211720118030

Real-time Data Processing

Homework 8

Q1. What is Flume?

Apache Flume is an open-source, powerful, reliable and flexible system used to collect, aggregate and move large amounts of unstructured data from multiple data sources into HDFS in a distributed fashion by it's strong coupling with the Hadoop cluster.

Q2. Explain the core components of Flume.

>Flume event

>Flume client

>Flume agent

Q3. What is an Agent?

Flume Agent is an independent daemon process. The agent receives events from clients or other agents and then forwards it to its next destination sink or agent.we can have multiple flume agents. There are three main components of Agent.

>Flume Source

>Flume Channel

>Flume Sink

Q4. What is a channel?

A channel is a temporary store which receives the events from the source and buffers them till they are consumed by sinks. These channels are fully transactional and they can work with any no.of sources and sinks. It acts as a bridge between the sources and the sinks.

Q5. What is Kafka?

Apache Kafka is a distributed data store optimized for receiving and processing streaming data in real-time.A streaming platform needs to handle this constant influx of data, and process the data sequentially and incrementally.It combines messaging, storage, and stream processing to allow storage and analysis of both real-time and historical data.

Q6. List the various components in Kafka.

>Topics

>consumers

>consumer groups

>Producers

>clusters

>brokers

>partitions

>leaders

>followers

>replicas

Q7. What is the role of the ZooKeeper?

Zookeeper acts a Kafka cluster coordinator that manages cluster membership of brokers, producers, and consumers participating in message transfers through Kafka. Zookeeper is used as the core cloud component for node membership and management, coordination of jobs executing among workers, a lock service and a simple queue service.

Q8. Why are Replications critical in Kafka?

Replication is the process of having multiple copies of the data for the sole purpose of availability in case one of the brokers goes down and is unavailable to serve the requests. Data replication is a critical feature of Kafka that allows it to provide high availability and durability.