SUMMARY-Day9

Name: Tejaswini Gokanakonda

Roll no:DE142

Date:18-11-2024

Big Data

1. Introduction to Big Data

- **Definition**: Big Data refers to data sets that are so large or complex that traditional data-processing software cannot manage them efficiently. It involves massive amounts of structured, semi-structured, and unstructured data generated from multiple sources at a rapid pace.
- **Significance**: Companies like Google, eBay, Facebook, and LinkedIn have built their platforms around Big Data, leveraging it for cost savings, faster decision-making, and enhanced service offerings.

2. Characteristics of Big Data (3 V's)

- **Volume**: The sheer size of data generated daily, measured in gigabytes (GB), terabytes (TB), petabytes (PB), and beyond. Examples include millions of daily customer transactions or social media data.
- **Velocity**: The speed at which data is generated and processed. High-velocity data sources include social media feeds, stock trading systems, and real-time sensors.
- **Variety**: Refers to the different types and formats of data, including:
    - **Structured Data**: Organized in a fixed schema, e.g., relational databases.
    - **Semi-Structured Data**: Contains tags or markers, e.g., XML, JSON.
    - **Unstructured Data**: No defined format, e.g., text, video, audio, images.

3. Sources of Big Data

- **Mobile Devices**: Data from calls, messaging, app usage.
- **Microphones and Cameras**: Audio and video recordings for surveillance or social media.
- **Readers/Scanners**: Barcodes, RFID systems, etc.
- **Science Facilities**: Research data from experiments.
- **Programs/Software**: Logs from application and system activities.
- **Social Media**: Posts, likes, shares, and other interactions.

4. Storing and Processing Big Data

Storage Techniques

- **Hadoop Distributed File System (HDFS)**: A distributed storage system that splits and stores data across multiple machines, making it scalable and fault-tolerant.
- **NoSQL Databases**: Such as HBase, designed to store and query large volumes of unstructured data.

Processing Techniques

- **Hadoop and MapReduce**:
  - **Hadoop** is an open-source framework that allows distributed storage and processing.
  - **MapReduce** processes large data sets by splitting tasks into smaller, parallel operations.
  - **Core Components of Hadoop**:
    - **HDFS**: Storage component of Hadoop.
    - **MapReduce**: Distributed data processing engine.

5. Applications of Big Data

- **Homeland Security**: Real-time threat detection.
- **Healthcare Analytics**: Personalized treatments, predictive analysis.
- **Multi-Channel Sales**: Targeted marketing based on user behavior data.
- **Telecom**: Call data analysis, churn prediction.
- **Manufacturing**: Predictive maintenance and quality control.
- **Traffic Control**: Real-time traffic flow optimization.
- **Trading Analytics**: High-frequency trading algorithms.
- **Search Quality**: Improving search engine accuracy and relevance.

6. Benefits of Big Data

- **Real-Time Insights**: Enables organizations to make data-driven decisions quickly.
- **Cost Reduction**: Optimizes data storage and computing costs through distributed systems.
- **Scalability**: Technologies like Hadoop allow businesses to scale data storage and processing easily from single servers to thousands of machines.
- **Flexibility**: Modern Big Data tools can store raw data, allowing analysis without prior structuring.

7. Overview of Hadoop

- **Definition**: An open-source framework used for storing and processing Big Data in a distributed environment across clusters of computers.
- **History**:
  - Inspired by Google's GFS (Google File System) and MapReduce algorithms.
  - Invented by Doug Cutting and widely popularized by Yahoo.
- **Core Concepts**:
  - **HDFS (Hadoop Distributed File System)**: Provides fault-tolerant storage and high throughput.
  - **MapReduce**: A programming model for distributed data processing, breaking jobs into map and reduce tasks.

8. Tools Used in Big Data

- **Distributed Servers/Cloud** (e.g., Amazon EC2): For hosting data processing.
- **Distributed Storage** (e.g., Amazon S3): For storing data in a scalable manner.
- **High-Performance Schema-Free Databases** (e.g., MongoDB): Allow flexible data storage without fixed schemas.

- **Distributed Processing Frameworks** (e.g., MapReduce): Enable parallel data processing.

9. Apache Spark Overview

- **Definition**: Spark is a fast, general-purpose cluster computing system optimized for Big Data. It performs in-memory computations to speed up data processing.
- **Key Features**:
  - **In-Memory Computation**: Reduces read/write times and increases processing speed.
  - **DAG (Directed Acyclic Graph)**: Optimizes task execution.
  - **Lazy Evaluation**: Transformations are executed only when an action (e.g., `collect()`) is triggered.

10. Spark Components

- **Spark Core**: The foundation for distributed data processing.
- **Spark SQL**: For querying structured data using SQL syntax.
- **Spark Streaming**: Processes real-time data streams.
- **MLlib (Machine Learning Library)**: For scalable machine learning algorithms.
- **GraphX**: For graph-based data computation.

11. Spark Architecture

- **Driver Node**: Manages and coordinates Spark tasks.
- **Worker Nodes**: Execute the tasks.
- **Cluster Manager**: Allocates resources (e.g., Standalone, YARN).
- **Spark Session**: Entry point for interacting with Spark features, creating DataFrames, and managing configurations.

12. Spark Toolset

- **RDD (Resilient Distributed Dataset)**: Low-level, fault-tolerant data structure for distributed processing.
- **DataFrames**: Higher-level abstraction providing optimized data processing.
- **Datasets**: Strongly-typed version of DataFrames with compile-time type safety.