Naan Mudhalvan

Big Data Analytics

Module 8 Homework

Real Time Data Processing

Q1. 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.

Q2. Explain the core components of Flume.

A Flume data flow is made up of five main components: Events, Sources, Channels, Sinks, and Agents.

Events

An event is the basic unit of data that is moved using Flume. It is similar to a message in JMS and is generally small. It is made up of headers and a byte-array body.

Sources

The source receives the event from some external entity and stores it in a channel. The source must understand the type of event that is sent to it: an Avro event requires an Avro source.

Channels

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.

Sinks

The sink removes the event from the channel and forwards it on either to a destination, like HDFS, or to another agent/dataflow. The sink must output an event that is appropriate to the destination.

Agents

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.

Q3. What is an Agent?

A Flume agent is a (JVM) process that hosts the components through which events flow from an external source to the next destination (hop**).** A Flume source consumes events delivered to it by an external source like a web server.

Q4. 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.

Q5. What is Kafka?

Apache Kafka is a real-time data streaming technology capable of handling trillions of events per day. Initially conceived as a messaging queue, Kafka is based on an abstraction of a distributed commit log. Since being created and open sourced in 2011, Kafka has since become the industry standard for working with data in motion.

Q6. List the various components in Kafka.

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

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?

The purpose of adding replication in Kafka is for stronger durability and higher availability. We want to guarantee that any successfully published message will not be lost and can be consumed, even when there are server failures.