● Explain the need of Flume.

Flume is basically used to handle dynamic data. E.g. twitter, Facebook, log files, etc.

The traditional method of transferring data into the HDFS system is to use the put command. Let us see how to use the put command.

But there are many disadvantages of the put command.

1. The file which is to be transferred should be closed which not possible in case of the dynamic data.
2. Also it can put only one file at a time.
3. Also transfer to the HDFS is slow in case of put command due to the packaging of the data.

Therefore we need a reliable, configurable, and maintainable system to transfer the log data into HDFS.

This is where the Flume comes into picture which handles such transfer of such dynamic files easily. The transfers are handled by the flume agents. Which has the basic components like source, channel ,sink.

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.

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.

Sink - A sink stores the data into centralized stores like HDFS. It consumes the data (events) from the channels and delivers it to the destination.

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

Flume event is basic unit of data transport.

And flume agent is the long run independent flume process which has the following components

1. SOURCE
2. CHANNEL
3. SINK

Source – It receives data from one or more data generators and transfers these data to one or channels.

The external source transfers the data format that is understood by the source e.g. Avro flume.

Channel – It acts as the buffer so that it remains there in the channel till it is consumed by the sink. 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. E.g JDBC channel.

Sink - A sink stores the data into centralized stores like HDFS. It takes the events from the channels and delivers to the destination which would be centralized HDFS or HBase. The flume agent can have multiple source, channels, sinks.

Important thing in flume is that all the transfers of the events are in transactional in nature. Hence the loss of the data is totally avoided. i.e every time the receiver gets the data it commits the received chunk and also sends the acknowledgement.



