UNIT I INTRODUCTION TO BIG DATA

Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis –Nature of Data - Analytic Processes and Tools - Analysis Vs Reporting - Modern Data Analytic Tools- Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

INTRODUCTION TO BIG DATA PLATFORM

What is Big Data?

Big data can be defined as a concept used to describe a large volume of data, which are both structured and unstructured, and that gets increased day by day by any system or business. However, it is not the quantity of data, which is essential. The important part is what any firm or organization can do with the data that matters a lot. Analysis can be performed on big data for insight and predictions, which can lead to a better decision and reliable strategy in business moves.

What Are the 3Vs of Big Data?

This conception theory gained thrust in the early 2000s when trade and business analyst Mr. Doug Laney expressed the mainstream explanation of the keyword big data over the pillars of 3v's:

Volume: Organizations and firms gather as well as pull together different data from different sources, which includes business transactions and data, data from social media, login data, as well as information from the sensor as well as machine-to-machine data. Earlier, this data storage would have been an issue - but because of the advent of new technologies for handling extensive data with tools like Apache Spark, Hadoop, the burden of enormous data has decreased.

Velocity: Data is now streaming at an exceptional speed, which has to be dealt with suitably. Sensors, smart metering, user data as well as RFID tags are lashing the need for dealing with an inundation of data in near real-time.

Variety: The releases of data from various systems have diverse types and formats. They range from structured to unstructured, numeric data of traditional databases to non-numeric or text documents, emails, audios and videos, stock ticker data, login data, Blockchains' encrypted data, or even financial transactions.

Importance of Big Data

Big Data does not take care of how much data is there, but how it can be used. Data can be taken from various sources for analyzing it and finding answers which enable:

- **Reduction in cost.**
- **Time reductions.**
- **New product development with optimized offers.**
- **Well-groomed decision making.**

When you merge big data with high-powered data analytics, it is possible to achieve business-related tasks like:

- **Real-time determination of core causes of failures, problems, or faults.**
- **Produce tokens and coupons as per the customer's buying behavior.**
- **Risk-management can be done in minutes by calculating risk portfolios.**
- **Detection of deceptive behavior before its influence.**

Things That Comes Under Big Data (Examples of Big Data)

As you know, the concept of big data is a clustered management of different forms of data generated by various devices (Android, iOS, etc.), applications (music apps, web apps, game apps, etc.), or actions (searching through SE, navigating through similar types of web pages, etc.). Here is the list of some commonly found fields of data that come under the umbrella of Big Data:

Black Box Data: Black box data is a type of data that is collected from private and government helicopters, airplanes, and jets. This data includes the capture of Flight Crew Sounds, separate recording of the microphone as well as earphones, etc. Stock Exchange Data: Stock exchange data includes various data prepared about 'purchase' and 'selling' of different raw and well-made decisions.

Social Media Data: This type of data contains information about social media activities that include posts submitted by millions of people worldwide.

Transport Data: Transport data includes vehicle models, capacity, distance (from source to destination), and the availability of different vehicles.

Search Engine Data: Retrieve a wide variety of unprocessed information that is stored in SE databases.

There are various other types of data that get generated in bulk from applications and organizations.

Types of Big Data (Types of Data Handled by Big Data)

The data generated in bulk amount with high velocity can be categorized as:

1. Structured Data: These are relational data.
2. Semi-structured Data: example: XML, JSON data.
3. Unstructured Data: Data of different formats: document files, multimedia files, images, backup files, etc.

Big Data Technologies

This technology is significant for presenting a more precise analysis that leads the business analyst to highly accurate decision-making, ensuring more considerable operational efficiencies by reducing costs and trade risks. Now to implement such analytics and hold such a wide variety of data, one must need an infrastructure that can facilitate and manage and process huge data volumes in real-time. This way, big data is classified into two subcategories:

Operational Big Data: comprises data on systems such as MongoDB, Apache Cassandra, or CouchDB, which offer equipped capabilities in real-time for large data operations.

Analytical Big Data: comprises systems such as MapReduce, BigQuery, Apache Spark, or Massively Parallel Processing (MPP) database, which offer analytical competence to process complex analysis on large datasets.

Challenges of Big Data

Rapid Data Growth: The growth velocity at such a high rate creates a problem to look for insights using it. There is no 100% efficient way to filter out relevant data.

Storage: The generation of such a massive amount of data needs space for storage, and organizations face challenges to handle such extensive data without suitable tools and technologies.

Unreliable Data: It cannot be guaranteed that the big data collected and analyzed are totally (100%) accurate. Redundant data, contradicting data, or incomplete data are challenges that remain within it.

Data Security: Firms and organizations storing such massive data (of users) can be a target of cybercriminals, and there is a risk of data getting stolen. Hence, encrypting such colossal data is also a challenge for firms and organizations.
