**1. Introduction to Big Data – Sital Sharma**

**Definition**

* **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
* The data is characterized by the **5 V's**:
  + **Volume**: The quantity of data generated.
  + **Velocity**: The speed at which data is generated and processed.
  + **Variety**: The different types of data (structured, semi-structured, unstructured).
  + **Veracity**: The quality and accuracy of data.
  + **Value**: The potential of data to be turned into insights.

**2. Evolution of Big Data**

**Key Milestones**

* **1970s-1980s**: Emergence of relational databases.
* **1990s**: Growth of the internet and data generation.
* **2000s**: Explosion of social media and mobile devices.
* **2010s**: Advancements in cloud computing and data analytics.

**Structuring of Big Data**

* **Structured Data**: Organized data in fixed fields (e.g., databases).
* **Semi-Structured Data**: Does not conform to a strict structure (e.g., XML, JSON).
* **Unstructured Data**: No predefined format (e.g., text, videos, social media posts).

**Elements of Big Data**

* **Data Sources**: Various origins of data (social media, IoT devices, transactions).
* **Data Storage**: Methods for storing vast amounts of data (databases, data lakes).
* **Data Processing**: Techniques for processing large data sets (batch processing, real-time processing).
* **Data Analysis**: Extracting meaningful insights from data using analytics tools.
* **Data Visualization**: Representing data in visual formats (charts, graphs).

**3. Big Data Analytics**

**Importance**

* Provides actionable insights for decision-making.
* Enhances customer experience and business operations.
* Facilitates predictive analytics and trend analysis.

**Techniques**

* **Descriptive Analytics**: Summarizes past data.
* **Predictive Analytics**: Uses historical data to predict future events.
* **Prescriptive Analytics**: Suggests actions based on data analysis.

**4. Distributed and Parallel Computing for Big Data**

**Distributed Computing**

* Involves multiple computers working together to solve a problem.
* Data and tasks are distributed across different nodes in a network.

**Parallel Computing**

* Simultaneous data processing using multiple processors.
* Improves processing speed and efficiency.

**5. Hadoop**

**Overview**

* An open-source framework for processing large data sets across clusters of computers using simple programming models.

**Components**

* **Hadoop Distributed File System (HDFS)**: Storage system for large data sets.
* **MapReduce**: Programming model for processing large data sets with a distributed algorithm.

**Ecosystem**

* **Hive**: Data warehousing and SQL-like query language.
* **Pig**: High-level platform for creating MapReduce programs.
* **HBase**: Distributed, scalable, big data store.
* **YARN**: Resource management and job scheduling.

**6. Cloud Computing and Big Data**

**Integration**

* Cloud computing provides scalable resources for storing and processing big data.
* Allows for flexible, cost-effective, and scalable big data solutions.

**Benefits**

* **Scalability**: Easy to scale resources up or down.
* **Cost-Efficiency**: Pay-as-you-go model reduces costs.
* **Accessibility**: Data and applications accessible from anywhere.

**7. In-Memory Computing Technology for Big Data**

**Definition**

* Stores data in the main memory (RAM) of the server to provide faster data access and processing.

**Advantages**

* Reduces latency and improves performance.
* Enables real-time data processing and analytics.

**8. Big Data Stack**

**Layers**

* **Data Ingestion**: Tools and processes for acquiring data.
* **Data Storage**: Databases and data warehouses.
* **Data Processing**: Batch and stream processing engines.
* **Data Analysis**: Analytics tools and frameworks.
* **Data Visualization**: Tools for visual representation of data.
* **Data Management**: Security, governance, and quality management.

**9. Virtualization and Big Data**

**Overview**

* **Virtualization**: Creating virtual versions of physical components like servers, storage devices, and networks.

**Role in Big Data**

* Facilitates efficient resource utilization and management.
* Enhances scalability and flexibility of big data environments.
* Supports multiple workloads and data processing tasks simultaneously.

**Benefits**

* **Resource Optimization**: Efficient use of hardware resources.
* **Cost Savings**: Reduced hardware costs and maintenance.
* **Isolation**: Better security and fault isolation.