Q1. Explain the need of Flume.

Ans:

-Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events from various sources to a centralized data store.

-The purpose of Flume is to provide a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store i.e. it is principally designed to copy streaming data (log data) from various web servers to HDFS.

-Flume is used to move the log data generated by application servers into HDFS at a higher speed.

-When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.

-Flume provides the feature of contextual routing.

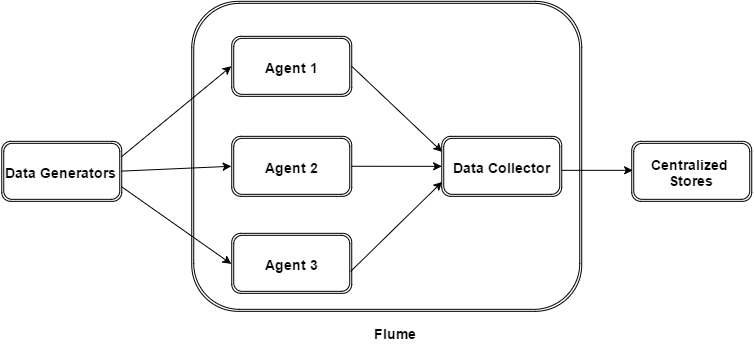
-The transactions in Flume are channel-based where two transactions are maintained for each message. It guarantees reliable message delivery.

-Using Flume, we can get the data from multiple servers immediately into Hadoop.

-Flume can be scaled horizontally.

Q2. Explain the working of Flume and its components in brief.

Ans:

Working of Flume: 

-Data generators (such as Facebook, Twitter) generate data which gets collected by individual Flume agents running on them.

-Thereafter, a data collector (which is also an agent) collects the data from the agents which is aggregated and pushed into a centralized store such as HDFS or HBase.

Apache Flume consists of the following components:

1. Event: A single data record entry transported by flume is known as Event.
2. Source: Source is an active component that listens for events and writes them on one or channels. It is the main component with the help of which data enters into the Flume. It collects the data from a variety of sources.
3. Sink: It is that component which removes events from a channel and delivers data to the destination or next hop. There are multiple sinks available that delivers data to a wide range of destinations. Example: HDFS, HBase, etc.
4. Channel: It is a conduit between the Source and the Sink that queues event data as transactions. Events are ingested by the sources into the channel and drained by the sinks from the channel.
5. Agent: An Agent is a Java virtual machine in which Flume runs. It consists of sources, sinks, channels and other important components through which events get transferred from one place to another.
6. Client: Events are generated by the clients and are sent to one or more agents.