CSA1526

CLOUD COMPUTING

Assignment - 5

Name : Sai Lokesh Nalabothu

Reg.No : 192365023

Branch : CSE - Cyber Security

Date : 24 - 02 - 2025

Serial.No : 19

Hadoop and Real-time Data Processing

Hadoop is traditionally designed for batch Processing and is not inherently suited for Real-time data Processing.

Hadoop and Real-time Data Processing

Problem Statement

Hadoop is widely used for large-scale data Storage and batch Processing but is not designed for Real time data Processing.

1. Understanding the Components

* Hadoop distributed file system (HDFS): A Scalable Storage system for handling large datasets.

* Map Reduce: A batch Processing Model that is not inherently Real-time but can Process streamed data

in Micro - batches.

Apache kafka

* kafka is a distributed Messaging System designed for handling Real time data Streams

* ACH as a buffer between data Procedures and Real time Processing Systems.

Apache Storm

* A Real time Stream Processing framework that Processes data as it arrives.

* Can Integrate with kafka to Consume and Process Real time data Streams.

→ Latency

→ Scalability

→ Fault tolerance

→ Fraud detection.

1. How is Hadoop used in Real - time data Processing?

* Hadoop is traditionally designed for batch Processing Rather than Real - time data Processing. It Can be Integrated with other technologies to Enable near - Real time or Real time data Processing.

1. Integration with streaming technologies.

* Apache katka

* Apache storm

* Apache spark

2. Real time data Processing workflow using Hadoop.

* Data Integration

* stream Processing

* storage in Hadoop

* Batch Processing for Deep Analysis.

3. Real world Use Cases

- Fraud detection
- Social Media Analytics
- IoT Data Processing

4. Challenges

- Latency
- Complex Architecture
- Scalability & Maintenance

5. Alternative technologies for Real-time Processing

- Apache flink
- Apache Druid
- Google BigQuery & AWS Kinesis

2. What role does Apache kafka Play in Real time data Processing?

APache kafka is a distributed Event Streaming Platform that Plays a Crucial Role in Real time data Processing by acting as a high - throushPut.

1. key Roles of kafka in Real time Data Processing

* Data Ingestion & streaming

* Messsage Buffering

* Data Storage for Replay & Fault tolerance

* Real time Processing With stream Processing FrameWorks.

* Integration With Big data & Databases.

2. Use Cases of kafka

* Fraud detection

* IOT Data Processing

* Social Media Analytics

* Log Monitoring

3. Why use kafka ?

* High ThroughPut

* Scalability

* Fault tolerance

* Low Latency

4. Work FLow

* Producers

* topics

* Consumers

* Databases

3. Explain how Apache storm integrates klith Hadoop for Real-time Processing

Apache storm is a Real time stream processing framenlork that integrates klith Hadoop to Process and analyze data as it arrives.

1. How Apache storm klorks in Real-time Processing

* Spouts

* Bolts

2. Integration of Apache storm klith Hadoop

1. Real time Data Ingestion

2. stream Processing

3. storing Processed Data in Hadoop.

4. Batch Processing & Analytics

3. Real World Use Cases of Storm With Hadoop

* Fraud detection

* IoT Data Analysis

* Real-time Log Processing

4. Advantages of using Apache Storm With Hadoop

* Real-time & Batch Hybrid

* scalability

* Fault tolerance

* Low Latency

4. Discuss the Challenges of using Hadoop for Real time analytics.

Hadoop is batch Processing framework designed for handling large-scale data storage and analytics.

1. High Latency in Processing

* Map Reduce

* large batches

* Low Latency

* Real time decision Making

2. Lack of Native streaming support

* Hadoop does not natively support stream processing.

* Live data streams.

3. Inefficiency in small file handling

* Hadoop's HDFS is optimized for large files but Perform Poorly with many small files.

* Each file's data is stored in NameNode.

WorkAround

* use APache HBase

Conclusion:

Hadoop alone is not well-suited for Real-time analytics due to its batch-oriented nature, high latency, and lack of native streaming support.