Unit-1

What is Big Data?

- Big Data is also a data but with huge size.
- Big Data is a term used to describe a collection of data that is **huge in volume and yet growing** exponentially with time.
- As a growing data it become so large and complex, so the traditional database management tools are unable to handle that kind of data.
- Here are some examples as who generates big data: Social Media, Big Companies like Flipkart and Amazon, Stock Exchange data, etc.
- Big Data can be found in 3 Types:
  1. **Structured**: Data is in a **structured format or fixed format**. We can also say that it can be present in a table in a row and column format.

  2. **Unstructured**: Any data with **unknown form or the structured** is classified as unstructured data.
     Ex- simple text file, video, image, audio, etc.

  3. **Semi-Structured**: Semi-Structure data can contain both forms of data.

We can see semi-structured data as structured but not in form.
Ex- Personal data stored in XML file.

Characteristics Of Big Data:

1. **Volume:**
   - The name big data itself related to the **size**.
   - Size of data plays a very crucial role in determining value out of data.
   - ‘Volume’ is the one characteristic which need to be considered while dealing with Big Data.
2. **Variety:**
   - Variety refers to **heterogeneous sources** and the nature of data, both structured and unstructured.
   - As simple, nowadays data can be received from emails, photos, videos, PDF, etc.
   - This variety of unstructured data poses issues with storage, mining and analyzing data.
3. **Velocity:**
   - The term velocity refers to the **speed of generation of data**. How fast is data generated and processed.
   - Big Data velocity deals with speed at which data flows in from sources like business processes, networks, social media, etc.
   - The flow of data is **massive and continuous**.
4. **Variability:**

- This refers to the **inconsistency** which can be shown by the data at times.
- Size of data plays a very crucial role in determining value out of data.

5. **Value:**
   - Value is an essential characteristic of big data.
   - It is not the data we process or store.
   - It is valuable and reliable data we store, process and also analyze.

Why Big Data Is Important?

- It is used to analysis of customer or business data by big companies to make a strategy or business-related decision.
- The company can take data from any source and analyze it to find answers which will enable.
  1. Cost Saving
  2. Time Reduction
  3. Understanding the market condition
  4. Control online reputation
  5. Better decision making

Enabling Technologies in Big-Data?

1. Predictive Analysis:
   - One of the prime tools for business to avoid risks in decision making predictive analytics can help business.

2. NoSQL Database:
   - These databases are utilized for reliable and efficient data management across a scalable number of storage nodes.
3. Distributed Storage
   - Distributed storage is a storage system that distributes data across multiple servers and locations, often in a decentralized manner.

4. Data Virtualization:
   - It enables application to retrieve data without implementing technical restriction such as data format, physical location of data.
5. Data Quality:
   - As important parameter for big processing is data quality.

Some other technologies:

- Stream Analytics
- Data Integration
- Knowledge discovery tool

Application Of Big-Data

I) Government Sector
II) Social Media Analytics
III) Fraud Detection
IV) Banking

V) Recommendation
VI) IoT
VII) Smart Traffic System
VIII) Education Sector
IX) Smart Phones
X) Marketing
XI) Healthcare
XII) Agriculture

Big Data Distribution Package:

A **big data distribution package** refers to a comprehensive set of tools, frameworks, and technologies designed to facilitate the storage, processing, and analysis of large volumes of data across distributed computing environments. These packages enable organizations to efficiently handle the challenges posed by big data, such as volume, velocity, variety, and veracity.

Key Components

1. **Data Storage:**
   a. **Distributed File Systems**: Technologies like **Hadoop Distributed File System (HDFS)** allow for the storage of massive datasets across multiple nodes, ensuring high availability and fault tolerance.
   b. **NoSQL Databases**: Systems like **Apache HBase** and **Cassandra** provide flexible schemas and can handle structured and unstructured data efficiently.

2. **Data Processing Frameworks:**
   a. **Batch Processing**: Frameworks like **Apache Hadoop** facilitate batch processing of large datasets using MapReduce.
   b. **Stream Processing**: Technologies like **Apache Spark** and **Apache Flink** support real-time data processing and analytics.

3. **Data Ingestion and Integration:**
   a. Tools such as **Apache Kafka** and **Apache NiFi** enable seamless data ingestion from various sources and integrate different data streams for processing.
4. **Data Management and Orchestration:**
   a. **Apache Airflow** and **Apache Zookeeper** help manage data workflows, scheduling, and coordination of distributed systems.
5. **Data Visualization and Reporting:**
   a. Tools like **Tableau** and **Apache Superset** assist in visualizing data insights and creating dashboards for better decision-making.
6. **Cloud Services:**
   a. Cloud-based solutions like **Amazon EMR**, **Microsoft Azure HDInsight**, and **Google Cloud BigQuery** provide scalable environments for big data processing, reducing the need for on-premises infrastructure.