**Session 12**

**Assignment 1**

**Part 1:** Explain the need of Flume.

Apache Flume used for the service of streaming logs into Hadoop. It 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). The architecture of Flume is very simple and flexible based on flow of streaming data.

Some of the notable needs of Apache Flume are as follows,

* Flume ingests log data from multiple web servers into a centralized store (HDFS, HBase) efficiently.
* Using Flume, we can get the data from multiple servers immediately into Hadoop.
* Along with the log files, Flume is also used to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart.
* Flume supports a large set of sources and destinations types.
* Flume provides the feature of contextual routing.
* Flume is reliable, fault tolerant, scalable, manageable, and customizable.
* Flume supports multi-hop flows, fan-in and fan-out flows, contextual routing, etc.
* Flume can be scaled horizontally to ingest new data streams and additional volumes as needed.
* Flume insulates the systems by buffering the storage platform from transient spikes, when the rate of incoming data exceeds the rate at which data can be written to the destination.

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

1. **Working of Flume:**

The working of Flume is basically defined as data flow in Hadoop.

* Flume is a framework which is used to move log data into HDFS.
* Generally events and log data are generated by the log servers and these servers have Flume agents running on them.
* These agents receive the data from the data generators.
* The data in these agents will be collected by an intermediate node known as Collector.
* Just like agents, there can be multiple collectors in Flume.
* Finally, the data from all these collectors will be aggregated and pushed to a centralized store such as HBase or HDFS.

1. **Components of Flume:**

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

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

* **Agent:**

An agent is an independent daemon process (JVM) in Flume that receives the data (events) from clients or any other sources and forwards it to its next destination (sink or agent). Flume may have more than one agent. It is a collection of sources, sinks and channels.

* **Source:**

A source is the entity of an Agent which receives data from the data generators and transfers it to one or more channels in the form of Flume events.

**Example** - Avro source, Thrift source, etc.

* **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.

**Example** - JDBC channel, Memory channel, etc.

* **Sink:**

A sink stores the data into centralized stores like HBase and HDFS and also it consumes the data from the channels and delivers it to the destination. The destination of the sink might be another agent or the central stores.

**Example** - HDFS sink.

* **Client:**

Client is the entity that produces and transmits the Event to the Source operating within the Agent.