Assignment-12.1:

Explain the need of Flume:

* 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)
* Flume is basically used to handle dynamic data. E.g. twitter, Facebook, log files, etc.
* Flume generally pulls the data from these services and automatically store it in the hdfs.
* Flume is designed for high-volume ingestion into Hadoop of event-based data.
* The usual destination (or sink in Flume parlance) is HDFS. However, Flume is flexible enough to write to other systems, like HBase or Solr.
* The traditional method of transferring data into the HDFS system is to use the put command

But there are many disadvantages of the put command.

1. The file which is to be transferred should be closed which not possible in case of the dynamic data.
2. Also it can copy only one file at a time.

Features of Flume:

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

Explain the working of Flume and its components in brief.

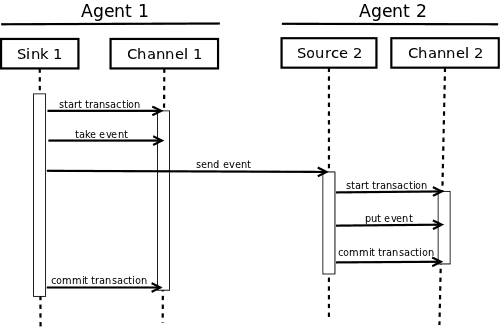
The following components make up Apache Flume:

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

Flume:

An **event** is the basic unit of the data transported inside **Flume**. It contains a payload of byte array that is to be transported from the source to the destination accompanied by optional headers

Architecture Of Apache Flume:



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. Flume Agent contains three main components namely, **source**, **channel**, and **sink**.

### **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. It acts as a bridge between the sources and the 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.