1. Explain the need of Flume.

A company has tons of services running on multiple servers. And lots of data (logs) are produced by them, now we need to analyze them altogether. In order process that logs, we need a reliable, scalable, extensible and manageable distributed data collection service which can perform flow of unstructured data (logs) from one location to another where they will be processed (say in [HDFS](http://goo.gl/26B0d3)).

[Apache flume](http://goo.gl/eZZbmE" \t "_blank) is an open [Apache Flume](http://data-flair.training/blogs/flume-quickstart-guide-learn-install-flume-copy-data-into-hdfs/) is the most reliable, distributed, and available service for systematically collecting, aggregating, and moving large amounts of streaming data (logs) into the Hadoop Distributed File System ([HDFS](http://goo.gl/AacJiN)). Based on streaming data flows, it has a simple and flexible architecture. It is highly fault tolerant and robust and with tunable reliability mechanisms for fail-over and recovery. Flume allows data collection in batch as well as streaming mode.source data collection service for moving the data from source to destination.

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

Flume is a framework which is used to move log data into HDFS. Generally events and log data are generated by the log servers and these servers have Flume agents running on them. These agents receive the data from the data generators.

The data in these agents will be collected by an intermediate node known as **Collector**. Just like agents, there can be multiple collectors in Flume.

Finally, the data from all these collectors will be aggregated and pushed to a centralized store such as HBase or HDFS

There are three types of components in Flume

a)Source

b)Sink

c)Channel

Source:

A **source** is the component of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.

Eg: agent\_name.sources. source\_name.type = value

Sink:

A **sink** stores the data into centralized stores like HBase and HDFS. It consumes the data (events) from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.

Eg: agent\_name.sinks. sink\_name.type = value

Channel:

A **channel** is a transient store which receives the events from the source and buffers them till they are consumed by sinks. It acts as a bridge between the sources and the sinks.

These channels are fully transactional and they can work with any number of sources and sinks.

Eg: agent\_name.channels.channel\_name.type = value