Assignment-12.1:

**Explain the need of Flume:**

In apache flume it is a efficient method for distributed and reliable type of services where a large amount of streaming data is moved for the process of streaming into HDFS (Hadoop Distributed File System).

It can handle dynamic data handling (for e.g. Facebook, gmail) .

All the data were pulled from the services and will be stored in HDFS automatically.

Like HBase or Solr in Apache were flume will be much flexible to write which acts as destination in HDFS.

The purpose of flume is for high-volume ingestion which is acts as a event based data in Hadoop.

Put command is a traditional one which is used for data transformation in Hadoop.

Still there will persist some sort of disadvantage in put command to be used:

1. In dynamic data it is not possible to close the file which need to be transferred.
2. One file can be copied at a time.

Features of Flume:

* Data Streaming
* Insulate systems
* Data delivery will be guaranteed
* Will Scale horizontally

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

Apache Flume is made up of following components:

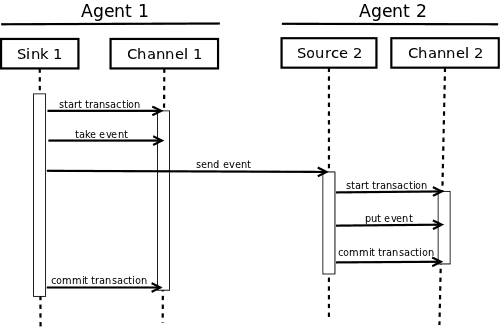
* Event
* Source
* Sink
* Channel
* Agent
* Client

***Flume*:**

All the events will be transferred inside flume based on data unit.

Byte array will be loaded inside them where it need to be transferred from the source to destination which will be accompanied by header(Optional).

Apache Flume Architecture:



***Flume Agent***

JVM is an independent daemon in flume

Where all the data will be received in the form of events from the client or agents and later it will be forwarded to the destination which follows (eg: sink, agent).

And more than one agent is possible.

And comprises of three main components and they are:

**source**, **channel**, and **sink**.

### ***Source***

### From the data generators data will be received to the source and from that the data will be transferred to more number of channels as flume events.

***Channel***

Channel acts as a store where all the data will be received as event from the source and will be buffered until the sink gets consumed.

It acts as a bridge in between the source and the sink.

***Sink***

In sink data will be stored in HDFS or HBase.

The data which all consumed from different channels will be delivered to a specific destination.

The central source or agent acts as a destination of sink.

### ***Interceptors***

They are used for altering or inspecting the flume which acts as transformation agent between the source and the destination.

### ***Sink Processors***

They are used to retrieve particular sink from a group of sink.

They are specially used to create a failover path or load balance from channel to sink..