211720243010

BIG DATA ANALYTICS
REAL-TIME DATA PROCESSING
HOMEWORK-8

1.What is Flume?

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

2. Explain the core components of Flume.

A Flume data flow is made up of five main components:

Events
Sources
Channels
Sink
Agents

3. What is an Agent?

An agent is the container for a Flume data flow. It is any physical JVM running Flume. The same agent can run multiple sources, sinks, and channels. A particular data flow path is set up through the configuration process

4. What is a channel?

A channel is an internal passive store with certain specific characteristics. An in-memory channel, for example, can move events very quickly, but does not provide persistence. A file based channel provides persistence. A source stores an event in the channel where it stays until it is consumed by a sink. This temporary storage lets source and sink run asynchronously.

5. What is Kafka?

Apache Kafka is a distributed data store optimized for ingesting and processing streaming data in real-time. Streaming data is data that is continuously generated by thousands of data sources, which typically send the data records in simultaneously. A streaming platform needs to handle this constant influx of data, and process the data sequentially and incrementally.

6. List the various components in Kafka.

The main Kafka components are topics, producers, consumers, consumer groups, clusters, brokers, partitions, replicas, leaders, and followers.

7. 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.ZooKeeper is part of the Hadoop package. The Hadoop package is assumed to be installed on all Hadoop nodes in the BDD cluster deployment. Even though ZooKeeper is installed on all Hadoop nodes in the BDD cluster, it may not be running on all of these nodes. To ensure availability of a clustered Dgraph deployment, configure an odd number (at least three) of Hadoop nodes to run ZooKeeper instances. This will prevent ZooKeeper from being a single point of failure.

8. 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.

Replication factor defines the number of copies of the partition that needs to be kept.