1. Compare Big Data and Data Mining?

Big Data

- **Big Data**: Refers to large, complex, and high-velocity datasets that cannot be easily processed, stored, or analyzed using traditional data processing methods. It encompasses structured, semi-structured, and unstructured data.

Purpose

Focuses on handling, storing, and processing massive amounts of data efficiently. Its aim is to provide infrastructure and tools to manage data.

Tools & Technologies

- Frameworks: Hadoop, Apache Spark, Apache Flink
- Storage: HDFS, NoSQL databases like MongoDB and Cassandra
- Processing: MapReduce, real-time analytics tools

Use Cases

- Real-time monitoring of social media sentiment
- Large-scale financial fraud detection
- Predictive maintenance in IoT

Data Mining

Refers to the process of discovering patterns, trends, and useful information from large datasets using algorithms and statistical techniques.

Purpose

Focuses on extracting actionable insights, patterns, and knowledge from the data. It's about interpreting and deriving value from data.

Tools & Technologies

- Algorithms: Decision trees, clustering, association rules, neural networks
- Software: RapidMiner, WEKA, KNIME, SAS, Python libraries like Scikit-learn

Use Cases

- Market basket analysis (e.g., finding frequently bought items together)
- Customer segmentation
- Predicting customer churn

2.Discuss the role of traditional on-disk storage devices (HDDs, SSDs) in Big Data environments. Evaluate the advantages and limitations of using

on-disk storage for managing large volumes of data. How do modern storage technologies, such as SSDs and hybrid storage solutions, impact the performance and scalability of Big Data platforms?

Advantages of On-Disk Storage in Big Data

1. **High Capacity**:
   - HDDs are cost-effective for storing massive amounts of data, making them suitable for data lakes and archival storage.
   - SSDs provide moderately high capacity with improved performance.
2. **Data Persistence**:
   - Both HDDs and SSDs offer persistent storage, ensuring that data remains available even after power loss.
3. **Scalability**:
   - On-disk storage solutions can be scaled horizontally by adding more devices to distributed storage clusters.

Limitations of On-Disk Storage in Big Data

1. **Performance Bottlenecks**:
   - HDDs suffer from slower read/write speeds due to mechanical components.
   - Random access times in HDDs are significantly slower compared to SSDs.
2. **Energy Consumption**:
   - HDDs consume more power, making them less energy-efficient in large-scale deployments.
3. **Latency**:
   - HDDs introduce higher latency, which can hinder the performance of real-time Big Data applications.
4. **Durability**:
   - HDDs are prone to mechanical failures, leading to potential data loss or downtime.
5. **Costs for High Performance**:
   - While SSDs offer superior performance, their higher cost per terabyte can limit their use for long-term or archival storage.

Solid-State Drives (SSDs)

1. **Performance**:
   - SSDs deliver significantly faster read/write speeds and lower latency, which improves data ingestion, query performance, and real-time analytics.
   - Suitable for workloads requiring frequent random access, such as transactional systems or NoSQL databases.
2. **Durability**:
   - SSDs lack moving parts, making them more reliable and less prone to failure under continuous operation.
3. **Energy Efficiency**:
   - SSDs consume less power compared to HDDs, reducing operational costs in large-scale data centers.
4. **Limitations**:
   - Higher costs per terabyte limit their widespread use for archival or cold storage.
   - Write endurance issues can reduce SSD lifespan under heavy workloads.

Hybrid Storage Solutions

1. **Combining Strengths**:
    - Hybrid solutions leverage HDDs for high-capacity, cost-effective storage and SSDs for fast caching or tiered storage.
    - Frequently accessed ("hot") data is stored on SSDs, while less frequently accessed ("cold") data remains on HDDs.
2. **Performance Optimization**:
    - Significantly improves overall performance by reducing data retrieval times without incurring the full cost of SSD-only systems.
3. **Scalability**:
    - Supports flexible scaling by adding more storage tiers as needed.
4. **Use Cases**:
    - Ideal for Big Data applications requiring a balance between cost and performance, such as e-commerce analytics and log processing.

Melbin Sabu

INTMCA S6

Roll:No:42