**#Assignment\_12.1**

**Que.1) Explain the need of Flume.**

In today’s Generation, amount of data produced is growing rapidly with each passing minute and it is produced by various data sources like applications servers, social networking sites, cloud servers, and enterprise servers. The same data has omnipresent in customer’s life. The data could be of any form like emails, shares, likes, hashtags, comments, logs or events. This data is unstructured and is impractical to store as archives.

Hadoop allows to store and process structured data which is in multiple TB storages. To analyse that data and to transform this multi format unstructured data to HDFS, flume is used. The advantages of using Flume are discussed below:

1) Flume is component in Hadoop ecosystem.

2) Flume handles correction and aggregation of streaming data which continuously records activities performed on data.

3) Flume is reliable, fault tolerant, scalable, manageable, and customizable.

4) Flume supports multi-hop flows, fan-in fan-out flows, contextual routing, etc.

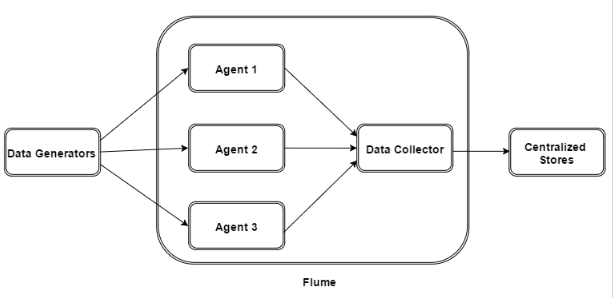
5) For each message flume maintains different channels to provide reliability between sender and receiver.

That’s why flume is very important.

**Que.2) Explain the working of Flume and its components in brief.**

As discussed above to handle and aggregate streaming data Flume is used. The streaming data is mostly produced by various 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, which is unstructured.

The architecture of flume is discussed below:

****

 In diagram above, data generators such as Facebook, Twitter or any platform that has huge continuously streaming data, generate data which gets collected by individual Flume agents running on them. Then, a data collector which is also an agent, collects the data from the agents which is aggregated and pushed into a centralized store HDFS.

* Flume event is a basic unit of data transported inside flume. It contains header and a payload of byte array.
* An agent is an independent daemon process in Flume. It receives the data from clients and transfer it to its next destination Flume may have more than one agent.
* The source of agent receives events from data generators and pass it to channel, 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.
* A sink stores the data into centralized stores like HDFS. It consumes the data from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.
* This is how Flume works.