**ASS-12.1**

**Explain the need of Flume.**

>Basically in hdfs to upload data ,we copy that into Hadoop cluster by single file which is tedious and complex for continuous streaming data .

>So to transfer the data at a high speed ,with high fault tolerant and to maintain the receiving data speed and sending data speed Flume plays a vital role.

>Flume is designed for high-volume ingestion into Hadoop of event-based data.

>The canonical example is using Flume to collect logfiles from a bank of web servers, then moving the log events from those files into new aggregated files in HDFS for processing.

>The usual destination (or sink in Flume parlance) is HDFS. However, Flume is flexible enough to write to other systems, like HBase or Solr.

**Explain the working of Flume and its components in brief**

**Flume Agent**

An **agent** is an independent daemon process (JVM) in Flume. It receives the data (events) from web servers or other sources and forwards it to its next destination (sink or agent). Flume may have more than one agent.

**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.

Apache Flume supports several types of sources and each source receives events from a specified data generator.

**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.

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.

### Interceptors

Interceptors are used to alter/inspect flume events which are transferred between source and channel.

### Channel Selectors

These are used to determine which channel is to be opted to transfer the data in case of multiple channels. There are two types of channel selectors −

* **Default channel selectors** − These are also known as replicating channel selectors they replicates all the events in each channel.
* **Multiplexing channel selectors** − These decides the channel to send an event based on the address in the header of that event.

### Sink Processors

These are used to invoke a particular sink from the selected group of sinks. These are used to create failover paths for your sinks or load balance events across multiple sinks from a channel.