UNIT 1 – INTRODUCTION TO BIG DATA

QUESTION BANK

PART A (2 Marks each)

1. What you mean by Big Data? Imp
2. What are the different sources of big data?
3. Explain different applications of Big Data. Imp
4. Write down the nature of data. Imp
5. What are the different stages in Intelligent Data Analysis?
6. Distinguish between Analysis vs Reporting. Imp
7. What you mean by statistical distributions?
8. What you mean by re-sampling?
9. Define prediction error.

PART B (5 Marks each)

10. Explain 5V's of Big Data.
11. Explain the challenges of conventional systems.
12. Explain Intelligent Data Analysis. Imp
13. Explain analytical process in detail.

PART C (15 Marks each)

14. Explain in detail nature of data.
15. Explain Intelligent Data Analysis. Imp
16. Explain Modern Data Analytical Tools. Imp

NOTES

INTRODUCTION TO BIG DATA PLATFORM

- Big Data is **high-volume**, **high-velocity** and/or **high-variety** information asset that requires new forms of processing for enhanced decision making, insight discovery and process optimization.
- A collection of data sets so large or complex that traditional data processing applications are inadequate.
- Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.
- A **single mobile phone user** will generate about **40 exabytes** of data on every month.
- This **massive amount of data** is termed as **Big data**.
- Hence these are a large amount of data.
- To classify any data as Big data, this is possible with the the concept of **5v.**
    1. **Volume**
    2. **Velocity**
    3. **Variety**

4. **Veracity**
5. **Value**

- Consider the following example of Health care industry.
  - **Volume**:
    - High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation, and management processes.
    - Hospitals and clinics across the world generate massive volumes of data.
    - 2314 Exabytes of data are collected annually.
  - **Velocity**:
    - In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time.
    - These data are patient records and test results.
    - All this data is generated at a very high speed. Which attributes to the velocity of big data.
  - **Variety**:
    -
    - Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions.
    - Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.
    - It referred to the various data types.
      - Structured data: Excel Records.
      - Semi-structured data: Log files
      - Un-structured data: X-ray Images.
  - **Veracity**:
    - Veracity refers to the quality or fidelity of data.
    - Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.
    - Accuracy and trustworthiness of the generated data termed as veracity.
    - Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information.
  - **Value**:
    - Value is defined as the usefulness of data for an enterprise.
    - Analysing all these data will benefit the medical sector by,
      - Faster disease detection
      - Better treatment
      - Reduced cost
    - These are known as the value of big data.
- To **store and process big data**, various **frameworks** are used.

  - **Cassandra**
  - **Hadoop**
  - **Spark**
- Big data is analysed for numerous **applications** in games like:
  - **HALO 3**
  - **CALL OF DUTY**
- Designers analyse users' data to understand at which stage users pause, restart, quit playing.
- This insight can help them to rework on the game and improve the user experience.
- Also, big data also helped with disaster management during hurricane in 2012 in USA and necessary measures were taken.

TYPES/SOURCES OF BIG DATA

- It is suggested by IBM and the Big Data task team:
  - **Social networks and web data**: such as Facebook, Twitter, e-mails, blogs, and YouTube.
  - **Transactions data and Business Processes data**: such as credit card transactions, flight bookings, etc. and public agencies data such as medical records, insurance business data, etc.
  - **Customer master data**: such as data for facial recognition and for the name, date of birth, marriage anniversary, gender, location and income category.
  - **Machine-generated data**: such as machine-to-machine or Internet of Things (IOT) data, and the data from sensors, trackers, web logs and computer systems log. Computer generated data is also considered as machine generated data from data stores. Usage of programs for processing of data using data repositories, such as database or file, generates data and also machine generated data.
  - **Human-generated data**: such as biometrics data, human-machine interaction data, e-mail records with a mail server and MySql database of student grades.

CHALLENGES OF CONVENTIONAL SYSTEMS

- Big data is huge amount of data, hence conventional systems cannot store, manage and analyse within a time interval.
  1. Data is collected in **large quantities** and it is **not possible** to process everything.
  2. Data must be **meaningful** and **collected in real time.** Where meaningful data means a data **without irregularity, mistakes** and **inconsistency**. Realtime data means the data **can't** be **old** or **outdated**.
  3. Data is **collected** from **multiple sources** (text, audio, video etc) and must be **categorized**. This is an important aspect of data collection. Doing manually, it is impossible for humans, because it is extremely challenging, time consuming, requires a lot of manpower.

4. **Collect correct data**. By eliminating flaws, mistakes, incompleteness, inconsistency, irregularity from data. Because wrong data will produce problems in future.
5. **Comparison of data using multiple tools**. The collected data must be represented in the form of **graphs, charts, statistics** etc.
6. The **data** must be **accessible** to the **respective person**.
7. Lack of knowledgeable professionals who can handle and deals with diverse and large data.

- To overcome these challenges and drawbacks, intelligent data analysis is used.
- The valuable information can be gathered with the help of machines, thus **reducing processing time, cost** and **errors**.

INTELLIGENT DATA ANALYSIS (IDA)

- Intelligent Data Analysis (IDA) discloses hidden facts that are not known previously and provides potentially important information or facts from large quantities of data.
- It also helps in making a decision.
- IDA helps to obtain useful information, necessary data and interesting models from a lot of data available online in order to make the right choices.
- The **main goal** of intelligent data analysis is to **obtain knowledge**.
- Data analysis is the **process** of a combination of **extracting data from data set**, **analysing**, **classification** of data, **organizing**, **reasoning** and so on.
- **Intelligent data analysis** is the combination of **advanced statistical techniques**, **human intuition** and **serious computing power** to address real world data intensive problems.
- **Analysis** is a **scientific process** to **discover the meaningful patterns** and **structures hidden** within the **mountain** of available **data** and **transform** this data into **information** for **better decisions**.
- **IDA** is one of the **major issues** in **Artificial Intelligence (AI)** and information.
- Intelligent data analysis **discloses hidden facts** that are **not known previously** and **provides** potentially **important information** or **facts from large quantities of data**.
- IDA **helps** to **obtain useful information**, **necessary data** and **interesting models**.
- In general, **IDA includes three stages**:
  1. **Preparation of data**
  2. **Data mining**
  3. **Data validation and explanation**

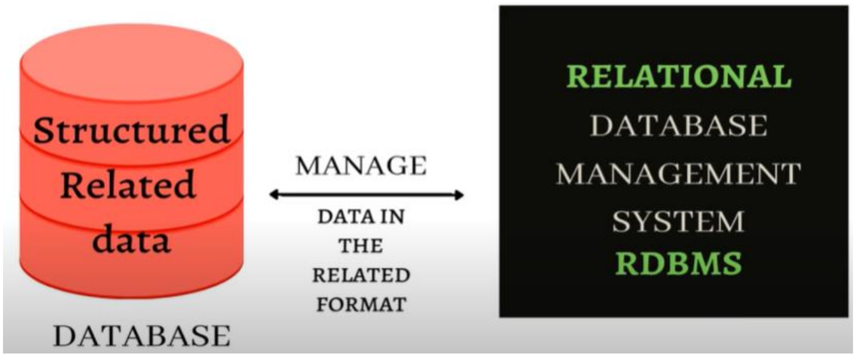
NATURE OF DATA

- **Data** are **raw facts** that have **not** been **processed** to explain their meaning.
- Data are stored in the **database** and **data base management system** manages data. i.e., it **stores**, **update** and **retrieve** from the database.
- There are **3 types of data**:
  1. **Structured data**
  2. **Semi-structured data**

3. **Unstructured data**

STRUCTURED DATA

- **Stored** in **tabular format**.
- i.e., in the **form of rows and columns**.
- Structured data **clearly defined** and **data is stored** in a **pre-defined data model**.
- Ex: **Excel files, SQL data bases**
- Data are stored in rows and columns are **related to each other.**
- Hence get a **proper view** and **understanding of data.**
- Real life example:



- Structured data are **stored** in **Relational Databases**.

UN- STRUCTURED DATA

- **No predefined structure**.
- **No data model**
- Data is **irregular and ambiguous**.
- Ex: text, numbers, images, audios, videos, messages, social media post etc.
- **Easy to extract data**.
- **80- 90%** of data are **unstructured data**.
- Real life example:
- **Face book, Instagram** & **YouTube** are unstructured data.
- It is complex task to analyse such data. hence **Artificial Intelligence** is used.
- Ex: **Face recognition by google**.
- Previously, only structured data was used extensively. But with the help of Artificial Intelligence, unstructured data are commonly used.
- So, unstructured data is the most useful kind of data. & It provides a lot of information.

SEMI-STRUCTURED DATA

- It falls **between structured and unstructured data**.
- It is a **combination of both**.
- Ex: **Emails, XML, WWW**.

ANALYTIC PROCESS

- The steps involved in data analytic process are:

1. Collecting data
2. Cleaning data
3. Manipulating data
4. Analysing data
5. Visualizing data

- Ex: Travel industry:

Collecting Data

- ✓ If one person is travelling to Delhi, he uses one of the Aviation website, provides basic details like destination, date of travel, price etc., He select one with his budget & make the payment confirm. These data are collected by travel company.
- ✓ Similarly, many people do the same thing, then it generates a lot of data.
- ✓ These data are stored in their web servers in tabular format. Hence it is easier for analyst to analyse this data

Cleaning data

- ✓ If there are some missing values or un-structed data in tabular format, by replacing missing values or deleting that row is called cleaning data.
- ✓ Now the data is clean and ready for analyse.

Manipulating Data

- ✓ Manipulates the data to create required features and variables.
- ✓ Ex: if the analyst adds new columns like return date etc.

Analysing Data

- ✓ Data analyse using logical methods and analytical techniques.
- ✓ Once it's ready, advanced analytics processes can turn big data into big insights.
- ✓ Some of these big data analysis methods include:

  ● **Data mining** sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.

  ● **Predictive analytics** uses an organization's historical data to make predictions about the future, identifying upcoming risks and opportunities.

  ● **Deep learning** imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

Visualizing Data

- ✓ Present data after complete analysis on data.
- ✓ Visualization means showing analysed data in visual or graphical format for easy interpretation.

DATA ANALYSIS TOOLS

- Data analysis tools provide the analysed results visually.
- They are:

- **Hadoop:**
  - ✓ It is an open-source framework that efficiently stores and processes big datasets on clusters of commodity hardware.
  - ✓ This framework is free and can handle large amounts of structured and unstructured data, making it a valuable mainstay for any big data operation.
- **Microsoft Excel** :
  - ✓ Developed by Microsoft.
  - ✓ It is a spread sheet program, used to create grid of numbers, texts and various formulas.
  - ✓ Easy to use & widely used tool.
  - ✓ Excel works with other office software. i.e., Excel spreadsheets can be easily added to Word document and Power point presentations.
  - ✓ The biggest benefits of Excel are ability to organize large amounts of data into orderly logical spreadsheets and charts.
- **RapidMiner:**
  - ✓ It is a data science software platform which helps with data presentation and analysis.
  - ✓ It is an integrated environment for:
    - Data preparation
    - Analysis
    - Machine learning
    - Deep learning
  - ✓ It is widely used in every business and commercial sector.
  - ✓ It has data exploration features such as:
    - Graphs
    - Descriptive statistics
    - Visualization which allows users to get valuable insights.
  - ✓ It has more than 1500 operators for data transformation and analysis tasks.
- **Talend:**
  - ✓ It is an open-source software platform which offers data integration and management.
  - ✓ It specializes in big data integration.
  - ✓ It is also available in open-source and premium versions.
  - ✓ It is one of the best tools for cloud computing and big data integration.
- **KNIME:**
  - ✓ It is a free and open-source data analysis tool to create data science applications and build machine learning models.
  - ✓ It is an analysing, reporting and integration platform.
  - ✓ KNIME has been used in pharmaceutical research and customer data analysis, business intelligence, text mining & financial data analysis.
  - ✓ It provides interactive graphical user interface to create visual workflows.

- Other tools are:
  - SAS (Statistical Analysis System)
  - R and Python
  - Apache spark
  - Power BI
  - Tableau

ANALYSIS VS REPORTING

REPORTING

- It is the process of organizing data in the form of graphs and charts.
- Reporting is used to provide facts, which can use to draw conclusions, avoid problems or create plans.
- Reporting presents the actual data to end-users, after collecting, sorting and summarizing it to make it easy to understand.
- Reporting offers no judgment or insight.
- It focuses on what is happening.
- High-level overview of data.

ANALYSING

- It is the process of exploring data in order to extract a meaningful insight.
- Analytics offers pre-analysed conclusions that a company can use to solve problems and improve its performance.
- Analytics doesn't present the data but instead draws information from the available data and uses it to generate insights, forecasts and recommended actions.
- It focuses on why is something happening.
- Data analytics focuses on "why" something is happening within an organization.
- Interpret data at a deeper level.

STATISTICAL CONCEPTS

- Statistics is a applied mathematics were we collect, organize, analyse and interpret numerical facts.
- Statistical methods are the concepts models and formulas of mathematics used in the Statistical analysis of data.
- It is the science of collecting, exploring and presenting large amounts of data to identify patterns and trends.
- It is also called **quantitative analysis**.

  ### SAMPLING DISTRIBUTIONS

  - In statistics, a population is the entire pool from which a statistical sample is drawn.
  - A population may refer to an entire group of people, objects, events, hospital visits, or measurements.
  - A population can thus be said to be an aggregate observation of subjects grouped together by a common feature.

- A lot of data drawn and used by academicians, statisticians, researchers, marketers, analysts, etc. are actually samples, not populations.
- A sample is a subset of a population.
- A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.
- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.
- Sampling Distribution is a statistic that aims to guess a large number of samples obtained from a specific group repeatedly.
- In statistics, the probability is used for calculating the likely occurrence of a phenomenon.
- This is done by collecting samples from populations.
- A lot of data that is collected over time, aim to calculate the probabilities of an event.
- This data is collected with utmost precision.
- Sampling distribution involves more than one statistical value of a sample.
- The primary purpose of Sampling Distribution is to establish representative results of small samples of a comparatively larger population.
- The significance of sampling distribution :
  - ✓ It provides accuracy.
  - ✓ Provides consistency.

RE-SAMPLING

- The problem with the sampling process is that we only have a single estimate of the population parameter, with little idea of the variability or uncertainty in the estimate.
- One way to address this is by estimating the population parameter multiple times from our data sample. This is called resampling.
- Re-sampling is the method that consists of creating or drawing repeated samples from the original samples.
- Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of a sample drawn has a number of cases that are similar to the original data sample.

STATISTICAL INFERENCE

- Statistical inference is the process of analysing the result and making conclusions from data.
- It is also called **inferential statistics**.
- Statistical inference is a method of making decisions about the parameters of a population, based on random sampling.
- It helps to assess the relationship between the dependent and independent variables.

- different types of statistical inferences are:
  - ✓ Pearson Correlation
  - ✓ Bi-variate regression
  - ✓ Multi-variate regression
  - ✓ Chi-square statistics and contingency table
  - ✓ ANOVA or T-test
- The statistical inference has a wide range of application in different fields, such as:
  - ✓ Business Analysis
  - ✓ Artificial Intelligence
  - ✓ Financial Analysis
  - ✓ Fraud Detection
  - ✓ Machine Learning etc.

PREDICTION ERROR

- Predictive analytical processes use new and historical data to forecast activity, behaviour, and trends.
- A prediction error is the failure of some expected event to occur.
- When prediction fails, humans can use different methods, examining predictions and failures and deciding some methods to overcome such errors in the future.
- Applying that type of knowledge can inform decisions and improve the quality of future prediction.