1. Explain the need of Flume.

The most of the data used in analysis is the data from data sources like applications servers, social networking sites, cloud servers, and enterprise servers. This data will be in the form of log files and events.

A log file is a file that lists actions that occur in an operating system. For example, web servers list every request made to the server in the log files.

The main challenge with the log files is moving it into HDFS. we have ‘put’ command but it is not helpful for log files . Using put command, we can transfer only one file at a time while the data generators generate data at a much higher rate .And for ‘put’ command the data is needed to be packaged

we can’t use ‘put’ command for log files. So we use new tool ‘FLUME’ to transfer streaming data into HDFS or HBASE .

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

The components of flume are:

  Agent

Data collector

Data generators are servers like facebook ,twitter etc which generate data which gets collected by individual Flume agents running on them. Thereafter, a data collector  collects the data from the agents which is aggregated and pushed into a centralized store such as HDFS or HBase.

Agent has again three components they are source ,channel and sink.

Working or Data Flow:

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. 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. The following diagram explains the data flow in Flume.