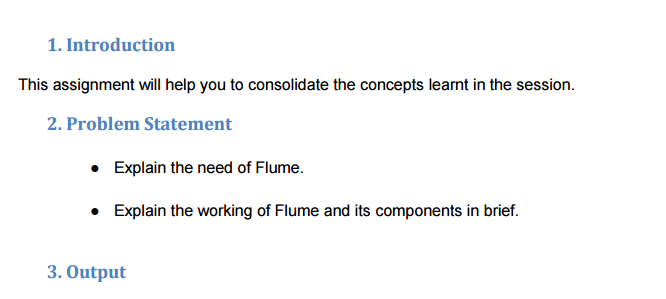
**Assignment 12.1 ---**



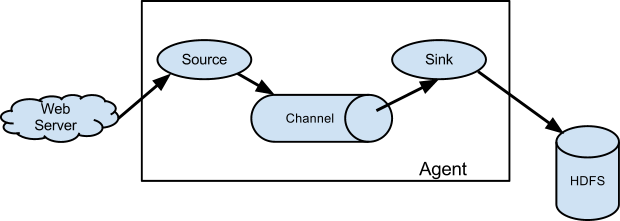
**>> The Need of Flume :**

* Flume can transport log files across a large number of hosts into HDFS.
* Using Apache Flume we can store the data in to any of the centralized stores (HBase, HDFS).
* When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.
* The transactions in Flume are channel-based where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.
* Flume can be used to transport massive quantities of event data including network traffic data, social-media-generated data, email messages and pretty much any data source possible.
* Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data.
* Flume is flexible enough to write to other systems, like HBase or Solr.

**>> Working of flume and its components :**

Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store.

A **Flume** event is defined as a unit of data flow having a byte payload and an optional set of string attributes. A Flume agent is a process that hosts the components through which events flow from an external source to the next destination (hop).



**Source**: A Flume source consumes events delivered to it by an external source like a **web Server**. The **external source** sends events to Flume in a format that is recognized by the target Flume source.

**Channel**: When a Flume source receives an event, it stores it into one or more channels. The channel is a passive store that keeps the event until it’s consumed by a Flume sink. The file channel is one example – it is backed by the local filesystem.

**Sink**: The sink removes the event from the channel and puts it into an external repository like **HDFS** (via Flume HDFS sink) or forwards it to the Flume source of the next Flume agent (next hop) in the flow. The source and sink within the given agent run asynchronously with the events staged in the channel.

**For example:**

In this case, agent1.sources.source1.type is set to spooldir, which is a spooling directory source that monitors a spooling directory for new files.

• The spooling directory source defines a spoolDir property, so for source1 the full key is agent1.sources.source1.spoolDir.

• The source’s channels are set with agent1.sources.source1.channels.

• The sink is a logger sink for logging events to the console.

• It must be connected to the channel (with the agent1.sinks.sink1.channel property).

• The channel is a file channel, which means that events in the channel persist with disk for durability.

