CSA1526

CLOUD COMPUTING

Assignment - 4

Name: Sai Lokesh Nalabothu

Reg.No: 192365023

Branch: CSE - Cyber security

Date: 22-02-2025

Serial.No: 19

BIG DATA TOOLS AND FRAMEWORKS

Big data Processing involves Managing vast volumes of structured, semi-structured, and unstructured data efficiently. several tools and frameworks have been developed to store, Process and analyze Big data.

Hadoop

APache Hadoop is an open-source framework that Enables distributed Storage and Processing of large datasets.

Key Concepts

* HDFS
* YARN
* MaPReduce
* HBase

Advantages

* scalable
* Fault-tolevant
* cost effictive.

use cases

* Log Analysis
* Data WareHousing
* Fraud detection

Apache Spark

Apache Spark is a fast and general-purpose cluster computing system that provides in-memory data processing capability.

key Features

* RDD (Resilient Distributed Datasets)
* Spark SQL
* Spark streaming
* MLlib
* GraphX

Advantages

* Faster than Hadoop
* Supports Multiple languages.

NoSQL Databases

NoSQL Databases are designed to handle large scale, distributed data storage with high availability and flexibility.

Types of NoSQL Databases

* key-value stores
* Document oriented
* Column-family stores
* Graph databases

Advantages

* scalable
* schema-less
* High Performance

use cases

* Social media analytics
* Recommendation systems
* IOT

1. What is Apache Hadoop, and how does it support Big Data Processing?

Apache Hadoop is an open-source framework designed for storing and processing massive datasets in a distributed environment. Hadoop is particularly useful for handling structured and unstructured data.

How Hadoop supports Big data Processing

Hadoop consists of several core components that enable efficient data storage and computation:

1. HDFS
2. MapReduce
3. YARN
4. HBase

WHY Hadoop

* Scalability
* Fault tolerance
* Cost- effective
* Flexibility

2. Discuss the Role of Apache Spark in Big data Processing.

Apache Spark is an open-source, It supports Real-time analytics, Machine learning and stream Processing, Making it a Popular Choice for Modern Big data applications.

key Features

1. In-Memory Processing
2. Distributed and fault tolerance
3. Supports Multiple WorkLoads.

4. Compatible with Hadoop & NO SQL Databases.

5. Multi-Language support.

Components of APache Spark

* Spark Core
* Spark SQL
* Spark streaming
* MLlib
* GraphX

Advantages

* Speed
* Flexibility
* Scalability
* Ease of use

Use cases

* Real time Analytics
* ETL
* Healthcare & Genomics

3. How do NOSQL databases support Big Data storage and Querying?

No SQL databases are designed to handle large-scale, distributed, and unstructured data efficiently. No SQL databases provide high scalability, flexible schema and Faster Performance.

Role of NOSQL in BigData Processing

1. Scalability
2. Flexible schema
3. High Performance
4. Real time Data Processing

Querying in NOSQL Databases

* key value stores: GET key, set key value

db.users.find({"age": { "$gt": 25}})

Select * From Users where age > 25;

Advantages

* Handles Large volumes : scales efficiently with large datasets.
* High Availability : Uses Replication for Fault tolerance.
* schema flexibility : No Need for predefined table structures.
* Optimized for Big data Analytics : Works with Real time and Batch Processing frame works like Hadoop & Spark.

Conclusion :

They are widely used in social media platforms, IOT systems, Recommendation engines, and analytics applications.

4. What are the advantages and challenges of using Big data tools?

Bigdata tools like Hadoop, APache spark, and NoSQL databases play a crucial Role in processing, storing and analyzing vast amounts of data.

Advantages of BigData Tools

1. Scalability
2. Speed and Performance
3. Cost-effectiveness
4. flexibility in Data Processing
5. Fault tolevance and Reliability
6. Real time Data Processing
7. Advanced Analytics & AI/ML SUPPort.

Challenges of Big data tools

1. Complexity in implementation
2. High Infrastructure Costs

3. Data security and Privacy Risks
4. Data quality & Integration issues
5. Lack of standardization
6. Maintenance & troubleshooting

Conclusion:

Bigdata tools Revolutionize data processing by offering scalability, speed and flexibility. but they come with challenges like complexity, security risks and high Infrastructure costs.