….*

Ready-to-eat pastries were purchased seven times more than normal. Top-selling item was beer. The store managers in the path of a storm would store these in more quantities.

*Source: https://www.rd.com/article/item-walmart-stocks-up-on-before-storms/, May 17, 2022, Accessed on 12/09/2022*

## 11.7 Multiple Linear Regression

In the preceding sections, the discussion was confined to only two variables. However, in business, we come across several situations where the relationship is not that simple. One variable may be affected by two or more independent variables. For example, the sale of a product may be related to a number of independent variables such as price, income, advertising expenditure, seasons, number, size and location of retail outlets, quality of the product, and so forth. If in such cases, we take cognizance of only one independent variable, then the magnitude of error in the result is likely to be high. In the light of this, it is desirable to use two or more independent variables in the estimating equation. The statistical technique of extending linear regression methods, so as to consider two or more independent variables is known as multiple linear regressions. Multiple regression, as a predictive analysis, is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummycoded as appropriate).

In multiple regressions, many formulae can be used to ascertain the relationships among variables taken. Tedious calculations are involved in multiple regression analysis; to overcome the problem, computers and other software applications are used. This facilitates us enormously as several independent variables can be handled. We also can ascertain whether adding another independent variable will improve our results or not.

Multiple regression analysis is useful in as much as it shows the degrees of association between one variable taken as a dependent variable while the remaining variables, two or more, are taken as the independent variables. It also serves as a measure of goodness of fit for a given series of data.

*Constraints*: A major limitation of multiple regression analysis is that it assumes that the relationship amongst the variables is linear. However, we find in practice a large number of relationships that are not linear and follow some other pattern. Another constraint is based on the assumption that the effects of independent variables on the dependent variable are quite separate from each other and hence, additive. Also, the amount of work involved in the calculation of multiple linear regressions is enormous.

Regression Analysis, a statistical technique, is used to evaluate the relationship between two or more variables. Regression analysis helps an organisation to understand what their data points represent and use them accordingly with the help of business analytical techniques in order to do better decision-making. In this analysis, it is understood how the typical value of the dependent variable changes when one of the independent variables is varied, while the other independent variables are held fixed. Business analysts and data professionals use this powerful statistical tool for removing the unwanted variables and select the important one. For example, Organisations collect data about sales, investments, expenditures and other parameters and analyse it for improvement. The regression analysis helps the organisations to make sense of the data which is then used for gaining insights into an organisation. Business analysts and data professionals use the regression analysis to make strategic business decisions.

**Example: Tata Motors uses Multiple Regression Technique to Measure Customer Satisfaction before and after Sales Service**

Tata Motors wanted to figure out what influences its customer satisfaction level on, before, during and after sales service of Tata Motors in Puducherry. The reliability of measurement items was established using the Cronbach's α reliability test. Multiple regression analysis revealed the sought after factors. Customers were more satisfied with after sales service and during the service. But the "before sales service" was found "not satisfactory".

*Source: Measuring the Customer Satisfaction Level Before and After Sales Service Provided by TATA Motors in Pondicherry - The Research Publication (trp.org.in), 2nd July, 2022, Accessed on 12/09/2022.*

**Activity 11.1**

**Data for Business Forecasting**

A company involved in selling the garments has hired you for giving the forecast of procurement of stock to supply against the demand. What type of data will be required for the exercise and which type of analysis you would be doing to decide on the forecast?

**Answer:**

## 11.8 Logistic Regression

The crucial limitation of linear regression is that it cannot deal with discrete variables that are dichotomous and categorical. Many interesting variables in the

business are categorical in nature: for example, consumers making a decision to buy or not to buy; a product may pass or fail the quality control, etc. A range of regression techniques has been developed for analyzing data with categorical dependent variables, including logistic regression.

Since the dependent variable is dichotomous, one cannot predict a numerical value for it using logistic regression. So, regression based on 'least squares deviations' criteria (fitting curves with least error square from original data to the new analytical fitted data)for best fit approach of minimizing error around the line of best fit is not suitable. Instead, the binomial probability theory based logistic regression (Logistic regression is a technique for making predictions when the dependent variable is a dichotomy, and the independent variables are continuous and/or discrete.) is used, in which there are only two values to predict, i.e., the event/person belongs to either one group or the other. Logistic regression forms a function, based on using the maximum likelihood method (Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the *known* likelihood distribution a maximum), which maximizes the probability of classifying the observed data into the appropriate category given the regression coefficients.

Generally, logistic regression is well suited for describing and testing hypotheses about relationships between one or more categorical or continuous predictor variables and the categorical outcome variable.

There are two main uses of logistic regression:

1. First is the prediction of group membership(group membership prediction problem involves predicting whether or not a collection of instances share a certain semantic property. For instance, in a verification given a collection of images, the goal is to predict whether or not they share a {\it familial} relationship.. Since logistic regression computes the probability of success over the probability of failure, the results of the analysis are in the form of an odds ratio.
2. Logistic regression also provides knowledge and strengths of the relationships among the variables.

Many procedures in SAS/STAT like CATMOD, GENMOD, LOGISTIC, and PROBIT can be used to perform logistic regression analysis. Every procedure has a special feature that makes it useful for certain applications.

**Application areas of Logistic Regression**

Marketing: A marketing consultant wants to predict if the subsidiary of his company will make profit, loss or just break even depending on the characteristic of the subsidiary operations.

Human Resources: The HR manager of a company wants to predict the absenteeism pattern of his employees based on their individual characteristic.

Finance: A bank wants to predict if his customers would default based on the previous transactions and history.

**Example: Fintech Start-up Simpl uses Analytics Tools Including Logistic Regression for Credit Decisioning**

SIMPL (fintech company) offers small value loans instantaneously without any documentation. The company realized a 35% increase in daily transactions. Machine learning decides the loan sanction. The model takes about 100 features like "the behaviour of the user at merchants", "historical behaviour of similar users and feeds to decision trees, gradient boosting, Bernoulli Naive Bayes classifiers and simple logistic regression" to arrive at the decision.

*Source: How Simpl Paved The Way For India's New-Age Digital 'Pay Later' Market (inc42.com), 21-January-2021, Accessed on 04/05/2022.*

## Check Your Progress - 1

1. Which of the following is referred to as Data mining?
    a. Knowledge Discovery in Databases
    b. Data Cleaning
    c. Data Extraction
    d. Data Management
    e. Datamart
2. For which of the following an OLAP tool is provided?
    a. Multidimensional Analysis
    b. Slicing and dicing
    c. Roll-up and drill-down
    d. Rotation
    e. Setting up only relations
3. Which of the following is not used by Business Analytics?
    a. Statistical Tools
    b. Quantitative Techniques
    c. Predictive Modeling
    d. Operations Research Modeling
    e. Median Analysis

4. If two variables are highly correlated, what can you infer?
    a. They always go together.
    b. High values on one variable lead to low values on the other variable.
    c. There are no other variables responsible for the relationship.
    d. You cannot make any of the casual claims, nor can you be sure they always go together.
    e. Both the variables take the same values.
5. In regression analysis, which of the following variables is predicted?
    a. Response or Dependent Variable
    b. Independent Variable
    c. Intervening Variable
    d. Usually the variable itself
    e. Neither Dependent nor Independent

## 11.9 Factor Analysis

The process of inspecting, cleaning, transforming, and modeling data for discovering useful information that helps to arrive at certain conclusions and support the decision-making process is called data analysis. There are multiple approaches with different techniques for data analysis. The data analysis in statistics is divided into descriptive statistics- used to describe the basic features of the data in a study, Exploratory Data Analysis (EDA) referred as Exploratory Factor analysis (looking for clues in data), and Confirmatory Data Analysis (CDA)- evaluate evidence using traditional statistical tools such as significance, inference, and confidence.

Factor analysis is a multivariate statistical procedure that has many uses. Firstly, factor analysis cuts down a large number of variables into a smaller set of variables (also referred to as factors). Secondly, it establishes fundamental dimensions between measured variables and hidden concepts, thereby allowing the formation and improvement of theory. Thirdly, it provides construct validity evidence of self-reporting scales.

If you are a banker and you are looking at increasing the reach and market share of your bank, you can do so by carefully analyzing the consumer perceptions about banks and their expectations regarding the same. Factor analysis is an important technique of doing that and it involves the effective utilization of the behavioral patterns of different consumers as well as their demographics. India being a vast and culturally diversified country, factor analysis can be much more effectively used here.

Factor analysis is considered to be better than other statistical tools that are available in the market today for analysis purpose owing to the suitability of conditions prevailing.

Factor analysis has certain strengths as well as shortcomings:

1. Factor analysis has a high degree of replicability, which means that the experiment can yield the same kind of results even in different environments.
2. Lot of underlying factors which cannot be explicitly brought out through various statistical analyses can be achieved through factor analysis.
3. If a research is quantitative, then it has lots of components that require subjective interpretation; factor analysis is the best in doing it.

Software like SPSS using Excel and Crosstabs can be very effective in doing factor analysis. The effectiveness is clearly visible in the case of tests involving quantitative research and quantitative data interpretation in various industries. It can also be used in the field data analysis involving research in specialized areas, and also in psychological studies, like intelligence, attitude, behavior, etc. Apart from the above, factor analysis is also used in market research projects in various fields like marketing and sales.

## 11.10 Exploratory Factor Analysis (EFA)

Factor analysis is a popular collection of heuristic techniques (self-learning) used by analysts as a part of behavioral science. Exploratory Factor Analysis is a primary technique for many researchers to conduct assessment-related studies. The goal of EFA is to maximize the amount of variance explained, by identifying factors based on data. The researchers need to have specific hypotheses about how many factors will emerge, and what variables these factors will be made up of.

Factor analysis is further composed of two subsets namely “common variance” and “specific variance”. Common method variance refers to variance attached to measurement method, rather than to the constructs supposedly represented by the measures. Specific variance is not explained by common factors; instead, it may be attributed to characteristics of various individual indicators, referred to as the particular stimuli that make up a task that also affect observed scores.

Unrestricted measurement models are estimated in EFA. There is no unique set of statistical estimates for unrestricted measurement models. This property relates to the rotation phase and is part of many applications of EFA. It is assumed in EFA that the specific variance of the individual indicator is not shared between them. Procedures for EFA are available in SPSS and SAS/STAT.

**Example: A Research Study Employing Exploratory Factor Analysis Identifies Five motivating Factors for Tourists to Choose Airbnb**

Airbnb has been witnessing rapid growth with millions of passengers using its service. A study was conducted to find out what factors motivate the people to stay at Airbnb and do customer segmentation. An online survey was done with around 800 tourists who had used its services. The most important factor for customer choice was the practical attributes, and not so much in terms of experiential attributes. An exploratory factor analysis showed five motivating factors—Interaction, Home Benefits, Novelty, Sharing Economy Ethos, and Local Authenticity.

*Source: https://www.mdpi.com/2071-1050/13/13/7493/pdf?version=1625734783, 2021, Accessed on 06/09/2022.*

## 11.11 Principal Factor Analysis (PFA)

Principal Factor Analysis (PFA), also referred to as the Principal Axis Factoring (PAF) and Common Factor Analysis, aims to identify the minimum number of factors that can lead to correlation between a given set of variables, whereas the more common Principal Components Analysis (PCA), in its full form, seeks the set of factors which can account for all the common and unique (specific plus error) variance in a set of variables. PFA is generally used when the research purpose is detecting data structure (latent constructs or factors) or causal modeling.

## 11.12 Confirmatory Factor Analysis (CFA)

Like Exploratory Analysis, Confirmatory Factor Analysis (CFA) is also widely used in statistical analysis. CFA is used to test a proposed theory or model and unlike EFA, it has assumptions and expectations based on priori theory regarding the number of factors, and which factor theories or models are the best fit. The major advantage of CFA is to study the relationships between a set of observed/continuous latent variables.

Confirmatory Factor Analysis (CFA) is a restricted measurement model. That is, the researcher must explicitly specify the indicator-factor correspondence to evaluate CFA. It is a measurement model. The model considers multivariate regression to describe the relationship between a set of dependent variables and latent variables. The observed dependent variables are referred to as factor indicators and the continuous latent variables are referred to as factors. CFA is a method to specify which variables load onto which factors. Based on the goodness-of-fit of the defined model, the result is taken, or modifications are made to the originally defined structure.

**An example for factor analysis**

"Outsource2India, an outsourcing solution company, gives a good example of the use of factor analysis by a financial institution in the business of home loans.

Since there are so many options for a customer with good credit, factor analysis would comply the list of variables that determine which financial institution a customer would choose for his loan. After that list is finished, then the analysis would determine the relevant factors -- a smaller list -- that really determine choice. Once the financial institution reviews those factors, it could then proceed to market its products based on those factors".

**Example: A Research Study Conducted on Students at University of Amsterdam Using Confirmatory Analysis Identifies Protective and Risk Factors Related to Covid-19 and Mental Health Related Issues**

The Covid 19 pandemic has severely impacted the mental health of young adults. The research study focussed on changes in mental health in at-risk University of Amsterdam students which were measured before and during the pandemic. The study is aimed at exploring the intercorrelations between mental health factors, and to identify risk and protective factors. Selection of variables from the COVID questionnaire responsible for mental health problems was based on Regression Analysis. The dependent variable, a composite scale for mental health problems was derived using confirmatory factor analysis. This comprised of depression, generalized anxiety, social anxiety, social avoidance, and insomnia.

The major findings are:

(1) At the group level, depression-anxiety and loneliness increased (2) lack of emotional support was responsible for this. (3) The pandemic induced stress led to increases in depression-anxiety; (4) Loneliness aggravated this.

*Source: https://onlinelibrary.wiley.com/doi/full/10.1002/mpr.1901, 21st December 2021, Accessed on 17/05/2022.*

## 11.13 Classification

Of late, a large amount of data is being collected and maintained in databases across the business world.

There is a lot of information and knowledge that can be extracted from such databases; and with automation for extracting this information, it is possible to mine the data. There are different methodologies to tackle such problems, such as classification, association rule mining, clustering, etc.

Classification is similar to clustering. It divides customer records into distinct segments referred to as classes. But unlike clustering, a classification analysis requires that the end-user/analyst knows how classes are defined. The objective of a classifier is to decide how new records should be classified, for instance, "is a new customer likely to default on the loan?"

Classifiers use approaches such as decision trees to partition and segment records. New records can be classified by traversing the given tree from its root through branches and nodes, to a leaf representing a class. The path any record takes

through a decision tree represents a rule. For example, If "income<₹ 30,000 and age<25, and debt=High", then Default Class=Yes.

### 11.13.1 Predictive Analysis

The organizations today need to know what is happening to their business, and also be able to predict what is likely to happen. The greatest challenge in the industry is sustainability in the market with persistent growth. Strategic planning has begun to play a heavy role in companies to decide on the future evolution.

Predictive analytics makes use of a variety of statistical and analytical techniques to build models predicting future events or behaviors. The form of the predictive model varies, depending upon the behavior or event that is being predicted. Most predictive models generate a score called the "credit score"; the higher score is the highest likelihood of the event occurrence.

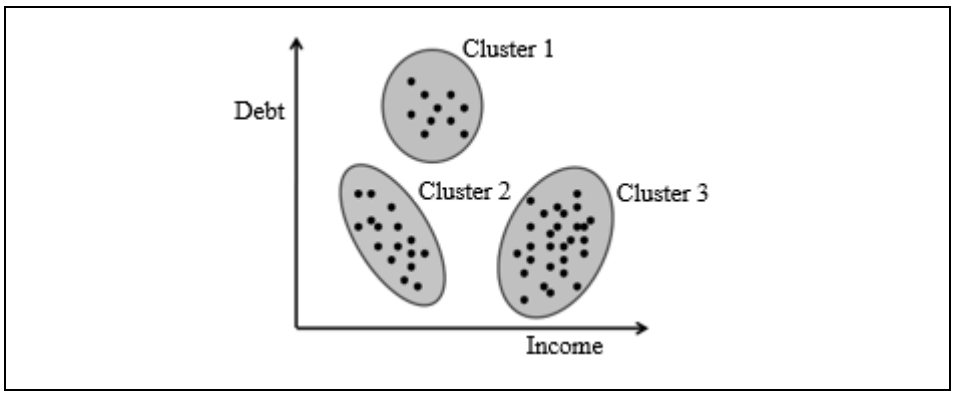
*Predictive analytics* is derived from the data mining model and focuses on predicting future possibilities and trends. Predictive analytics, along with predictive models and data mining techniques, depends on high-end statistical methods, which include multivariate analysis techniques such as advanced regression and time-series models. The insurance industry has always relied on forecasting. The use of predictive analytics has, therefore, quickly become the industry's best practice. Insurers use predictive analytics techniques to focus on potential clients and to identify potential fraudulent claims.

The applications of predictive analytics are spread over many sectors like CRM, healthcare, cross-selling, fraud detection, risk management, telecommunications, travel, etc.

### 11.13.2 Cluster Analysis

Researchers in many areas are working on how to organize observed data into meaningful structures, i.e. categorization. Clustering is the tool to group a set of objects in a way that objects in the same cluster group are more similar to one another than to those in the other groups (clusters) as shown in Figure 11.4.

**Figure 11.4: A Sample of Cluster Analysis**




*Source: ICFAI Research Center.*

Cluster Analysis is a technique of data analysis, where data or information is broken down or clustered into manageable sizes so that interpretation becomes easier. This method of data segregation makes it easy to comprehend and resolve the data issues. Cluster analysis helps in understanding the homogeneity of the data universe (like markets), and the extent of heterogeneity that exists in a datum (market). This can be effectively used in retail, pharmacy, tourism, and healthcare sectors, etc., in a more accurate and broader scale.

For example, the tourism sector uses cluster analysis for the following reasons:

1. To identify the homogeneity of the market in a vast country like India, where different heterogeneous income groups exist.
2. To package a brand communication to be delivered to each segment, so that the communication is clear, specific and unambiguous.

Hierarchical Cluster Analysis is a statistical method to find relatively homogeneous clusters based on the measured characteristics. It starts with each case as a separate cluster, i.e., there are as many clusters as there are cases. Then it combines the clusters sequentially. This reduces the number of clusters at each step. The process is done until only one cluster is left. The other methods of cluster analysis are *k*-means clustering, two-step clustering and Ward's method.

Cluster analysis is used in a variety of applications. For example it can be used to identify consumer segments, or competitive sets of products, or groups of assets whose prices co-move, or for geo-demographic segmentation, etc. In general it is often necessary to split our data into segments and perform any subsequent analysis within each segment in order to develop (potentially more refined) segment-specific insights. This may be the case even if there are no intuitively "natural" segments in our data.

### 11.13.3 Association Analysis

Using association analysis, one can identify groups of customers with similar interests and buying similar products. Using this information, recommendations can be developed that customers who purchased some book of interest also purchased other related books. This can frequently be seen on Amazon interface, where after the initial order for a book or any other article, the recommendation of similar items will start flowing in.

Association analysis identifies relationships or correlations between observations and/or between variables in the datasets. These relationships are then expressed as a collection of so-called association rules. The approach has been proved very successful in mining very large transactional databases like shopping baskets and online customer purchases. Association analysis is one of the core techniques of data mining.

For the online bookselling example, historical data is used to identify that customers who purchased two particular books also intended to purchase another particular book. The historical data might indicate that the first two books are purchased by only 0.5% of all customers. But 70% of these customers also purchase the third book. This is an interesting group of customers. As a business, we must take advantage of this observation by targeting advertising of the third book to those customers who have purchased both of the other similar books.

Association rules assist in marketing, targeted advertising, floor planning, inventory control, churning management, etc. In data mining, association rules are useful for analyzing and predicting customer's behavior.

**Example: Bank of America Saves $15 Million with Retention Analytics**

Bank of America noticed 40% attrition in call centres which affected customer satisfaction. The bank deployed data analytics using "classification analysis" to arrive at the root cause for the high level of attrition. The finding pointed out to the insight that the call centres which encouraged inter-office collaboration are faced with lower attritions. Based on this, the bank allowed all staff to take breaks together. This policy change resulted in 23% faster call handling time and 18% increase in cohesiveness. The company saved $15 million.

*Source: Advanced Analytics In HR: Applications and Examples | Blog (acuvate.com), 07-February-2022, Accessed on 05/05/2022.*

## Check Your Progress - 2

6. Which of the following is required for variables in Factor analysis?
    a. Measured at nominal level
    b. Abstract concepts
    c. Not related to each other
    d. Related to each other
    e. Standardized
7. What is the missing element in the list given here for data analysis: cleaning, transforming, and modeling data?
    a. Inspecting
    b. Collating
    c. Extracting
    d. Loading
    e. Transformation

8. If you have multiple predictor variables and a dichotomous dependent variable, which of the following is the most appropriate multivariate test?
    a. Stepwise regression
    b. Canonical correlation
    c. Logistic regression
    d. Factor analysis
    e. Predictive analysis
9. What is Predictive Analytics?
    a. Research aimed at anticipating the likely outcome of a course of action.
    b. Designed to generate insights into cause-and-effect relationships.
    c. Research that attempts to provide information on what exists.
    d. Designed to find out what happened in the past.
    e. Nothing related to decision making.

## 11.14 RFM (Recency Frequency Monetary) Analysis

RFM stands for Recency, Frequency and Monetary Value. It is a database-driven marketing technique that has been used by catalogers to increase conversion rates and reduce the expensive cost of mailing catalogs. Online retailers use RFM analysis to increase conversion rates, personalization and revenue. RMF provides answers to a number of business questions like:

- Can organizations identify their best customers?
- Do companies know who their worst customers are?
- Do companies know which customers they have lost, and which customer they are about to lose?
- Can companies identify loyal customers who buy often, but spend very little?
- Can companies target customers who are willing to spend the most at their store?

### 11.14.1 How Does RFM Analysis Work?

The goal of RFM Analysis is to divide customers based on buying behavior. One needs to understand the historical actions of individual customers for each RFM factor. Customers can be ranked based on each individual RFM factor, and finally, all these factors together are used collectively to create RFM segments for targeted marketing. The terms in RFM are:

1. R – Recency
2. F – Frequency
3. M – Monetary Value

i) R- Represents recency of the last purchase. This gives the interval between the time that the last consuming behavior happens and the present one that has taken place. The shorter the interval, the bigger the R.

ii) F- Represents frequency, which refers to the number of transactions in a particular period, for example, five times in a year, five times in one quarter or five times in one month. The more the number of times the customer purchases in a given limited period of time, the bigger the F.

iii) M- Represents monetary, which refers to the consumption of an amount of money in a particular period. The more the monetary value, the bigger the M.

It was observed that the bigger the value of Recency (R) and Frequency (F) are, the more likely the related customers are to produce a new business with enterprises. Furthermore, the bigger the monetary value (M), the more chances customers buy products or services with the same enterprise again. RFM analysis supports the Pareto axiom: "80% of business comes from 20% of your regular customers".

RFM has become an important tool that customers are consigned with a ranking of 1, 2, 3, 4, or 5 (5 being the highest) for each parameter in RFM. The three scores are together referred to as an RFM unit. Later, while doing the analysis, the database is sorted to determine which customers are having the unit ranking of '555' and are concluded as the ideal customers.

The limitation of the RFM analysis tool is that the company must be cautious while giving the ranking to the customers. They should also consider that the customers with low ranking should not be neglected, but instead should be nurtured to become improved customers.

**Example: BPI-Philam Achieves 200% Uplift in Unique Email Open Rates & 15x Improvement in Clicks by Using Recency Frequency Monetary Analysis technique**

The marketing team of BPI-Philam (Philippines based insurance company) was looking for creative ways to improve their customer engagement with emails. Personalized emails received better engagement compared to generic emails. The Recency-Frequency-Monetary methodology was used by the company to understand how users interacted emails, frequency of engagement, and the value of individual business. This should help in segmentation of customers.

*Source: BPI-Philam - Netcore Cloud, 2020, Accessed on 05/05/2022.*

## 11.15 Market Basket Analysis (MBA)

Today, many companies are trying to improve business performance with faster, better decision-making by applying advanced predictive modeling techniques to

their huge and growing volumes of data. Business analytics helps in areas like marketing, CRM, operations, with valuable insights drawn from their data.

Market Basket Analysis is a data modeling technique used to find associations between items by determining the likelihood for them to occur together. It is the concept of identifying associations between products that the customers are putting into their shopping baskets. MBA (Market Basket Analysis) is also popularly known as Product Affinity Analysis (PAA) or Association Rule Learning (ARL).

Market Basket Analysis (MBA) is one of the advanced models to leverage voluminous amounts of customer data to determine products, which are most commonly purchased together. Understanding customer's purchasing patterns helps marketing and sales organizations to make more informed decisions about how to deploy their efforts and resources. One of the classic examples of MBA is found in Amazon.com portal which shows "Customers who bought a specific item also bought allied items A, B and C".

### 11.15.1 MBA: Understanding Customer Purchase Behavior

Market Basket Analysis (MBA) is a data mining technique which is widely used in the consumer packaged goods and looks at purchase coincidence. It studies whether any two products are purchased together, and also whether the purchase of one product increases the likelihood to purchase the other.

Market Basket Analysis helps understanding customers and their purchasing behaviors by allowing companies to explore product associations. It helps in predicting the likelihood of a customer's subsequent purchase behavior based on the associations. MBA is an advanced business analytics tool that can help companies optimize marketing and sales operations for improved performance.

**Activity 11.2**

**Marker Research Process**

You are employed by a mobile manufacturing company to give an analysis of the purchase pattern of customers of their product line. It helps in launching a new product in the market. Which type of analysis will be used and how you will go ahead with this process?

**Answer:**

**Check Your Progress - 3**

10. In RFM analysis, what does F stand for?

a. Factor

b. Frequency

c. Fraudulent

d. Format

e. Fiscal

## 11.16 Summary

- Business Intelligence (BI) has become an expected business competency to improve decision-making effectiveness. It is for all workers, managers and executives to take the most effective action in a given business situation.
- Focus on business analytics has increased steadily over the past decade as evidenced by the continuously growing business analytics software market.
- Business analytics is reaching more organizations and extends to a wider range of users, from executives and number of business managers to analysts and other knowledge workers, within the organizations.
- While the main concern of database technologists was to find efficient ways of storing, retrieving and manipulating data, the main concern of the machine-learning community was to develop techniques for decoding and grasping knowledge from data.
- Many statistical tools were adopted like Correlation analysis, Regression analysis and Logistic regression to find the relationship between the decision variables.
- Other statistical methods are used in data mining for finding the underlying relationships and structures among a large set of variables. Factor Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis, Predictive Analytics, Cluster Analysis, Association Analysis, Market Basket Analysis, etc., are analytical techniques used in many applications to find the connection between variables.

## 11.17 Glossary

**Bivariate variable:** Bivariate data has two variables and involves relationships between the two variables.

**Categorical Variable:** Categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.

**Continuous Predictor Variable:** A continuous predictor variable is a continuous variable used in regression to predict another variable.

**Correlation:** Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

**Data Mining:** Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs or both. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

**Dichotomous Variable:** A dichotomous variable is one that takes on one of only two possible values when observed or measured.

**Discrete Variables:** Variables that can only take on a finite number of values are called discrete variables.

**Factor Analysis:** Factor analysis is a type of statistical procedure that is conducted to identify clusters or groups of related items (called factors) on a test. Factor analysis is used to analyze large numbers of dependent variables to detect certain aspects of the independent variables (called factors) affecting those dependent variables, without directly analyzing the independent variables.

**Linear Regression Model:** In simple linear regression, a single independent variable is used to predict the value of a dependent variable.

**MOLAP:** Multidimensional Online Analytical Processing is a kind of Online Analytical Processing (OLAP) that uses a multidimensional data model to analyze data.

**OLAP:** OLAP (Online Analytical Processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view. OLAP allows users to analyze database information from multiple database systems at one time.

**Regression:** A statistical measure that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables (known as independent variables). In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect.

**Time Series Models:** Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

## 11.18 Self-Assessment Test

1. What is Business Intelligence? What are its components?
2. What are the basic requirements for implementing business analytics? Explain.
3. How is cluster analysis different from association analysis? Explain.
4. Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables. What is the core difference between both the techniques?
5. Write short notes on the application of data mining in financial analysis.
6. Explain how Market Basket Analysis can be applied in understanding credit card purchases.

## 11.19 Suggested Readings / Reference Material

1. Rodney Heisterberg and Alakh Verma (April 2022). "Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video and Big Data Enables Competitive Advantage," Narrated by Stephen Graybill.
2. Jonathan S Walker (2021). Social Media Marketing For Beginners - How To Make Money Online: Guaranteed Strategies To Monetizing, Mastering, & Dominating Any Platform For Your Brand, JW Choices.
3. Barry Connolly (2020). Digital Trust: Social Media Strategies to Increase Trust and Engage Customers, Bloomsbury Business.
4. Seema Gupta (6 August 2020). Digital Marketing McGraw Hill; Second edition.
5. Tracy L. Tuten, Michael R (15 June 2020). Solomon et al, Social Media Marketing, SAGE Publications Pvt. Ltd; Third edition.
6. Paul Martin Thomas Erickson (2019). Social Media: Usage and Impact, Global Vision Publishing House, 2 edition.
7. Steve Randazzo (2019). Brand Experiences: Building Connections in a Digitally Cluttered World, Paipen publishing.

## 11.20 Answers to Check Your Progress Questions

**1. (a) Knowledge Discovery in Databases**

Data mining extracts the data patterns and derives knowledge from large databases using different analytical and statistical techniques for better organizational decision-making.

**2. (b) Slicing and dicing**

An OLAP (Online Analytical Processing) tool provides for slicing and dicing of the database.

**3. (d) Operations Research Modeling**

Business analytics makes extensive use of data, statistical and quantitative analysis, explanatory and predictive modeling, and fact-based management to drive decision-making.

**4. (d) You cannot make any of the casual claims, nor can you be sure they always go together.**

Correlation does not indicate any causal and functional relationship.

**5. (a) Response or Dependent Variable**

In regression analysis, the variable that is being predicted is usually the dependent variable.

**6. (d) Related to each other**

Factor analysis is often used to determine a linear relationship between variables.

**7. (a) Inspecting**

The process of inspecting, cleaning, transforming, and modeling data for discovering useful information that helps to arrive at certain conclusions and support the decision-making process is called data analysis.

**8. (c) Logistic regression**

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

**9. (a) Research aimed at anticipating the likely outcome of a course of action.**

Predictive analytics focuses on predicting future possibilities and trends.

**10. (b) Frequency**

RFM stands for Recency, Frequency and Monitory Value.

## Unit 12

# Business and Marketing Intelligence Using Analytics

**Structure**



*"The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd, and that makes it easier for humans to do what they are good at — explain those anomalies."*

- Daniel Gruhl, IBM Researcher

## 12.1 Introduction

New generation tools of machine learning and deep learning can analyse rapidly growing data and provide insights to the decision makers. Management needs to evaluate these insights in the context of company vision and strategic plan and take decisions which will optimize outcomes for all the stakeholders.

In the previous unit, we have studied the methods used in analytics at length. Here, in this unit, we will visit the application part of the analytics. We will specially refer to business and marketing intelligence in this unit, as it is well associated with analytics. Business Intelligence (BI) can be used in different industries such as airline, retail, manufacturing, financial services, healthcare, bioinformatics, and hospitality industry. The current day business intelligence systems are replacing DSS (Decision Support Systems), MIS (Management Information Systems) and EIS (Executive Information Systems). Organizations such as Tesco, Capital One, CEMEX, and Netflix have made better decisions based on business intelligence.

In the current unit, the need for business intelligence and the definition of business intelligence are explained. The distinction between data, information, knowledge, and wisdom are explained. Data warehouse, business intelligence architecture, business intelligence components, business intelligence methodologies, data mining techniques, and business intelligence tools are described at length. The usage of business intelligence in market knowledge collection and its applicability to decision making is highlighted. Data in organizations is growing much faster than the computing speed in the world. Hence, the importance of big data, Hadoop and big data analytics are also explained in the unit.

## 12.2 Objectives

After going through this unit, you will be able to:

- Define different components of business intelligence
- Explain the business intelligence architecture
- Define data mining techniques used in business intelligence
- Explain the application of business intelligence in market intelligence and decision-making
- Discuss the utility of various business intelligence tools commercially available in the market
- Define big data architecture and Hadoop

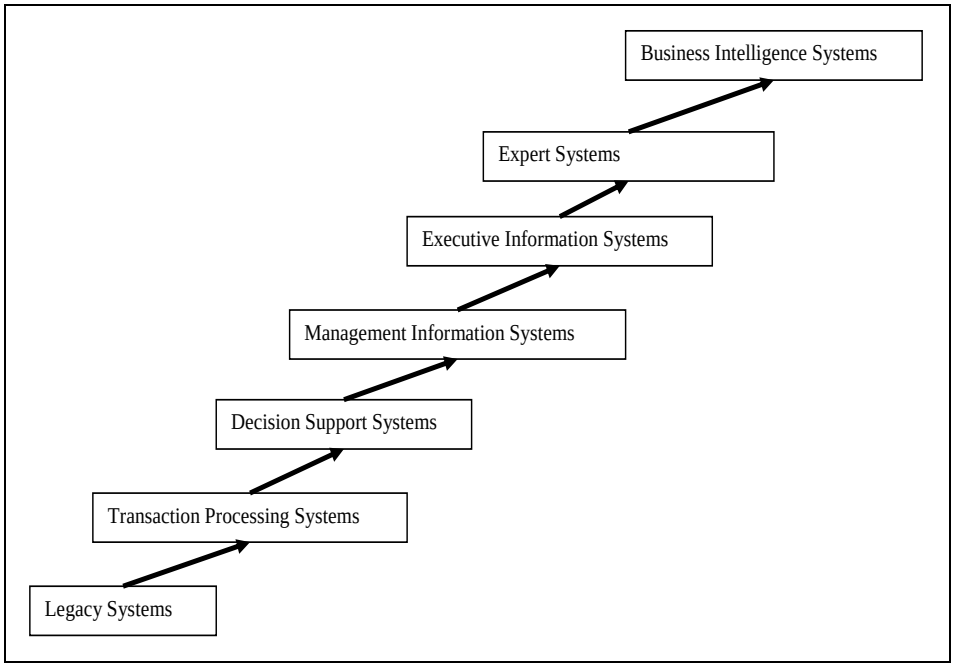
## 12.3 Need for Business Intelligence

What is the need and application of business intelligence systems in an organization? In the absence of data for decision-making in an organization, it would be just guessing, instead of judging the current status of performance of a business. Hence, the organizations looking for performance improvements prefer

to go for business intelligence. The organizations initially have to think: why do they require business intelligence application? If they decide to have business intelligence, then they should think: which stakeholders get benefited using business intelligence tools and what are the investments for the acquisition of such applications? Business intelligence is required for effective decision-making, operational and strategic excellence.

The evolution of information systems over a period of time is shown in Figure 12.1. Legacy systems used complex algorithms, being developed using procedural and functional programming languages and were mainly used in scientific computing.. Transaction Processing Systems (TPS) used some form of data and file-based systems and these were used for business purposes. Decision Support Systems (DSS) used data models and interfaces for the users. Management Information Systems (MIS) used relational databases and business logic. Executive Information Systems (EIS) were used in management reporting and data visualization. Expert Systems (ES) were rule-based and search knowledge bases for analysis. The current day Business Intelligence Systems (BI) use data mining techniques, data warehouses and business analytics tools useful for managerial decision-making.

**Figure 12.1: Evolution of Information Systems**




*Source: ICFAI Research Center*

### 12.3.1 Overview of Business Intelligence

Business intelligence is the application of methodologies, processes and technology in acquiring, integrating, storing, accessing, analyzing, and interpreting the data to make enterprise level decisions.

Business intelligence is about extracting the needed information and transforming that information into knowledge. Business intelligence extracts large amounts of data, analyzes it and generates reports needed for daily decision-making. Senior management and top management can be benefited from the insights and reports generated using business intelligence. Business intelligence technologies support efficient business operations. Business intelligence uses the technologies such as data warehouses, data mining tools, OLAP (Online Analytical Processing) tools, web services, XML, J2EE, and .Net. Business intelligence includes several software tools for extraction, transformation, load, querying, visualization, and reporting.

Business intelligence is different from competitive intelligence. Competitive Intelligence (CI) concentrates only on the external factors of the organization, whereas business intelligence considers internal factors such as operational details of the organization as well. Business intelligence capabilities include data mining, online analytical processing, decision support system, forecasting, and statistical analysis.

Business intelligence facilitates effective communication in an organization. The organizations can change their strategies and decisions based on the changing economic conditions, customer preferences, product sales, financial situation, and supply chain operations using business intelligence. Using business intelligence, the organizations can find who their loyal customers, most profitable customers and potential customers are. One can also find out the reasons for customer loyalty using business intelligence. Business intelligence enables us to identify the business trends, anomalies, obtain insights, and run simulations.

**Example: HDFC Bank Deploys Business Intelligence Using SAS Solution to Lower Credit Risk**

HDFC Bank was facing challenges in administering credit-underwriting policies. Thousands of credit decisions are taken daily in the bank. Around 1,000 applications are processed daily by bank officers. Majority customers are looking for an instantaneous loan decision. The bank uses SAS technology to sanction instant loans. Using SAS, HDFC Bank accesses credit-bureau report, to validate the customer's identity. The application is processed using several s models – such as probability to default.

*Source: Improve cross-selling capabilities | SAS., 2022, Accessed on 05/05/2022.*

**Activity 12.1**

**Business Intelligence: Retain Chain**

A retail chain is operating in different locations in India. It collects data of products being sold, who purchased them, along with the customer profiles.

| |
|---|
| The organization is currently collecting data based on customer surveys. But this process of finding results is taking a lot of time. The organization would like to know who their profitable customers are and what products are being mostly sold. The organization currently maintains a relational database. However, the top management needs a quick solution every time. What would you like to suggest to the organization? Which technologies, processes and approaches will solve their problem?<br>**Answer:** |
| |
| |
| |

## 12.4 Data, Information, Knowledge and Wisdom

The human brain contains four types of data. They are raw data, information, knowledge, and wisdom. The journey of data from information to knowledge and to wisdom is shown in Figure 12.2.

**Data:** Raw data is the figures and numbers. Data alone cannot make any sense. It cannot give any meaning to the individuals. It has to be processed following certain rules to understand. The data in a spreadsheet or a flat file is an example of raw data. Data is frequently shared between the organization and other stakeholders of the organization such as customers, suppliers and partners. The characteristics of good quality data include completeness, correctness, timeliness, and consistency.

**Information** is processed data. Status reports, trend reports and progress reports in the organization are processed data which gives information to the executives. The data tables with column names and row values in Relational Database Management Systems (RDBMS) provide useful information.

After reading the information, the individual understands it, interprets it and stores it. That becomes knowledge.

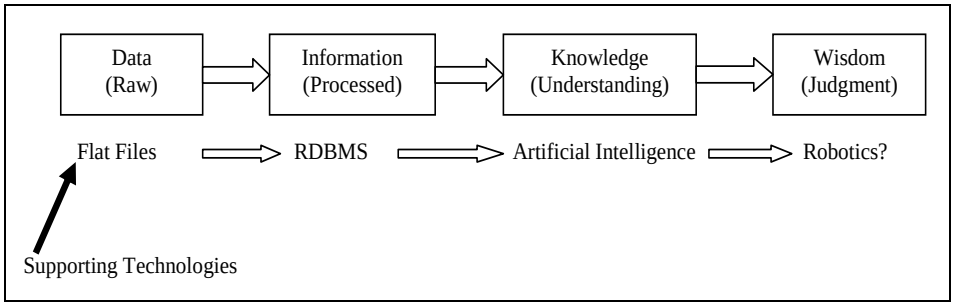
**Knowledge** is the information well understood. Further, the individual does not leave the understood information there itself. He applies his judgment, values and ethics into it and makes it wisdom. For example, remembering the number of defects per release of a software product, the number of test cases written per release of the product and making decisions based on that information is the knowledge of the product. This remembered knowledge is useful in the future journey of the product. Similarly, remembering the normal temperature of a human being and temperature of boiling water, etc., are examples of knowledge. Human beings remember this knowledge and apply whenever a need occurs.

Artificial intelligence systems take decisions based on the compiled knowledge and logic. Expert systems search the knowledge bases.

**Wisdom** state represents the judgmental level of knowledge. Based on this, the individual takes decisions in the organization and decides whether it is right or wrong, acceptable or unacceptable, and ethical or unethical. Wisdom also involves future thinking and vision. As data transforms into information, knowledge and wisdom, the level of understanding increases in the individual.

Note: It is to be noted here that intelligence is an application of knowledge. Thus, a book contains information but the reader converts the concepts explained to intelligence and applying this knowledge contents to real life problems.

**Figure 12.2: Evolution of Knowledge**




*Source: ICFAI Research Center*

Wisdom comes through systematic practice. Machines may not reach this level because judgment has to happen based on the facts, data, ethics, vision, values, and culture. Inculcating or embedding ethics, culture and values into machines is still the grey area. Robots work based on the knowledge fed to it beforehand and artificial intelligence. Artificial Intelligence (AI) is the process of application of knowledge supplied to it. AI work as rule-based systems. AI systems do not have the wisdom as a human being has. Thus, Business Intelligence (BI) systems provide knowledge and information useful for decision-making in the organization. Business intelligence system is becoming part of knowledge management practices of the organization as well. They are useful not only for business decision-making but also for knowledge management in the organization.

**Activity 12.2**

**Data, Information and Knowledge**

A pharmaceutical company would like to improve the sales of its different products in rural areas. For this, their marketing strategy team recommended developing an Expert System (ES) which can be deployed for this purpose. Using the proposed expert system, without a doctor's physical intervention,

| |
|---|
| prescriptions can be generated by supplying the symptoms of the disease as input. The IT department is given the job of developing such an expert system. The project manager wonders whether to collect data, information or knowledge. Suggest how the project manager can go ahead with this project.<br>**Answer:** |
| |
| |
| |

## Check Your Progress - 1

1. For which of the following Business intelligence is required?
    a. Organizational Performance
    b. Decision-Making
    c. Organizational performance and decision-making
    d. Loss of Productivity
    e. Knowledge
2. Business intelligence systems make use of which of the following?
    a. Data Warehouse
    b. Data Marts
    c. Data warehouse and data marts
    d. Data Loss
    e. Raw Data
3. What is information?
    a. Useful data
    b. Collected data
    c. Processed data
    d. Data about people
    e. Text without data
4. What does DSS stand for?
    a. Demand Supply Support
    b. Divided Supply Systems
    c. Direct System Source
    d. Decision Support System
    e. Decision Supply System

5. On which of the following Knowledge depends?
   a. Understanding
   b. Withstanding
   c. Outstanding
   d. Application
   e. Data management

## 12.5 Data Warehousing

Data warehouse is the major component of business intelligence. It helps in the propagation of data in an organization. It extracts, cleanses, integrates, transforms, and stores the data and transmits it for query processing and analysis as and when required. The sources of data for data warehouse can be the internal enterprise systems, operational databases, relational databases, spreadsheets, historical databases, unstructured data, from point of sale terminals, or the data from the Internet or emails. It integrates the data required for organization's strategic, tactical and operational planning and decision-making. Data received can be in any of the following two forms: Structured and Unstructured Data.

i) Structured data can be from the organization's relational databases such as tables, forms and spreadsheets. It is the data which can fit into an organizational relational database. Structured data is relatively easy to search.

ii) Unstructured data can be email messages, charts, graphs, memos, movies, images, telephone conversations, letters, news items, marketing flyers, presentations, spreadsheet files, web pages, white papers, discussion forum messages, pictures, biometrics (fingerprints, facial images), plain text files, audio and video files, etc. Some researchers have used the term semi-structured data to mean unstructured data. It is the data which cannot fit into a relational database or structured data. It cannot be represented in rows and columns. Semi-structured data analysis requires classification and taxonomy. It contains the important information needed for organizational decision-making. Data warehouse consolidates the data collected from various enterprise systems and the external data.

## 12.6 Business Intelligence Components

The essential components of business intelligence systems include data warehouse, data marts, corporate performance management systems, ETL tools, OLAP, analytical tools, data visualization, data mining, geographic information system, and a well-defined workflow. Data warehouse is the important component of business intelligence. However, it should be a real-time data warehouse.

Data mart is an organized collection of data specific to given departments. It is a subset of a data warehouse. That is, a data mart is formed by extracting data from a data warehouse based on the department, specific business function, business process or business unit. This is helpful in making decisions specific to that department. For example, there can be different data marts for marketing, sales, finance, operations, and HR. Each data mart is useful for efficient decision-making for that department. There can be multiple data marts in one enterprise. Each data mart is formed to achieve operational excellence through decision-making. Functional executives can take decisions based on data extracted from the data mart. Virtual data marts can also be created using database 'views'. Cubes are to be created from a data mart.

Corporate Performance Measurement can be done using organizational web portals, dashboards and scorecards. Key performance indicators (KPI) are also components of business intelligence. KPIs are the metrics collected weekly, monthly, quarterly, and yearly in the organization. Extract, Transform, and Load (ETL) tools are also components of business intelligence. Current day ETL tools extract the data very quickly.

**Activity 12.3**

**Business Intelligence Components**

A manufacturing company has many locations with location-specific databases. The organization would like to consolidate all the data and would like to have consolidated reports. For that purpose, the company decided to go for a centralized data warehouse with business intelligence capabilities. The business intelligence objective is to acquire data, organize data and analyze data. In that direction, the project manager thinks of how to acquire data from different sources. Suggest to the project manager how he can extract data from different data sources. Suggest certain commercially available tools for this purpose.

**Answer:**

### 12.6.1 OLAP

OLAP (Online Analytical Processing or OLAP server) provides multi-dimensional views, analyzing, visualizing, reporting and modeling the data. They can be used to optimize business operations. They work with data warehouses and data marts in course of accessing the data. They process queries which are

needed to find the trends in the organization. Current day OLAP tools access the data and generate reports very quickly. OLAP tools take 0.1% of the time that a traditional relational database system takes for answering a query. Popular OLAP tool vendors include Cognos and Business Objects.

Analytics tools do the statistical analysis needed for forecasting, data mining and predictive analysis. They predict or provide insights based on certain facts for the organization. The business intelligence components include business process model, business function model, business data model, metadata repository, and application inventory.

**Example: Bank Al Etihad (A Jordanian Bank) Implements Data Warehouse Through Data Marts to Reduce Load on Transaction Systems and Pave Way for Future Integration**

Bank al Etihad – (a Jordanian bank) planned to establish a Data warehouse based on solid architecture. The bank badly needed validation of the data from various reports, clarification on new reporting needs, and enhance storage architecture towards achieving a consistent workflow. The idea was to develop basic analytic and reporting dashboards. The company took the help of a partner for establishing the data warehouse which analyses data from 15 sources.

*Source: Top 7 Data Warehouse Solutions | DICEUS, 9th September, 2021, Accessed on 06/05/2022.*

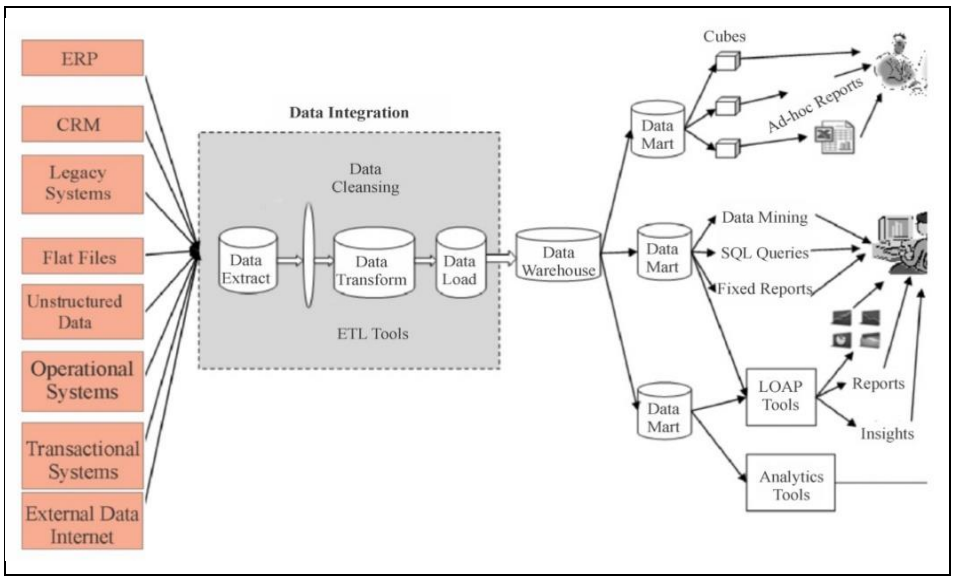
## 12.7 Business Intelligence Architecture

The objective of business intelligence (BI) systems is to provide quality and timely input data for decision-making in an organization. They provide information on demand. They combine the operational systems and data with analytical tools in order to provide complete information for decision-making and planning in the organization. Business intelligence architecture consists of the data sources, data integration, data storage, data management, operational processes, presentation tools and applications, querying, and reporting as shown in Figure 12.3. This architecture should fit into the enterprise system's architecture. Business intelligence architecture is also part of the enterprise system's architecture. It is not an isolated entity in the organization. It is but a part of the enterprise network of systems.

The data sources for business intelligence are heterogeneous. It includes both internal and external data sources. Internal sources of data for business intelligence include Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), legacy systems, flat files, operational systems, decision support systems, executive information systems, knowledge management systems, OLAP, visualization systems, transaction systems, and geographical

information systems. The unstructured data used are messages, video, audio and external sources of data include the Internet, e-mails, blogs, social networking sites, and media. Data warehouse pulls the data from all these sources. The data type can be structured or unstructured data.

**Figure 12.3: Business Intelligence Architecture**




*Source: ICFAI Research Center*

**Data integration** is done through data extraction from different sources, data cleansing, data transformation into a required format, and then data loading into the data warehouse. Here ETL tools known as Extract, Transform and Load tools can be used for data integration purpose. There are many commercially available ETL tools in the market. Some of the ETL tool vendors include Informatica, Trillium, Ascential, and Ab Initio.

**Data warehouse** provides the access, storage and integration to the data. The data from the data warehouse is loaded into data marts specific to the business function. Data mart is the tiny database specific to a department, business unit, business process, or business function. The advantage with data mart is – it provides quick access to data for specific purposes of the group. There can be multiple data marts in the organization. Data marts can also be used for SQL querying, fixed report generation and data mining purposes. There are different data mining techniques which can be applied on data marts. The outputs are pulled from the data marts.

**Cubes** are derived from data marts. A cube can be a logical view of the data. It provides structured information to the users. It is useful for querying and reporting purposes. The developers can derive multiple cubes from a single data mart. The developers and users access the data cubes. Data cubes can be used to generate ad-hoc reports.

**OLAP** (Online Analytical Processing) tools and other analytical tools can access the data marts in the business intelligence framework. These tools apply statistical techniques and derive insights and findings useful for managers. They can also be used to generate graphs and reports from the data. They can even generate trend reports, progress reports and status reports.

Overall business intelligence makes use of both internal and external data, analyzes it and prepares reports, graphs, insights, and knowledge useful for decision-making at different levels in an organization. Business intelligence systems should be transparent, reliable, accessible, and secure. They should be able to handle different data types, data formats and data sources. Business dominates the technology in business intelligence architecture. Business intelligence architecture also includes the metadata, standards, business rules, and policies. Technical architecture consists of hardware, database management systems and middleware. The security and scalability of the business intelligence systems are also to be taken into consideration while architecting business intelligence for the organization. Business intelligence architectures should comply with the regulatory requirement of the Sarbanes-Oxley Act of 2002.

Metadata repository contains details about the source of data, bibliographic information, data definition, and how it is processed. It also contains details about the reliability and accuracy of data.

**Example: Pfizer uses a Novel 'Incubation Sandbox' to Speed Up Data Analysis in its COVID-19 Vaccine Trial**

Pfizer scientists used an AI tool in their journey of developing COVID-19 vaccine at record speed. Normally, more than 30 days are consumed cleaning up patient data after a trail before the analysis can start. Manual inspection of data by the scientists for coding errors and other discrepancies was the past norm. A machine learning tool ( Smart Data Query (SDQ) could reduce this to 22 hours.

*Source: How a Novel 'Incubation Sandbox' Helped Speed Up Data Analysis in Pfizer's COVID-19 Vaccine Trial | Pfizer, 2022, Accessed on 06/05/2022.*

## Check Your Progress - 2

6. Which of the following can be Data?
    a. Structured data
    b. Unstructured data
    c. Structured or unstructured data
    d. Error data
    e. Scattered data

7. Which of the following is included in data warehouse functionality?
    a. Cleansing data
    b. Storing data
    c. Transforming data
    d. Restructuring data
    e. Extracting data
8. Which of the following is included in business intelligence architecture?
    a. Data warehouse
    b. ETL tools
    c. Data marts
    d. Data warehouse, Data marts, ETL tools
    e. Software
9. Which of the following should be true for Business intelligence architecture?
    a. Scalable
    b. Secure
    c. Scalable, secure
    d. Unreliable
    e. Loosely coupled
10. What does ETL stand for?
    a. Extract, Transform and Load
    b. Enter, Transfer and Leave
    c. Early Transfer Level
    d. Electronic Transfer and Leave
    e. Execute, Translate and Leave

## 12.8 Business Intelligence Methodologies

Business intelligence can be viewed as an application of data mining technique, usage of complex algorithms and statistical analysis on data. Business intelligence methodologies include predictive analysis, statistical analysis, reporting, and ad-hoc analysis. Business intelligence involves the detailed analysis of huge data, application of technologies and analysis practices.

The methodology to deal with structured and semi-structured data in business intelligence is to acquire the data, clean up the data and integrate the data. Then, search the data, analyze the data, identify trends, changes and incorrectness, and deliver the results. The management action is based on the provided information and results. The reports generated can be validated, structured and summarized.

For structured data, analysts use ETL tools, data warehouse, OLAP, and data mining. Semi-structured data requires different and less sophisticated tools to analyze. Semi-structured data can be gathered from business processes and news items.

Complex Analysis makes use of fast and user-friendly OLAP queries. OLAP queries are used in marketing, financial reporting, business process management, sales, and forecasting. The OLAP analyst traverses through the data warehouse, data marts and changes the data orientation. The operations possible in OLAP include slice and dice — a capability to combine and recombine different combinations of data; drill down/up — to navigate through data, pivot which changes the dimensions from rows to columns and vice-versa; nesting — displays one dimension inside another dimension of data.

Business intelligence analytical techniques include modeling, visualization, embedding, monitoring, reporting, data mining, scorecards and investigating.

## 12.9 Data Mining Techniques

Data mining is the process of finding patterns in huge data using statistical techniques, artificial intelligence and Database Management System (DBMS). Data mining can be used in fraud detection, marketing and supervision, etc. It finds correlations, patterns and trends in data from data warehouse using statistical and mathematical techniques. Data mining can be used for hypotheses proving and knowledge creation.

The main objective of data mining is to find the earlier undetected patterns in large data sets (of the organization). Data mining techniques include classification, multidimensional analysis, correlation, regression, associations, prediction, clustering, time series analysis, and outlier analysis. Exploratory data analysis and sequential pattern analysis are other data mining techniques.

Classification determines the characteristics of a particular group. Each group characteristics can be used to design a model. Clustering creates the group of observations having certain common characteristics. Time series analysis finds the associations based on time. Association finds the relationship between events. Correlation finds the relationship between two different variables or events. Regression finds the impact of one event on other event. Regression is of two types such as linear regression and non-linear regression. Prediction finds the future values based on huge data sets.

Multidimensional analysis requires a multidimensional database. Multidimensional analysis can be done on three-dimensional cubes. A Cube can also be called as a multidimensional database. Cubes are useful to generate ad-hoc reports and ad-hoc queries. The multidimensional setup with cubes should support pivot analysis useful for generating ad-hoc reports. An example of a three-dimensional cube consists of Products, Customers and Time of product purchased. Master data is used to generate different dimensions in the database.

Master data is important for business intelligence since it can be drilled to facilitate analysis and creation of various other data and information. Examples of master data include Customer Master File, Product Master File, Supplier Master File, etc.

Fixed reports can also be generated using multidimensional cubes. However, this process does not use standard SQL (Structured Query Language), but it uses a multidimensional language known as Multi-Dimensional Expressions (MDE). Hence, the best practice to generate fixed reports is to use SQL commands directly from data warehouse and use cubes for multidimensional reports. Data mining tools are based on artificial intelligence, statistical and mathematical techniques, decision trees, neural networks, and Bayesian network theory. Commercially available data mining tools include IBM Intelligent Miner, SAS Enterprise Miner, DBMiner, R, SGI Mine Set, and MS SQL Server. There are some text mining tools as well.

Data mining is used to construct six types of models which are intended to solve business problems. They are classification, regression, time series, clustering, association analysis, and sequence discovery. The first two, classification and regression, are used to make predictions, while association and sequence discovery are used to define behavior. Clustering can be used for either forecasting or classification.

Companies of various sectors can gain a competitive edge in the era of globalization and digitalization by mining their growing databases for valuable, detailed transaction information. Examples of such uses are explained below.

Each of the four applications below makes use of the first two activities of data mining: discovery and predictive modeling. The discovery process, though not cited clearly in the examples (except in the retail description), is used in customer segmentation. This is done through conditional logic, analysis of affinities and associations, and trends and variations. Each of the application types defined below leads to a kind of predictive modeling. Each business is interested in predicting the behavior of its customers through the knowledge gained in data mining.

**Retail**

The transactions happened in a store and the records of all the products purchased are recorded using store-branded credit cards and point-of-sale systems. This allows them to understand their several customer segments efficiently. Some of the retail applications are as follows:

- Performing Market basket analysis—Also known as affinity analysis, basket analysis discloses which are the combination of items purchased by the customers. The knowledge of the basket of goods purchased by a customer helps the stores to improve stocking, store layout strategies, and promotions.

- Sales forecasting—Examining time-based patterns helps retailers make stocking decisions and maintaining the inventory levels in the stores and warehouses. The stores will also be able to predict the most likely time of purchase of matching items if the items are purchased today.
- Database marketing—Retailers can develop profiles of customers with certain behaviors, for example, those who purchase designer labels clothing or those who attend sales. This information can be used to focus cost–effective promotions.
- Merchandise planning and allocation—When retailers add new stores, they can improve merchandise planning and allocation by understanding the patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

**Banking**

Banks can utilize knowledge discovery for various applications, which are as follows:

- Card marketing—By identifying customer segments, the card departments and acquirers can improve profitability with more effective achievement and retaining programs, targeted product development, and customized pricing.
- Cardholder pricing and profitability—The cards department can plan for risk-based pricing of their products in order to maximize profit and minimize the loss of cutomers using the data mining techniques. Includes risk-based pricing.
- Fraud detection—based on the past repayment behavior of the customer, the banks can identify the fraudulent patterns which is extremely costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns.
- Predictive life-cycle management—Data mining helps banks predict each customer's lifetime value and to provide service each section of the customers suitably such as offering special deals and discount.

**Telecommunications**

Telecommunication companies around the world face mounting competition which is compelling them to aggressively market special pricing programs in order to retain the existing customers and attracting new ones. Knowledge discovery in telecommunications include the following:

- Call detail record analysis—Telecommunication companies gather detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.

- Customer loyalty—Some customers repeatedly switch providers, or "churn", to take advantage of attractive incentives by competing companies. The companies can use data mining to identify the characteristics of customers who are likely to remain loyal, whoa are going to switch, thus enabling the companies to target their spending on customers who will produce the most profit.

**Other applications**

Knowledge discovery applications are emerging in a variety of industries. They are as follows:

- Customer segmentation—All industries can take advantage of data mining to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.
- Manufacturing—Based on the demands and expectatons of the customers, manufacturers are commencing to provide customized products for customers which suit their requirements. Therefore, they must be able to predict which features should be bundled to meet customer demand.
- Warranties—Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- Frequent flier incentives—Airlines can identify groups of customers that can be given incentives to fly more. Based on the data, the Airlines can provide even combo offers to their customers based on the information of the mode of booking their tickets.

In the application examples discussed above, the use of forensic analysis was not as common. The banking example is the only one that was looking for deviations in the data. Banks and other financial institutions use data mining for fraud detection, which was not referred to in the other examples even though there are similar uses of deviation detection in the other industries.

Relationship management has not been left out of the disruption taking place across industries. The adoption of advanced technology such as data warehousing and data mining technologies has become a big part of the commotion. Firms and corporate organizations are taking advantage of these technologies to develop a competitive edge in the marketplace. They achieve this by gaining expertise in extracting hidden predictive information from a large dataset using advanced statistical & mathematical algorithms.

Data mining can support the identification of valuable customer understandings, predict future behaviours of customers, and help make data-driven decisions. Data mining has improved over the years and can be described as matured enough to deliver unprecedented benefits to organizations. Companies can achieve

automated, future-oriented investigations, which can help them gain competitive advantage. An automated Customer Relationship Management is now imaginable to achieve with the usage of Data Mining techniques. Hence we can say that with the advent of data minning business questions that were time-consuming in the past are now very much achievable.

There are different techniques and approaches to data mining. However, an organization should utilize the method and strategy that will help in achieving its overall objective of establishing a productive and successful customer relationship management.

**Check Your Progress - 3**

11. What does OLAP stand for?
    a. Online Application Processing
    b. Online Application Performing
    c. Online Analytical Processing
    d. Offline Analytical Procedure
    e. On Load Analytical Procedure
12. What is data mart?
    a. Sub-set of data warehouse
    b. Specific to business function
    c. Specific to department
    d. Specific to department, business function and subset of data warehouse
    e. Total database
13. From which of the following Cubes are derived?
    a. Data Fields
    b. Data Marts
    c. Data Entries
    d. Data Summaries
    e. Storage Blocks
14. Which of the following is included in Business intelligence methodologies?
    a. Predictive Analysis
    b. Statistical Analysis
    c. Ad-hoc Analysis
    d. Predictive, statistical and Ad-hoc analysis
    e. Risk Analysis

15. Which of the following is not a data mining technique?
    a. Classification
    b. Multidimensional analysis
    c. Clustering
    d. Cloning
    e. Factoring

## 12.10 Market Intelligence and Decision Making

Business intelligence implementation in an organization involves hardware, software, human resources, and costs of implementation. Training is also required for analysts to use the system. System upgrades also initiate training need in the organization. Business intelligence can be used in deriving competitive intelligence for the organization. Competitive intelligence is about gathering and analyzing external information useful for devising organizational plans, strategies, operations, and decisions. Market intelligence includes competitive intelligence, competitive strategies, pricing strategies, sales strategies and competitive advantages.

The sources of market intelligence include government websites and portals, online databases, government publications and reports, online databases, surveys, trade associations' periodicals and reports, user groups, consumer groups, industry bodies, industry consortiums, competitors, suppliers, vendors, partners, customers, distributors, interviews with industry experts, journals, newspapers, magazines, financial reports and private sector organizations.

The data collected from the above mentioned sources is to be fed into the organizational business intelligence system to gain market intelligence in the industry. The output reports, graphs, knowledge and information of business intelligence are useful in organizational decision-making.

What is Market Intelligence?

In order to make Strategic planning, the top management need to know the best usage of reliable information for decision making. Even in this scenario, there are no structured processes to collect this information and analyze marketing information.

The dependence on this type of information is grown which is also known as Competitive as the access to information has become easier and technology has facilitated data generation and analysis.

The most common mistake is to think that Market Intelligence comes down to the analysis of competitors. Although this really is one of the main focuses, there is much more. The goal is to understand what goes on outside the company in order to increase competitiveness. Collecting all this intelligence can help support decision making.

So, in rder to understand the increased competitiveness, the companies are relying on Market intelligence which involves the business environment in which the company fits and everything that can impact performance and development. That includes product/service, customers, competitors, the market segment and the economy. It also depends in understanding the following so that the companies can gain competitive advantage.

- Know how the product/service is seen by customers and the market;
- Assess whether the price charged is consistent with the market;
- Identify product differentiators;
- Figure out how consumers perceive the brand;
- Have in depth knowledge of current and potential customers;
- Understand competitors including their strategies and growth;
- Understand the market segment including problems, trends, players and influencers.

For this, market intelligence uses tools such as benchmarking, SWOT analysis, mystery shopping, media monitoring, and, especially, data analysis. The intelligence generated through this helps to make the company more competitive.

**Example: Bank of America Uses AI based Market Intelligence to Identify Investors for IPOs of its Corporate Customers**

Bank of America has created a Predictive Intelligence Analytics Machine - PRIA which uses a set of supervised machine learning algorithms to understand relationship trends between ECM deals and investors. Some 15 crore data points were used to train the model. The data included information from around 50,000 past ECM deals, investor data, and market data. The model can process around 1000 investors in seconds, with an 80% accuracy.

*Source: Bank of America brings AI to equity capital markets (cio.com), March 2, 2020, Accessed on 06/05/2022.*

## 12.11 Make the Last Mile in Data Analytics

The 'last mile' is the group of people who deliver the results of the data analysis. The group gives this result to the business so that they can easily understand the trend. The 'last mile' group has expertise in data analytics and knows enough about the business. It requires experience in data analytics and also the confidence to present the results to the CEO. The 'last mile' group can handle big issues and help in developing and guiding business strategies.

For example, Yahoo Mail: 250+ million people are using yahoo mail. When people signup for Yahoo Mail account, they see the news preview module first.

The news preview module has become popular because it helps in retaining active users. Data analytics noticed that new users like to read the news when they read email. By adding a news preview window, Yahoo mail was able to increase the return rate by 40%. In addition to new users, the other users also liked reading news while looking at their email.

### 12.11.1 Geospatial Intelligence

Geospatial Intelligence means using data related to space and time to improve the quality of predictive analysis.

For example, smartphone: the smartphone helps to look at traffic and it shows streets in red and yellow color. It observes the average speed of travel and calculates the aggregate speed to travel and then it helps us in avoiding traffic. Therefore, geospatial analytics has become a standard part of life today.

For advertisers, geographical intelligence helps in a different way. It makes the users feel ads less like spam and more like information. The following are the examples of geospatial intelligence:

- Healthcare organizations will be able to predict movements of disease outbreaks over time and adequately prepare for potential epidemics before they occur.
- Police departments can study past geospatial data to see where crimes occurred frequently and understand where and when future crimes are most likely to happen.
- Insurers can incorporate geospatial information into their risk calculations to optimize pricing for known risk factor.

### 12.11.2 Consumption of Analytics

Consumption of analytics means making analytics consumable in an organization. There are different stages in the consumption of analytics. They are:

i) Communication

ii) Implementation

iii) Measurement

iv) Align incentives

v) Develop cognitive repairs

**i) Communication**

In the first stage, the business analytics from the core team will be sent to the wider group of decision makers and the daily consumers of analytics in your organization. It helps the team to create a platform for analytics in an organization.

**ii) Implementation**

Implementation means to get all the ingredients in place to consume analytics successfully. Strong leadership can be the most important trigger in adapting analytics in an organization.

**iii) Measurement**

Measurement means testing of consumption. It uses analytics to test itself. A successful business decision can be taken only with a combination of business experience and analytics.

**iv) Align incentives**

Successful consumption of analytics leads to the creation of structured decision-making processes which is produced by data analysis.

**v) Develop cognitive repairs**

Creation of business insights based on data and then going and proving it right for all to see is by far the most effective to both expose biases and create repairs.

### 12.11.3 From Creation to Consumption

Various organizations have created analytics but have failed in consumption. Creating analytics does not automatically result in consumption.

The following are some key questions:

- Do you have experience in creating analytics but failed in consumption?
- Do you have any problem in maintaining the balance between analytics creation and consumption?

If the answer to any of these questions is 'yes', that means your organization suffers from the creation-consumption gap. Organizations should be able to manage this creation-consumption gap and capitalize analytics as a source of consumptive advantage.

### 12.11.4 Analytics for Business and Market Intelligence

Big data analytics uses three types of business analytics. They are:

i) Descriptive analytics

ii) Predictive analytics

iii) Prescriptive analytics

i) *Descriptive analytics:* It describes the previous business analytics. It uses SAS and SPSS for descriptive statistics.

ii) *Predictive analytics:* It uses the previous business analytical information and predicts future outcomes with some degree of likelihood.

iii) *Prescriptive analytics:* It uses previous business information to direct future activities to achieve optimal results.

These three techniques have been used for decades, combining with big data in shifts. Some of the important factors/aspects to be considered while dealing with analytics include:

- Using more or all of the data for predictive model
- Combining analytical models to improve the results
- Using new learning in predictive models
- Making predictive model close to real-time analytics
- Applying predictive models rather than new techniques

**Example: CVS is Promoting Learning Opportunities to help its Professionals Consume the Output of Analytics in its Analytics Initiatives**

CVS has put in place a process to enhance the company's organizational capability to understand how the outside world is changing in terms of AI, analytics, data science, and machine learning. The leadership team realized that this understanding is just not sufficient with the core analytics team. The entire organization needs to get the shared understanding. This is being addressed through a massive training program covering entire workforce. If everyone has this understanding, the technologies will be part of every business process at every level.

*Source: CVS Health Paves Path to Better Care With Data, AI – WSJ, March 10, 2021, Accessed on 06/05/2022.*

## 12.12 Correlation Analysis

What is correlation?

Correlation is the most useful statistics which describes the degree of relative-predictive-and-prescriptive-analytics.

After calculating correlation, we determine the probability of observed correlation by conducting a test of significance. If one thing causes another, then we say that the two will be correlated. Two things that are correlated are not necessarily related by cause because one is a subset of another.

With big data comes great responsibility because advanced algorithms are developed to help us consistently address questions we use to analyze vast amounts of data. We must continue to rely on the expertise of data scientists to ask the right questions and draw the correct conclusions.

### 12.12.1 Logistic Regression

Logistic regression is a statistical method for analyzing a data in which there are one or more independent variables that determine an outcome.

- Logistic Regression is a predictive model.
- Logistic regression model does not involve decision trees.

- Logistic regression can be used only with two types of target variables:
- A categorical target variable
- A continuous target variable

### 12.12.2 Factor Analysis

Factor analysis is a tool used to measure the relationship between large numbers of variables. It allows researchers to use psychological scales to measure directly by collapsing a large number of variables.

The main concept of factor analysis is to measure the variables which are associated with a latent variable (which is not measured directly). For example, people may respond similarly with regard to income, education and occupation—all of which are associated with latent variable socioeconomic status.

In every factor analysis, the number of factors and variables are the same and the factors are always listed in the order of variation. Therefore, each factor captures overall variance in the observed variables.

The eigenvalue is a measure of the variance of the observed variables. Any factor with an eigenvalue ≥1 explains more variance than a single observed variable.

**Factor loading**

Factor loading means the relationship of each variable under each factor. Here is an example of the output of a simple factor analysis with just six variables and two resulting factors. Table 12.2 shows example factor loadings of different factors.

**Table 12.2: Factor Loadings**

| Variables | Factor 1 | Factor 2 |
|---|---|---|
| Income | 0.65 | 0.11 |
| Education | 0.59 | 0.25 |
| Occupation | 0.48 | 0.19 |
| House value | 0.38 | 0.60 |
| Number of public parks in neighborhood | 0.13 | 0.57 |
| Number of violent crimes per year in neighbourhood | 0.23 | 0.55 |

*Source: http://www.theanalysisfactor.com/factor-analysis-1-introduction/*

The variable with the strongest association with the underlying latent variable Factor 1, is income, with a factor loading of 0.65. So that we can say that the variable income has a correlation of 0.65 with Factor 1. This would be considered a strong association for a factor analysis in most research fields.

## 12.13 Marketing Intelligence Using Analytics

Today, leading companies are looking to improve business performance via faster, better decision-making by applying advanced predictive modeling to their vast and growing volumes of data. Business analytics, whether for marketing, CRM, loyalty or operations, provides organizations with valuable insights from their data — allowing them to uncover and act on new opportunities to increase revenue and profitability.

Market Basket Analysis is a data modeling technique used to find associations between items or events by determining the likelihood for them to occur together. Taking its name from the concept of identifying products that customers are putting into their shopping cart, Market Basket Analysis is also commonly referred to as Product Affinity Analysis or Association Rule Learning.

Market Basket Analysis is one of the many advanced models. A typical approach of Market Basket Analysis is to leverage large amounts of customer transaction data to determine products that are most commonly purchased together. Understanding these purchasing patterns empowers marketing and sales organizations to make more informed decisions about how and where to deploy their efforts and resources.

An obvious application of Market Basket Analysis is in the retail sector where retailers have large amounts of transactional data and often thousands of products. One of the recognizable examples is the Amazon.com recommendation system: "Customers who bought this item also bought items A, B and C".

In order to capitalize on big data value, big data apps have started to emerge. The horizontal big data apps (machine log analytics) and vertical big data apps (telecommunications analytics) are emerging.

These emerging techniques are designed to solve specific business problems which incorporate deeper and more complex prescriptive analytics.

Top emerging technologies include:

- Fuel-cell vehicles (Cars that run on hydro zed)
- Next generation robotics (Rolling away from the production line)
- Recyclable thermoset plastics (A new kind of plastic to cut landfill waste)
- Precise genetic-engineering techniques (A breakthrough; offers better crops with less controversy)
- Additive manufacturing (Making things from printable organs to intelligent clothes)
- Emergent artificial intelligence (What happens when a computer can learn on-the-job?)
- Distributed manufacturing (The factory of the future is online and on your doorstep)
- Neuromorphic technology (Computer chips that mimic the human brain)

**Example: Leading Indian Media-Tech Company Explores Data in Real-Time and Makes Faster Decisions with the Big Business Intelligence Platform**

NEWJ (video creation and curation company) chose the business intelligence platform from BIPP Inc. to provide analytical insights which can help the company in its expansion plans in India. The analysis needed to connect rows from multiple tables to generate complex daily management reports. The report generation is faster as the queries need not be rewritten. The reports enable the editorial team to select and publish appropriate content on various platforms in many languages.

*Source: NEWJ Adopts bipp Business Intelligence Platform to Drive Growth (martechseries.com) 4th May, 2022, Accessed on 06/05/2022.*

## 12.14 Customer Experience Management Using Analytics

Market Basket Analysis (MBA) is a data mining technique which looks at purchase coincidence and is widely used in the consumer packaged goods. It investigates whether two products are being purchased together and whether the purchase of one product increases the likelihood of purchasing the other.

Market Basket Analysis results in a better understanding of your customers and their purchasing behavior by allowing you to explore associations and predict the likelihood of a customer's future purchase behavior based on associations. It is one of many advanced business analytics tools that can help organizations optimize marketing and sales operations for improved performance.

Marketing and sales organizations across all industries are looking to analyze, understand and predict customer purchase patterns towards achieving strategic goals to reduce churn rates and maximize Customer Lifetime Value (CLV). Selling additional products and services to existing customers over their lifetime is the key to optimizing revenues and profitability. Market Basket Analysis association rules identify the products and services that customers typically purchase together, empowering organizations to offer and promote the right products to the right customers.

Moreover, with predictive analytics, organizations are able to promote their most profitable products and services to the most likely buyers. They can also encourage

additional purchases by introducing new targeted products, products with high margin, or high performing products which may not have otherwise been an obvious next purchase.

**Example: L.L. Bean turned to Qualtrics Experience Management to Understand how Customers Purchase Products**

Based on many studies, L.L. Bean found out a key segment of customers to target. The segment is outdoor family enthusiasts. They needed to understand and strategize around the segment in each department. Qualtrics Customer Experience management tool was used for this. The insight was: gardening is the most preferred activity for this group in March and May. The spring product selection is done now with this new insight in mind.

*Source: L.L. Bean + Qualtrics // Experience Management, 2022, Accessed on 06/05/2022.*

## 12.15 Business Intelligence Tools

Business intelligence tools include AQL (Associate Query Logic), Decision Support Systems (DSS), Executive Information Systems (EIS), Management Information Systems (MIS), Query and Reporting Tools, OLAP (Online Application Processing) Tools, Data Mining Tools, and ETL (Extract, Transform and Load) tools.

Most influential commercially available business intelligence tools are from organizations such as Business Objects, Microsoft, SAS, Teradata, PeopleSoft, ORACLE, IBM, Manhattan Associates, Insight Software and OutlookSoft.

Oracle business intelligence applications include Oracle Financial Analytics, Oracle Project Analytics, Oracle Sales Analytics, Oracle Price Analytics, Oracle Marketing Analytics, Oracle Procurement and Spend Analytics, Oracle Supply Chain and Order Management Analytics, Oracle Human Resources Analytics, Oracle Service Analytics, Oracle Loyalty Analytics, and Oracle Call Center Telephony Analytics. Oracle business intelligence applications are capable of integrating with Oracle E-Business Suite, JD Edwards Enterprise One, PeopleSoft Enterprise and Siebel CRM. It consists of more than 3,000 pre-built reports. A proper analytics package comes with data schemas, dashboards, predefined reports, business views, and an integrated set of tools. The business intelligence tools for semi-structured data are still maturing.

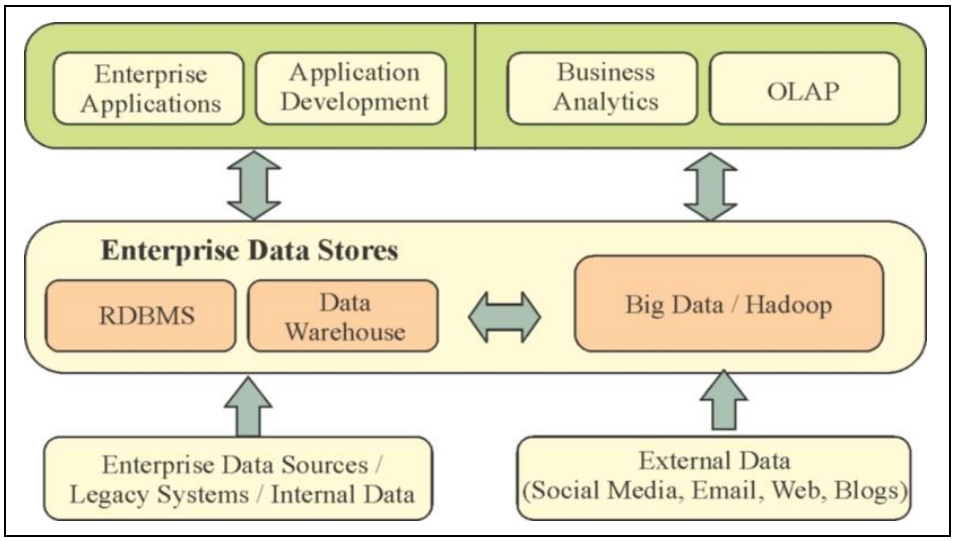
### 12.15.1 Big Data

The amount of data growing in organizations is huge and also unmanageable sometimes.

Big data can be used in wide areas such as retail, mobile services, e-commerce, education, financial services, scientific research, manufacturing, life sciences, bioinformatics, physical sciences, and astronomy. Big data applications include traffic management, urban planning, environmental modeling, smart materials, computational social sciences, financial risk analysis, security and intelligent transportation, seismic data analysis, meteorological data analysis, etc.

Organizations may use big data for customer retention, risk assessment and brand management. Big data requires different data mining techniques other than the traditional statistical techniques. Big data analysis process is explained in Figure 12.4.

**Figure 12.4: Big Data Architecture**




*Source: ICFAI Research Center*

### 12.15.2 Hadoop

Apache Hadoop is an open source platform to manage big data. It addresses the issues such as low cost, reliable storage and tools for analyzing unstructured data, etc. It is a project of Apache Software Foundation. Hadoop consists of a fault-tolerant system known as Hadoop Distributed File System (HDFS). It provides storage infrastructure that can hold data without loss. It creates clusters of machines and coordinates between them. These clusters are built with less expensive hardware machines. HDFS stores three copies of each block of data in three different servers in the cluster. Even if two servers go down in the cluster, the file can still be retrieved without any data loss.

Hadoop tool works at the whole quantity of data in the relational database. It uses a technique known as MapReduce that splits the task and pushes it on to different servers and later collates the results achieving the operational parallelism. It automatically restarts the work if any node goes down in the cluster. Hadoop Distributed File System (HDFS) and MapReduce are the key features of Hadoop providing a reliable and low-cost solution to big data.

### 12.15.3 Big Data Analytics

An organizational business analytics can include customer analytics, supply chain analytics, IT analytics, HR analytics, financial analytics, etc. For example, customer analytics can find what type of customers are profitable to the organization, supply chain analytics can find what inventory optimization levels

are suitable for the organization, IT analytics can find whether IT services are efficient in the organization or not, HR analytics can find out what is the cost of recruitment, financial analytics can find what are the cost and revenue drivers of the organization and their impact on the profitability.

Some of the analytical technologies include neural networks, genetic algorithms, swarm intelligence, information extraction, text categorization, text mining, audio mining, video mining, rule-based engines, data mining tools, simulation tools, spreadsheets, and OLAP tools.

Predictive analytics is a technique for predicting the future scenarios for the organization. It gets data from the data warehouse and applies mathematical algorithms to predict the future trends of a business. It uses the techniques such as regression, logistic regression, time-series analysis, duration analysis, situational analysis, multivariate analysis, classification, association, and machine learning techniques such as neural networks and radial basis functions.

**Example: Mitsui Sumitomo Insurance Uses Teradata Business Intelligence Tool to Achieve the Much-Needed Data Integration**

Mitsui Sumitomo Insurance has huge volume of data on its databases spread over different databases. This made it difficult for the users to access. Data integration was found critical for digital transformation. The first step was to decommission the data from various databases. Then the data has to be migrated to an integrated cloud storage. The integration and centralization can only handle the growing data analysis and offer easy access by end users for decision making.

*Source: Improving CX Demands Digital Transformation & Analytics (teradata.com), 2022, Accessed on 06/05/2022.*

**Activity 12.4**

**Prescriptive Analytics**

A nationalized bank in India decided to use analytics for finding customer loan requirements. The board of the organization decided that with analytics capabilities they would like to find out the type of customers and the type of various loans, customers are looking for. Can an alternative loan scheme be suggested to a specific customer? What is the best alternative among the available loans? What can be suggested to the prospective customer? The IT team decided to use big data analytics for this purpose. Suggest the team the type of analytics algorithms which suit their requirement.

**Answer:**

## 12.16 Moving beyond the tools to Analytics Applications

Organizations are using Data Visualization as a way to take immediate action.

The companies are collecting billion rows of data a day and generating tableau for analytics and reporting. Everyday product managers analyze hundreds of millions of rows to understand the user dynamics and problems. Data visualization is making big data analytics iterative. It is also reducing the cycle time of big data analytics so that immediate action can be taken. Big data visualization is still in the early stages and commercial vendors are using open-source projects who are leading the charge. The following are some open-source projects:

- Qlikview – www.qlikview.com
- Tableau – www.tableausoftware.com
- Micro strategy – www.microstrategy.com
- SAS – www.sas.com
- Cubism (a plug-in for D3 for visualizing time series) – http://cubsim.com
- Arbor JS, a java-based graph library, http://arborjs.org
- Java Script Info Vis Toolkit, http://thejit.org
- Many Eyes, data visualization tools from IBM Research.

**Activity 12.5**

**HR Analytic**s

The HR manager would like to know the impact their organizational culture has on the employee productivity in the organization. For this purpose, he wants to have data-based evidence. He would like to use the organizational data warehouse maintained by the IT department for this purpose. He seeks the help of IT project manager in finding the impact of organizational culture on employee productivity. The IT project manager tells the HR manager that some tools can be used to do this work. What are those tools? What statistical technique needs to be applied in this case?

**Answer:**

## 12.17 Introduction to Google big Query, Google dataflow and Apache Spark

There are many cloud-based big data managing services provided by different companies. Some of the popular ones are Big Query and Dataflow by Google Cloud platform, Spark technology by Apache, etc.

### 12.17.1 Google Big Query

Big Query is a subscription-based data analytics service by Google Cloud platform to manage Big Data. The main advantage of using this service is that there is no need to manage IT infrastructure or hire a database administrator. Minimum knowledge of SQL is essential to query the data from the cloud. Organizations with high volumes of real-time data on which analysis is required can make use of this service. Query operations on massive datasets need specialized expensive hardware and is time-consuming; such requirements can be handled using the processing power of Google's infrastructure. There are many Fortune 500 companies and even startups that use these services. Big Query has the following components:

- Projects*:* These are top-level containers in Google Cloud platform used to store Big Query data and information regarding billing and authorized users. Every project has a unique ID.
- Tables: The tables contain data in Big Query, schema with field names, data types details
- Datasets: The datasets allow to organize and control access to your tables; tables are entities within the datasets.
- Jobs: Jobs is a Set of action tasks like load, export data, query data, or data which is executed by Big Query. Jobs execute independently and they consume time.

### 12.17.2 Google Dataflow

Google Cloud Dataflow is a tool to perform data-processing tasks on data, irrespective of its size and type.

Cloud Dataflow consists of two major components:

- It consists of a few SDKs which allow defining data processing jobs. Dataflow SDKs are based on a unique programming model to handle large-scale cloud data processing. Data processing jobs are defined by writing programs with the help of Dataflow SDKs.
- The Dataflow service integrates a set of Google Cloud Platform technologies, like Google Compute Engine, Google Cloud Storage, and Big Query. These are used to execute data processing jobs on Google Cloud Platform resources.

### 12.17.3 Apache Spark

Apache Spark is a fast open-source cluster computing framework used for big data processing, with a built-in library to support streaming, SQL, machine learning, and graph processing activities. This framework is maintained at AMP Lab at UC Berkeley. Compared to Hadoop, it has dual stage MapReduce method. Apache Spark provides much faster performance for some special applications. The major advantage is its ease of use. Due to the availability of high-level operators, parallel apps can also be built. Its high-performing tools like Spark SQL, MLlib for machine learning, GraphX, and Spark Streaming allow the user to perform streaming and complex analytics activities.

**Example: Ford uses Google Tools Like Big Query to Consolidate Its Data and move into Next Generation Manufacturing**

The number of sensors capturing data from machines located at different locations grew rapidly. The management felt the need for analysing the data in a holistic way. Ford and Google joined hands to address this. The platform developed can handle hundreds of machines connected between two plants. Some 2.5 crore records were collected weekly. Ford can now use predictive analytics for predictive and preventive maintenance.

*Source: Google intelligently modernizes Ford manufacturing | VentureBeat, May 5, 2022, Accessed on 06/05/22.*

**Check Your Progress - 4**

16. Which of the following organizations provide business intelligence tools?
    a. Business Objects
    b. Microsoft
    c. Peoplesoft
    d. Business objects, Microsoft and Peoplesoft
    e. Wipro
17. Which of the following is not a big data characteristic?
    a. Size
    b. Speed
    c. Data type
    d. Data Variety
    e. Cost
18. Which of the following features included in Hadoop?
    a. HDFS (Hadoop Distributed File System)
    b. MapReduce

c. HDFS and MapReduce

d. High-Cost Solution

e. Storage

19. Which of the following is not included in Big data analysis process?

a. Data Acquisition

b. Data Integration

c. Data Analysis

d. Inconsistency

e. Data Restructuring

## 12.18 Summary

- The challenges facing business intelligence include the volume of data (size), security, data retention, performance targets, and benchmarking. However, business intelligence systems are useful for strategic, tactical and operational planning and decision-making.
- Data warehouse is the major component of business intelligence. It helps in the propagation of data in the organization. It extracts, cleanses, integrates, transforms, and stores the data and further transmits it for query processing and analysis.
- The essential components of business intelligence systems include data warehouse, data marts, corporate performance management, ETL tools, OLAP, analytical tools, data visualization, data mining, geographic information system, and a well-defined workflow.
- Business intelligence tools include AQL (Associate Query Logic), Decision Support Systems (DSS), Executive Information Systems (EIS), Management Information Systems (MIS), Query and Reporting Tools, OLAP tools, Data Mining Tools, and ETL tools.
- There are many cloud-based big data managing services provided by different companies; some of the popular ones are Big Query and Dataflow by Google cloud platform, Spark technology by Apache.

## 12.19 Glossary

**Customer Life Time Value:** It is a prediction of the net profit attributed to the entire future relationship with a customer**.**

**Data Cleansing:** It is the process of removing errors in the data. Incomplete, erroneous and inconsistent data is removed from the data. This process is known as data cleansing. It is required after data acquisition.

**Expert Systems:** These are the systems developed using artificial intelligence techniques. They can be developed using rule-based programming language. For example, Prolog (Programming in Logic) can be used to develop expert systems. Expert systems can be used in medical diagnosis as well.

**Legacy Systems:** These are the systems developed using first-generation programming languages. The extension and reusability of these systems were very complex. They are not extendable like object-oriented systems. They were developed using programming languages such as BASIC and FORTRAN.

**Metadata:** It is the data about the data. The date and time of data creation, who created the data, size of data, the date and time last modified, who has recently accessed the data, etc., are maintained in the metadata. There are metadata repositories in business intelligence systems.

**Pivot Analysis:** Pivot analysis is done to find out the projected support and resistance levels of data. It is used to generate ad-hoc reports useful for managerial decision- making.

**Taxonomy:** It is a technique of analyzing semi-structured or unstructured data. Usually, classification is applied to semi-structured data analysis.

**Transaction Processing Systems:** These systems were developed using business-oriented programming languages such as COBOL. They used to have a master file and a transaction file. They also support batch processing instead of online processing. These were used in banks in the early days of computing for transaction processing. Now, these transaction processing systems are replaced by online distributed systems and cloud computing in multi-national banks.

## 12.20 Self-Assessment Test

1. Distinguish between data, information and knowledge. How are business intelligence systems helpful in knowledge management in the organization?
2. What are the features of a data warehouse? Explain.
3. Explain the business intelligence architecture. What are its main components?
4. Briefly describe the data mining techniques. Which techniques can be used to find associations between two events?
5. Write a short note on Google Big Query.
6. What is big data? What are its characteristics? What are the important features of Hadoop?

## 12.21 Suggested Readings / Reference Material

1. Rodney Heisterberg and Alakh Verma (April 2022). “Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video and Big Data Enables Competitive Advantage,” Narrated by Stephen Graybill.

2. Jonathan S Walker (2021). Social Media Marketing For Beginners - How To Make Money Online: Guaranteed Strategies To Monetizing, Mastering, & Dominating Any Platform For Your Brand, JW Choices.
3. Barry Connolly (2020). Digital Trust: Social Media Strategies to Increase Trust and Engage Customers, Bloomsbury Business.
4. Seema Gupta (6 August 2020). Digital Marketing McGraw Hill; Second edition.
5. Tracy L. Tuten, Michael R (15 June 2020). Solomon et al, Social Media Marketing, SAGE Publications Pvt. Ltd; Third edition.
6. Paul Martin Thomas Erickson (2019). Social Media: Usage and Impact, Global Vision Publishing House, 2 edition.
7. Steve Randazzo (2019). Brand Experiences: Building Connections in a Digitally Cluttered World, Paipen publishing.

## 12.22 Answers to Check Your Progress Questions

**1. (c) Organizational performance and decision-making.**

Business intelligence is required for both organizational performance and decision-making.

**2. (c) Data warehouse and data marts.**

Business intelligence systems make use of data warehouse and data marts.

**3. (c) Processed data**

Information is processed data.

**4. (d) Decision Support System**

DSS stands for Decision Support System.

**5. (a) Understanding**

Knowledge depends on understanding.

**6. (c) Structured data or unstructured data**

Data can be structured data or unstructured data.

**7. (d) Restructuring Data**

Data warehouse functionality includes cleansing data, storing data and transforming data.

**8. (d) Data warehouses, data marts, and ETL tools**

Business intelligence architecture consists of data warehouses, data marts and ETL tools.

**9. (c) Scalable and Secure**

Business intelligence architecture should be scalable and secure.

**10. (a) Extract, Transform and Load**

ETL stands for Extract, Transform and Load.

**11. (c) Online Analytical Processing**

OLAP stands for Online Analytical Processing.

**12. (d) Data warehouse, specific to business function, specific to department**

Data mart is a sub-set of data warehouse, specific to the business function, specific to the department.

**13. (b) Data Marts**

Cubes are derived from data marts.

**14. (d) Predictive analysis, statistical analysis and ad-hoc analysis**

Business intelligence methodologies include predictive analysis, statistical analysis and ad-hoc analysis.

**15. (e) Factoring**

Classification, multidimensional analysis and clustering are data mining techniques.

**16. (d) Business objects, Microsoft and Peoplesoft**

Microsoft, Peoplesoft and Business Objects provide business intelligence tools.

**17. (e) Cost**

Big data characteristics include size, speed and data type.

**18. (c) HDFS and MapReduce**

Hadoop features include Hadoop Distributed File System (HDFS) and MapReduce.

**19. (d) Inconsistency**

Big data analysis procedure includes data acquisition, data integration, data analysis, etc.

# SMACS (Social, Mobile, Analytics, Cloud, and Security) Technologies for Business

## Course Structure

| | |
|---|---|
| **Block 1: Introduction to Digitization** | |
| Unit 1 | Introduction to SMACS (Social, Mobile, Analytics, Cloud, and Security) Technologies for Entrepreneurship Development |
| Unit 2 | Social Networking Platforms and Stakeholders |
| Unit 3 | Product Development Using Social Media |
| Unit 4 | Customer Relationships through Social Media |
| **Block 2: Mobile Technologies for Business** | |
| Unit 5 | Mobile Devices and Platforms |
| Unit 6 | Mobile Operating Systems |
| Unit 7 | Mobile Apps for Business Organizations |
| Unit 8 | Mobile Business Process Management |
| **Block 3: Business Analytics** | |
| Unit 9 | Decision Making Using Big Data |
| Unit 10 | Handling Unstructured Data |
| Unit 11 | Data Analytics for Top Management Decision Making |
| Unit 12 | Business and Marketing Intelligence Using Analytics |
| **Block 4: Cloud for Business** | |
| Unit 13 | Cloud Architectures and Services |
| Unit 14 | Enterprise Systems Development Using Cloud Technologies |
| Unit 15 | Clouds for Social Marketing |
| **Block 5: Security Technologies for Business** | |
| Unit 16 | Data Security in Organizations |
| Unit 17 | Network Security in Organizations |
| Unit 18 | Information Security in Cloud Environment |
| **Block 6: Applications of SMACS** | |
| Unit 19 | SMACS Applications to Top Management |
| Unit 20 | SMACS for Marketing |
| Unit 21 | SMACS for Operations |