*Assignment-12.1:*

*Question 1 : Explain the need of Flume:*

1. *It generally pulls the data from these services and automatically store it in the HDFS.*
2. *The traditional method of transferring data into the HDFS system is to use the put command.*
3. *Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS)*
4. *Flume is designed for high-volume ingestion into Hadoop of event-based data.*
5. *Flume is basically used to handle dynamic data. E.g. twitter, Facebook, log files, etc.*
6. *The usual destination (or sink in Flume parlance) is HDFS. However, Flume is flexible enough to write to other systems, like HBase or Solr.*

*Features of Flume:*

* *Stream data*
* *Insulate systems*
* *Guaranteed data delivery*
* *Scale horizontally*

*Question 2 : Explain the working of Flume and its components in brief.*

*The following components make up Apache Flume:*

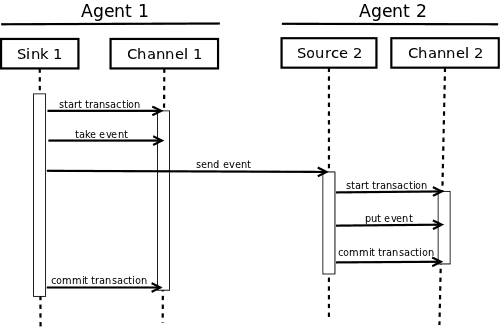
* *Event*
* *Source*
* *Sink*
* *Channel*
* *Agent*
* *Client*

*Flume :*

*It contains a payload of byte array that is to be transported from the source to the destination accompanied by optional headers.*

*Flume Agent :*

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

**

*Architecture of Apache Flume*

### *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.*

### *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.*

### *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.*