211720118011

Homework 8

Q1. What is Flume?

Apache Flume is a distributed,reliable and available software for effciently collecting,aggregating and moving large amount of log data. it 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. 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.the project aims to provide a unified,high throughput,low-latency platform for handling real data feeds

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?

The ZooKeeper utility provides configuration and state management and distributed coordination services to Dgraph nodes of the Big Data Discovery cluster.

It ensures high availability of the query processing by the Dgraph nodes in the cluster.

Q8. Why are Replications critical in Kafka?

Data replication is a critical feature of Kafka that allows it to provide high durability and availability.The Replication Factor (RF) is equivalent to the number of nodes where data (rows and partitions) are replicated. Data is replicated to multiple (RF=N) nodes. An RF of one means there is only one copy of a row in a cluster, and there is no way to recover the data if the node is compromised or goes down.