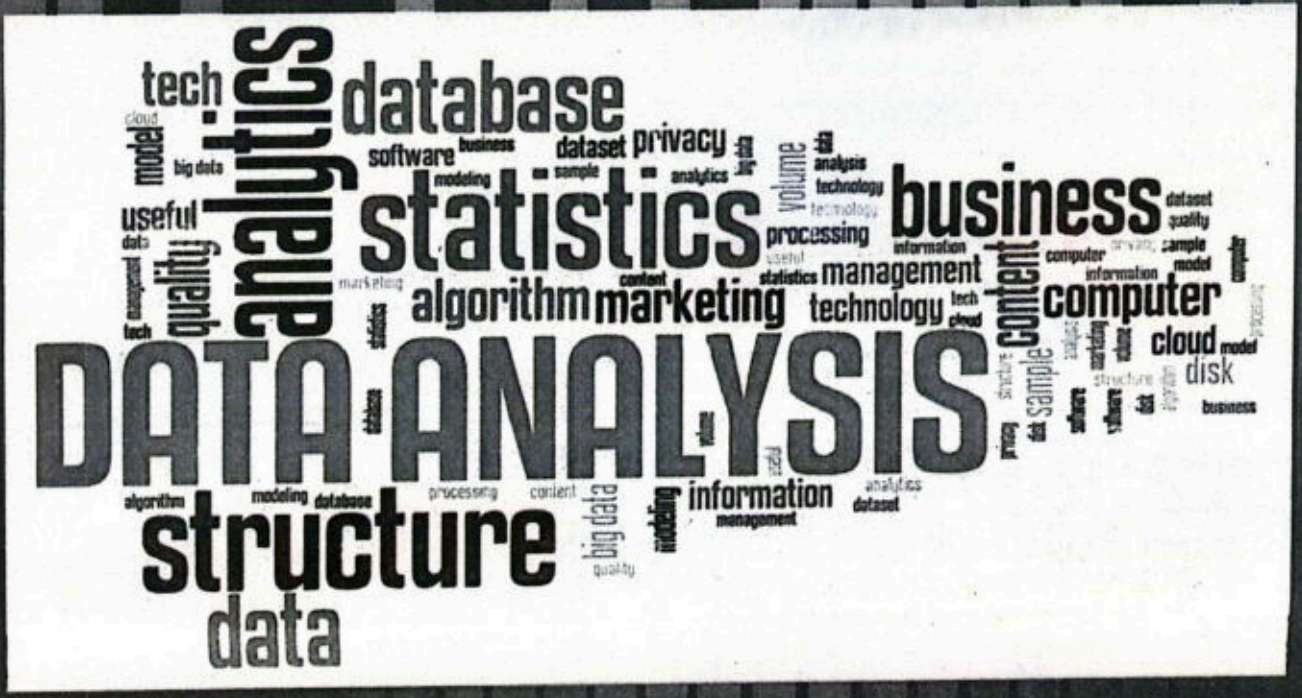
UNIT 4

Data and Analysis

After completing this lesson, you will *be able to:*

- explain the scope of the data science field as an interdisciplinary field (computer sciences, mathematics & statistics, and business knowledge & understanding).
- define and explain data types, data collection, and data storage.
- define and explain big data, and applications of big data in real-world business.



Introduction

A computer system is a fundamental and important part of modern life. It has revolutionized the way we work, communicate, learn, and entertain ourselves. . In today's world, we are surrounded by a lot of data, which may be on our computer system or otherwise. This data is continuously growing and to get meaningful information from it we need to follow some discipline. Data science is the branch of knowledge, in which computer programming skills along with mathematics and statistics is used to extract meaningful information from the collection of data.

4.1 Data and Analysis

Data Analytics: Data analytics is the process of examining raw data to draw conclusions about the information it contains. It's like solving a puzzle or retrieving meaningful results from the given or collected data.

To analyze data, you can use different techniques like mathematical calculations, statistical techniques, charts etc. These tools are helpful to find patterns, trends, and insight from the data. For example, after recording hourly temperature data in a science experiment you can create a graph to see how it changed over time. From graphical representation of data, you draw a conclusion that it got warmer as the day went on, that information will be the result of your data analysis.

The goal of data analytics is to transform raw data into actionable knowledge that can inform decision-making, improve processes, and solve problems. It can involve working with both quantitative (numerical) and qualitative (descriptive) data.

4.1.1 Data Science

Data Science refers to an interdisciplinary field of multiple disciplines that uses mathematics, statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it. It is like a pipeline from data to insights. This insight or knowledge is used to find patterns in the data. The result drawn can be used for making informed decisions to solve real world problems e.g., medical, education, scientific research, and business etc.

4.1.2 Concepts of Data Science

Data science consists of many components, theories, and algorithms. To understand data science and make its productive usage, following are some key concepts or components that lay the foundation of data science:

Data: As mentioned earlier, data is a collection of observations, facts or information collected from different sources. This data can be in the form of numbers, measurements, words, observations, or in audio or video form. It could be structured(processed) data which is in the form of tables or unstructured(unprocessed) data in the form of audio, video, tweets, pdf files etc.

Dataset: Dataset is a structured or processed collection of data usually associated with a

unique body of work. This collection of data is related to each other in some way, for example a collection of brain CT scan of brain tumor patients is a dataset which can be used to evaluate certain pattern or trend common in the entire dataset.

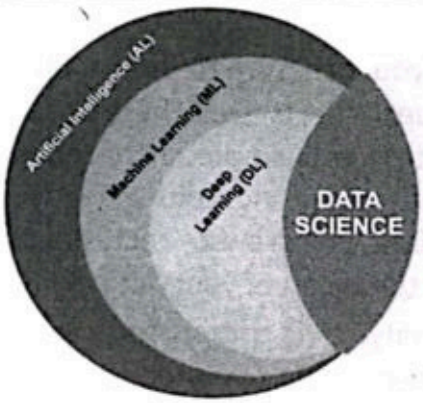
Statistics and Probability: Statistics is the analysis of the frequency of past events and probability is to predict the likelihood of future events. Data scientists use statistics and probability to find patterns and trends in the data.

Mathematics: Mathematics is a fundamental part of data science which helps to solve problems, optimize the model performances, and interpret huge complex data into simple and clear results, for decision making.

Machine Learning: Machine learning is a branch of Artificial Intelligence and computer science which emphasis on the use of data and algorithms to imitate human learning by the computers.

Deep Learning: Deep learning is the subset of Machine learning, with emphasis on the simulation or imitation of human brain's behavior by using artificial neural networks.

Data Mining: Data mining is the subset of data science which primarily focuses on discovering patterns and relationships in existing datasets. The usage of techniques and tools is limited in data mining as compared to data science.



The relationship of some of the key concepts is shown in Fig. 4.1

Data Visualization: Data visualization is the graphical representation of data using common charts, plots, infographics, and animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Big Data: Big data refers to handling large volumes of data. Data scientists use big data to find patterns and trends in datasets, to obtain more accurate and reliable results. The huge size of data provides more opportunities for machine learning and provides better results.

Predictive Analysis: Predictive analysis is the use of data to predict future trends and events based on historical data.

Natural Language Processing(NLP): It is the ability of computers to understand, interpret, and generate human language. It enables computers to perform human-like tasks e.g. language translation, content(text) summarization, and sentiment analysis. The common examples of NLP are chatbots, language translators, sentiment analysis tools and conversational AI Plateforms etc.

4.1.3 Scope and application of Data Science

DO YOU KNOW?
Sentiment analysis is the term used to identify the sentiments of a customer by analyzing the review about the product. The sentiment can be positive, negative, or neutral. Sentiment analysis can be performed on reviews, text, opinions etc.

Data science is used for a wide range of applications which includes predictive analytics, machine learning, data visualization, recommendation systems, sentiment analysis, fraud detection, and decision-making in various industries like healthcare, finance, marketing, and technology. A business problem is a gap between the existing and desired state of a situation. It is a desired action or series of actions to achieve an objective. Various business problems can be solved through data science, some of them are as follows:

- To decide the best routes for shipping of goods or passenger airplanes.
- To choose the best product among many, which one to buy A or B.
- To foresee delays for flight/ship/train etc. (through predictive analysis).
- To create promotional offers (which products are more popular than others)
- To find the best suitable time to deliver goods to reduce cost.
- To forecast next year's revenue for a company.
- To analyze health benefit of physical training programs.
- To predict some fore coming event like who will win elections.

4.1.4 Business problems and Data Science

Data science can be applied to various businesses after analyzing the available data, some of them are:

Industry: Data science can be used to make data driven decisions by analyzing historical data and predicting future trends. It can also help in effective marketing an improving quality control.

Consumer goods: Data science skills can be used to optimize inventory according to the demand forecasting of particular goods in particular social groups, communities, and demographics.

Logistic companies: These companies can apply data science for their rout optimization, demand forecasting, real-time tracking, load balancing, carrier selection, cost reduction and global trade optimization.

Stock markets: Data science techniques and tools can be helpful in algorithmic trading, market sentiment analysis ,volatility predictions, quantitative analysis, machine learning based trading, market surveillance and risk management etc.

E-commerce: In e-commerce data science helps in recommendation systems, customer segmentation, shopping cart analysis, fraud detection, supply chain optimization and customers sentiment analysis etc.

4.2 Data types in Data Science

In data science we can mainly classify data into two main types qualitative(categorical) and quantitative(numeric).

Qualitative or Categorical data describes an object or a group of objects that can be labeled according to some group or category. It cannot be represented in numerical form. For example, data including colors, places, etc. it is further subdivided into two types:

i. Ordinal data
ii. Nominal data

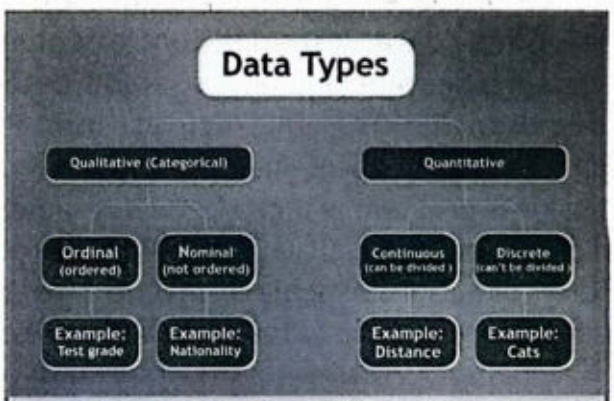


Fig. 4.2 Data types in Data Science

Ordinal Data:

Ordinal data sees a specific order or ranking, it uses certain scale or measure to group data into categories. Such as in test grades, economic status, or military rank.

Nominal Data:

Nominal data does not have any order, it can be labelled into mutually exclusive categories, which cannot be ordered meaningfully. For example, if we consider the categories of transportation as car, bus or train. Similarly, gender, city, color, employment status are also examples of nominal data.

Quantitative or Numerical data deals with numeric values, that can be computed mathematically to draw some conclusions. Examples of numeric data are height, weight, number of students in a school, fruits in a basket etc. Quantitative data can be further divided into two types:

i. Discrete data
ii. Continuous data

DO YOU KNOW?

Table (Relation)

A table is a combination of rows and columns, which can store data about a particular person, place, thing or event.

Rows (Record, Tuple, Object)

A row is a single entry or record in the table. It is also called an instance of the table.

Column (Object Attributes, Properties)

A column is a single characteristic or field in the table.

Example: A student table can contain attributes name, age, grade. One row will contain data about one particular student.

Discrete Data: It includes data which can only take certain values and cannot be further subdivided into smaller units. This data can be counted and has a finite number of values. For example, the number of product reviews, ticket sold, computers in certain departments, employees in a company etc.

Continuous Data: It refers to the unspecified number of possible measurements between two realistic points or numbers. For example, daily wind speed, weight of newborn babies, freezer's temperature etc.

Continuous data is further subdivided into two types:

- Interval Scaled data
- Ratio scaled data

Interval Scaled data: In this type of data the differences between consecutive measurements are equal, but there is no true zero point. For example, Temperature in Celsius or Fahrenheit (0°C or 0°F) does not mean the absence of temperature. Interval scaled data can be added, subtracted and compared.

Ratio scaled data: In this type of data the differences between the consecutive measurements are meaningful. There is a true zero point in this sort of data. For example, 0 Kg means absence of weight. This data can be added subtracted and compared.

4.2.1 Sources of data

To analyze data for predictive analysis and decision making, the initial step is data collection through various reliable sources. Data can be divided into two categories, primary data, and secondary data. Primary data is collected directly by questionnaires, surveys, and interviews. Primary data can also be collected through experiments and recording observations. secondary data is collected from some previously recorded from primary data. Following are some sources of data:

Website: Collecting tweets regarding some topic or thread.

Surveys: Collecting firsthand data by performing surveys about some event, movie, or anything else.

Sensors: Collecting seismic data regarding changes under the earth which cause earthquakes.

4.2.2 Dataset and Database

A dataset is a structured or organized collection of data , which is usually associated with a unique body of work. However, a database is an organized collection of data stored in multiple datasets or tables. These tables can be accessed electronically from the computer system for further manipulation and update.

To perform actions on the data stored in a database, we need a Database Management System (DBMS). DBMS is the interface between the database and the end user, providing a platform to create, modify, and retrieve data. There are many different database management systems available, depending on the type of database being used.

For example, relational databases, which store data in tables, can be managed by database management systems such as MySQL, Oracle, MS-Access and IBM Db2. These are the most used databases in Data Science, for the data which is

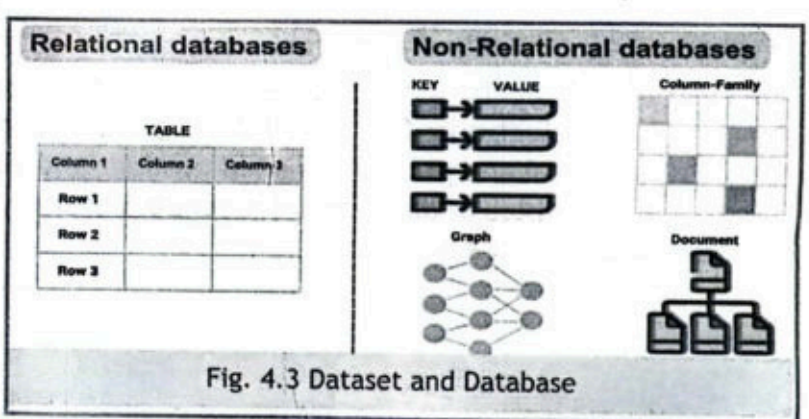


Fig. 4.3 Dataset and Database

presented in a tabular format. Non-relational databases, which store data in forms such as key-value pairs, column families, or graphs, can be managed by database management systems like MongoDB and Cassandra. Non-relational DBMS are also called NoSQL DBMS.

To further understand, how database is helpful to store data, consider online shopping websites. The products shown on the websites are stored in a database. To manage the products, customers and their orders a specific database is maintained. The following is the structure of an online website database:

Tables: Product, Customer, Orders

We can display each table as follows:

Product Table

Product ID	Product Name	Price	Stock Quantity	Expiry Date
Pd001	Chocolate 50Gm	Rs. 250	700	June 2026
Pd002	Honey 500Gm	Rs. 1000	150	December 2027

Customer Table

Cust. ID	Customer Name	Email ID	Contact Number	Address
Cust001	Ali Ahmad	ali@gmail.com	03901597535	152 main road Islamabad
Cust005	Shagufta Asad	sasad@gmail.com	03624569852	78 DHA Phase-II Karachi

Order Table

Order ID	Cust. ID	Product ID	Qty.	Order Date	Status	Payment
Os987	Cust005	Pd002	2	4 February, 2025	Pending	COD
Os873	Cust001	Pd002	5	6 February, 2025	Shipped	Paid

The customers visit the website, choose the desired product/products and add them to the cart. They select quantity and mode of payment and place the order. In this way, the database is helpful and provides ease of product and order management for the website. Some other examples of databases are school library system, hotel management system, student management system etc.

4.2.3 Role of database in data science

Before the advent of database systems, computer scientists relied on file management systems to store and manage data. However, without a structured method of storing data, it would be of little use. This is why databases were introduced to manage and store large amounts of data. The first database management system was developed in 1960s.

There are two key reasons why databases have become so popular in recent years:

- The rapid increase in data generation
- The dependence of data science on data

To better understand the importance of databases in our daily lives, let's take an example of supermarkets' evolution. In this case study you will learn how data science had impact on the shopping in current age.

Case study for the use case of Database and Data Science

In the old days, people used to buy their necessities from various shops. For example, if you had to buy a calculator, a box of yogurt, shoe polish and a pair of socks of school uniform, you were supposed to visit four various shops. Such shopping was never an enjoyable experience because shops often had less space for customers, and they had to wait for the shopkeeper to find their desired item. The introduction of supermarkets, however, changed that, as they made shopping much more pleasant by displaying all the products in a large space and making them easily accessible to customers. As the number of products and customers in supermarkets increased, the need for a database system to keep track of all the purchases became critical.

DO YOU KNOW?
Survey: It is a method of collecting information from individuals. The basic purpose of a survey is to collect data to describe different characteristics such as usefulness, quality, price, kindness, etc. It involves asking questions about a product or service from many people.

Data Science plays a crucial role in determining the place of various products in various shelves of the supermarket. For example, the information gathered from the database will guide us to place the products with less shelf life in the most easily accessible shelves. Similarly, predictive analysis provides adequate guidelines that which products would be in high demand in which season/month. For example, in Pakistan during the months of religious and national festivals, the demand of food items and clothing increases as compared to the rest of the year. By analyzing sales data from different supermarket branches, supermarket owners can identify which products need to be stocked in larger quantities and during which months the sales are highest.

To determine the months with the heaviest customer traffic, a graph was plotted between the month and gross income. The analysis showed that the sales were highest in the months of festivals. In this way data science provides maximum benefits to the supermarket owners as well as customers, who can find their desired items easily.

Month	Food Items	Clothing and Footware
Jan-23	11.85	8.95
Feb-23	11.75	9.54
Mar-23	16.53	10.52
Apr-23	15.01	20.25
May-23	10.75	9.52
Jun-23	16.11	18.73
Jul-23	11.56	7.66
Aug-23	10.52	8.71
Sep-23	11.25	8.15
Oct-23	12.75	9.57
Nov-23	11.23	8.41
Dec-23	11.89	10.63

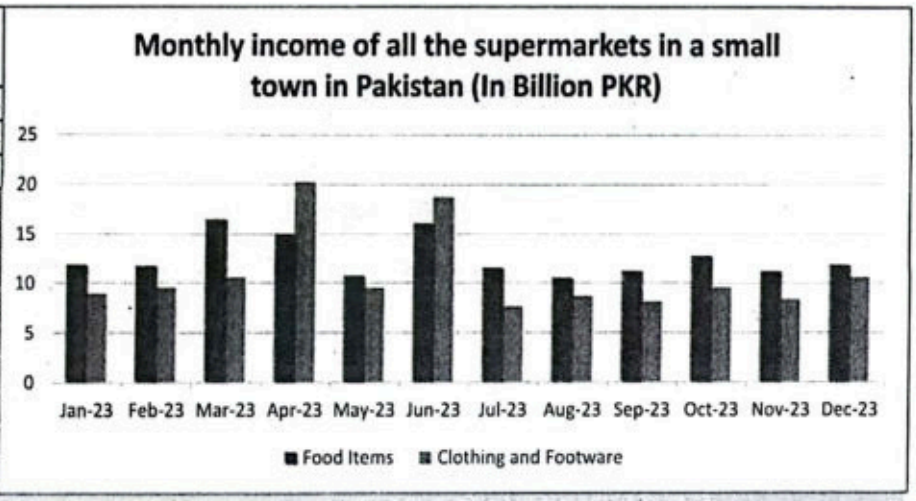


Fig. 4.4 Monthly income of all the supermarkets in a small town in Pakistan

4.2.4 Data Collection in Data Science

Data Collection is the process of collecting information from relevant sources to find a solution to the given statistical enquiry. Collection of Data is the first and foremost step in a statistical investigation. Data collection methods are divided into two categories:

i. Primary data collection

ii. Secondary data collection

Primary data collection methods:

It involves the collection of original data directly from the data source or via direct interaction with the respondent. A respondent is a person from whom the statistical information required for the enquiry is collected. Some common primary data collection methods are as follows:

i. Surveys and Questionnaires

ii. Interviews

iii. Observations

iv. Experiments

v. Focus groups

vi. Sensors

vii. IoT devices

viii. Biometric devices

Secondary data collection methods:

It involves data collection using existing data collected by someone else for some purpose. Such data is usually available in the form of published material like research papers, books, websites etc. Some common secondary data collection methods are as follows:

i. Published sources
ii. Online databases
iii. Government and institutional records
iv. Surveys and Questionnaires conducted in the past
v. Social media data/posts
vi. Publicly available data
vii. Past research studies

DO YOU KNOW?

Investigator:

An investigator is a person who conducts the statistical enquiry.

Enumerators: To collect information for analysis, an investigator needs the help of some people. These people are known as enumerators.

4.2.5 Data Storage

After data collection, effective storage of data is an essential step for managing and analyzing the large volumes of data. There are various data storage methods according to the nature of data. Some common data storage methods are as follows:

i. Relational/NoSQL databases
ii. Data warehouse
iii. Distributed file systems
iv. Cloud based data storage
v. Blockchain

4.2.6 Data Visualization

Data visualization is graphical representation of data to get meaningful insight, trends, and patterns from data. The visual elements which help in data visualization are charts, graphs, maps, figures, and dashboard etc.

4.2.7 Summary statistics

It is information about the data in a sample. It can help understand the values better. It may include the total number of values, minimum value, and maximum value, along with the mean value and the standard deviation corresponding to a data

DO YOU KNOW?

Outliers are unusual data points that are significantly different from the other values in the dataset. They can be errors or unusual patterns.

Trends are the patterns or directions in the data over time or across different categories. e.g. increase/decrease in temperature

The distribution of values is the way data points are spread out across the dataset.

collection. Summary statistics help to understand the trends, outliers, and distribution of values in a data set.

4.2.8 Requirement of Summary Statistics

The summary statistics provide a quick overview of characteristics of data. It leads towards a better understanding of data cleaning, data preprocessing, feature selection and data visualization.

4.3 Big Data

Big data contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Big data is larger, more complex datasets, especially *from* new data sources. These data sets are so voluminous that traditional data processing software cannot manage them. These massive volumes of data can be used to address business problems which were difficult to handle before.

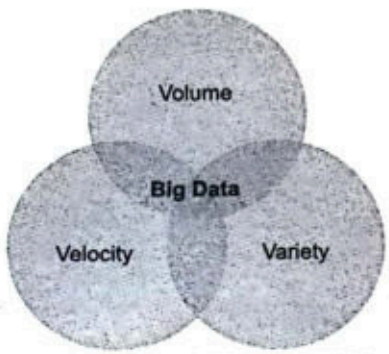


Fig. 4.5 Big Data

The three Vs of big data are:

Volume: It refers to the amount of data. Big data deals with huge volumes of low-density, unstructured data. The size/volume of data may vary from system to system. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

Velocity: It refers to the speed of data. Velocity is the fast rate at which data is received. Normally, the highest velocity of data streams directly into memory rather than being written to disk. Some internet-enabled smart products operate in real-time and will require real-time evaluation and action.

Variety: It refers to the various formats and types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new data types. These unstructured data (text, images, videos) and semi structured data (JSON, XML) types require additional preprocessing to derive meaningful insight.

4.3.1 The history of big data

The term big data emerged in the early 2000s as a term to describe exponential growth of data. Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. In 2005 a tool called Hadoop (an open-source framework created specifically to store and analyze big datasets) was developed, which helped store and manage huge data.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data. The analysis of this huge data provides business insight for optimized decision making.

4.3.2 Advantages and benefits of big data

Big data contains more information therefore it helps individuals, organizations, and businesses to optimize and generate cost effective solutions. Big data has many advantages for the betterment and progress of business, some of them are as follows:

Product development: Developing and creating new products, services or brands is much easier when based on data collected from customers' needs and wants. Companies use big data to anticipate customer demand. They build predictive models for new products and services by classifying key attributes of past and current products.

Predictive maintenance: It is a proactive maintenance strategy that uses the analysis of existing data to predict when equipment, machinery or product is likely to fail. Therefore, it indicates the potential issues before the problems happen.

Customer experience/satisfaction: A clearer view of customer experience is more possible now than ever before. Big data enables the businesses to gather data from social media, web visits, call logs, and other sources to improve customer satisfaction.

Fraud and compliance: Big data analytics can identify and detect unusual suspicious patterns and anomalies. As a result provides an effective tool to detect fraudulent activities and enhance cybersecurity measures.

4.3.3 Big data challenges

Since there are many advantages of big data, businesses encounter many challenges of big data. Some of them are as follows:

i. **Data Quality:** Poor quality of data may lead to errors, inefficiency, and misleading insight after data analysis.
ii. **Data Security and privacy:** It is difficult to manage the protection and privacy of massive datasets to prevent unauthorized access.
iii. **Rapid growth of data:** Making systems that can handle more and more data as it keeps on growing without slowing down is challenging.
iv. **Big data tool selection:** Ensuring compatibility and seamless interaction between different big data tools and platforms.
v. **Data integration:** To create harmony among diverse data formats and structures is a difficult task.

4.3.4 Application of big data in business

Big data applications can help companies to make better business decisions by analyzing large volumes of data and discovering hidden patterns. The following are a few business

domains where big data can be applied:

1. Healthcare
2. Media and Entertainment
3. IoT
4. Manufacturing
5. Government

Health care

Big data is making a major impact on the huge healthcare industry. Wearable devices and sensors collect patient data which is then fed in real-time to an individual's electronic health records. Healthcare providers are now using big data to predict epidemic outbreaks, real-time alerting, predict and prevent serious medical conditions etc. Researchers analyze the data to determine the best treatment for a particular disease, side effects of the drugs, forecasting the health risks, etc.

Media and entertainment

The media and entertainment industries are creating, advertising, and distributing their content using new business models. The media houses are targeting audiences by predicting what they would like to see, how to target the ads, content monetization, etc. Big data systems are thus increasing the revenues of such media houses by analyzing viewer patterns.

Internet of Things (IoT)

Big data plays an important role in enhancing the capabilities of IoT devices. IoT devices generate continuous data. The analytics based on this huge data helps in personalized customer experience. In brief, big data is essential for unlocking the full potential of IoT by providing meaningful insight derived from the massive amount of data generated by IoT devices.

DO YOU KNOW?
IoT (Internet of Things)
It is the network of physical devices, home appliances or other gadgets which allow to collect and exchange data with other devices over the internet.

Manufacturing

Big data helps the manufacturing companies to make better products and smarter decisions. It helps in predicting when machines might need a break (predictive maintenance), making sure they don't unexpectedly stop working. Big data also looks at how products are made better and cheaper. It is like having a smart assistant that guides the whole manufacturing process, making things more efficient and helping companies build the best products. The following are some of the major advantages of employing big data applications in manufacturing industries:

- High product quality
- Tracking faults
- Supply planning

- Predicting the output
- Increasing energy efficiency
- Testing and simulation of new manufacturing process
- Large scale customization of manufacturing

Government

Analytics through big data management techniques allows governments to understand the needs of their citizens, combat fraud, minimize system errors and improve operations, reducing costs and improving the services of any government entity. By adopting big data systems, the government can attain efficiency in terms of cost, output, and novelty.

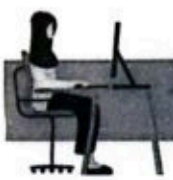
Big data applications can be applied in each and everywhere big data finds applications include:

- Agriculture
- Aviation
- Cyber security and intelligence
- Crime prediction and prevention
- E-commerce
- Fake news detection
- Fraud detection
- Pharmaceutical drug evaluation
- Scientific research
- Weather forecasting
- Tax compliance

Summary

- Data is a collection of information or facts that we gather about something. It can be represented by numbers, measurements, descriptions, sounds or pictures.
- Data Analytics refers to the process of carefully examining and studying data to identify patterns, draw conclusions, or make the data meaningful.
- Data Science refers to an interdisciplinary field of multiple disciplines that uses mathematics, statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.
- Dataset is a structured or processed collection of data usually associated with a unique body of work.
- Machine learning is a branch of Artificial Intelligence and computer science which emphasizes the use of data and algorithms to imitate human learning by the computers.
- Deep learning is the subset of Machine learning, with emphasis on the simulation or imitation of human brain's behavior by using artificial neural networks.
- Data mining is the subset of data science which primarily focuses on discovering patterns and relationships in existing datasets.
- Data visualization is the graphical representation of data using common charts, plots, infographics, and animations.
- Predictive analysis is the use of data to predict future trends and events based on historical data.
- NLP is the study of interaction between human language and computers. The common uses of NLP are chatbots, language translators and sentiment analysis.
- Qualitative or Categorical data describes an object or a group of objects that can be labeled according to some group or category.
- Ordinal data sees a specific order or ranking, it uses certain scale or measure to group data into categories. Such as in test grades, economic status, or military rank.
- Nominal data does not have any order, it can be labelled into mutually exclusive categories, which cannot be ordered meaningfully.
- Quantitative or Numerical data deals with numeric values, that can be computed mathematically to draw some conclusions.
- Discrete data includes data which can only take certain values and cannot be further subdivided into smaller units.
- Continuous data refers to the unspecified number of possible measurements between two realistic points or numbers.
- Primary data can also be collected through experiments and recording observations. secondary data is collected from some previously recorded from primary data. Following are some sources of data:

- Primary data collection methods involve the collection of original data directly from the data source or via direct interaction with the respondent.
- Secondary data collection methods involve data collection using existing data collected by someone else for some purpose.
- Data visualization is graphical representation of data to get meaningful insight, trends, and patterns from data.
- Big data contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. The three Vs of big data are Volume, Velocity and Variety.
- Hadoop In 2005 a tool called was developed, which helped store and manage huge data.



Exercise

Q1. Select the suitable answer for the following Multiple-choice questions.

i. _____ is a structured or processed collection of data usually associated with a unique body of work.

a) Database b) Dataset
c) Data and Information d) Information

ii. _____ refers to the process of carefully examining and studying data to identify patterns, draw conclusions, or make the data meaningful.

a) Data analytics b) Data Predictions
c) Dataset d) Database

iii. _____ is the graphical representation of data through use of common charts, plots, infographics, and animations.

a. Data cleaning b. Missing values
c. Data visualization d. Data hiding

iv. _____ is subset of Machine learning, with emphasis on the simulation or imitation of human brain's behavior by using artificial neural networks.

a. Data visualization b. Computer vision
c. Deep learning d. Big Data

v. _____ is the use of data to predict future trends and events based on historical data.

a. Statistical analysis b. Predictive analysis
c. Graphical analysis d. Deep learning

vi. _____ is the fast rate at which data is received and acted on.

a. Volume b. Velocity

c. Variety d. Vision

vii. _____ includes the data which can only take certain values and cannot be further subdivided into smaller units.

a. Discrete data b. Continuous data

c. Ordinal data d. Referral data

viii. _____ is limitation of big data.

a) Statistical data b) Unlimited growth of data

c) Data visualization d) Predictive maintenance

ix. Customer satisfaction level such as satisfied, dissatisfied, and neutral are examples of _____ data type.

a) Ordinal data b) Continuous data

c) Numeric data d) Discrete data

x. _____ is a method of collecting information from individuals.

a. Survey b. Data hiding

c. Data visualization d. Data finding

II. Give Short answers to the following short response questions (SRQs).

1) Define data analytics and data science. Are they similar or different? Give reason.
2) Can you relate how data science is helpful in solving business problems?
3) Database is useful in the field of data science. Defend this statement.
4) Compare machine learning and deep learning, in the context of formal & informal education.
5) What is meant by sources of data? Give three sources of data excluding those mentioned in the book.
6) Differentiate between database and dataset.
7) Argue about the trends, outliers, and distribution of values in a data set?
8) Why are summary statistics needed?
9) Express big data in your own words. Explain three V's of big data with reference to email data. (Hint: An email box that contains hundreds of emails)
10) Illustrate the purpose of data storage?

III. Give Long answers to the following extended response questions (ERQs).

1) Sketch the key concepts of data science in your own words.
2) Develop your own thinking on the various data types used in data science.

3) Compare how big data is applicable to various fields of life. Illustrate your answer with suitable examples.
4) Relate the advantages and challenges of big data?
5) Design a case study about how data science and big data has revolutionized the field of healthcare.

Lab Activities

Activity 1

Conduct a survey:

Objective: To help students understand the process of data collection through survey and questionnaires.

i. Provide the class with a sample questionnaire about the sentiments analysis of small enterprises. Sample questions may include, why did you choose this business/occupation? How do you manage to meet the market competition? What limitations do you feel about your business? What short term and long-term goals/objectives have you set regarding your business?
ii. Ask the students to modify the questionnaire according to their favorite business/occupation. Help the students in modifying the questionnaire.
iii. Arrange a field trip or ask the students to get it filled by asking questions from relevant shopkeepers around them.
iv. Let the students arrange the collected data and organize it for the next activity.

Outcome: This activity will help students understand various businesses/occupations around them as well as understand the opportunities and shortcomings regarding various businesses.

Activity 2

Data Visualization:

Objective: To help students understand the process of data visualization.

I. Ask the class to make groups according to the professions about which they conducted survey. Group similar professions according to their nature.
ii. Ask the students to plot the data by using charts and graphs. They can use software tools if available.
iii. Let the students draw conclusions about various professions and what are the sentiments of business owners about their businesses.
iv. Ask the students to add modifications to the current businesses model to overcome the existing shortcomings.

Outcome: This activity will help students understand data visualization and how it is helpful for decision making to improve it and make new strategies or modifying the existing one.

Activity 3

Data collection about various businesses:

Objective: To help students understand the importance of data collection to get insight about data.

v. Divide the class into small groups (5 to 6 students) and assign each a business topic, preferably from their own surroundings, so that students have appropriate information about it.
vi. Instruct each group to write a case study, that how that business can be related to big data.
vii. Each group will present their case study and other groups can ask questions about it, try to keep it a supervised interactive session.

Outcome: This activity will help students understand the relationship and impact of big data on various businesses.

Teacher's Guide

Learning Data Science Concepts through games (https://www.iggi-phd.org/themes/game-data)

Teacher's Guide

'Our World in Data' put emphasis on Coronavirus data, employing interactive charts and graphs to illustrate the rapid speed at which data is gathered, particularly in scenarios like a pandemic. (https://ourworldindata.org/coronavirus)