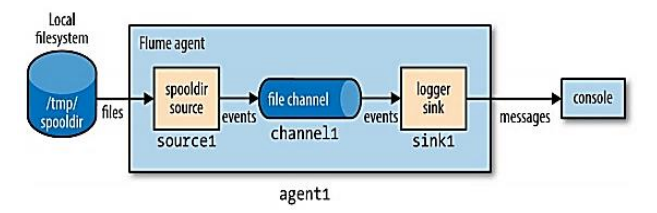
**Problem Statement-**

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

* Normally while working with HDFS, it is assumed that the data is already present in HDFS or can be copied there in bulk. But there are systems that do not meet this assumption.
* There are some systems which produces streams of data and works on these streamed data.
* They produce these streamed data and would like to aggregate, store, and analyze using Hadoop, so for processing these streamed data Apache Flume is needed.
* Flume is designed for high volume ingestion into Hadoop of event based data.
* For example Flume collects log files from any server let say banks servers, then move this log events from those files into new aggregated files in HDFS for processing.
* Flume allows to work with Stream data, Insulate systems, it also guarantee data delivery and can be scaled horizontally.
* Flume is basically used to move the log data generated by application servers into HDFS at a very high speed.
* So that it can be processed for decision making.

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

* Flume is a tool which helps to listen streaming data or data which is coming from server logs.
* To use Flume, we need to run a Flume agent, which is a long-lived java process that runs sources and sinks, connected by channels.
* A source in Flume produces events and delivers them to the channel, which stores the events until they are forwarded to the sink.
* You can think of the source-channel-sink combination as a basic Flume building block.
* Flume installation is made up of a collection of connected agents running in a distributed topology.
* Agents on the edge of the system (co-located on web server machines, for example) collect data and forward it to the agents that are responsible for aggregating and then storing the data in its final destination.
* Agents are configured to run a collection of particular sources and sinks, so using Flume is mainly a configuration exercise in wiring the pieces together.
* 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 JVM process that hosts from an external sources to the next destination.



**Flume architecture-**

* 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.

1. **Flume Event-**

Flume event is the basic unit of data transmission inside Flume which contains a payload of byte array transported from source to destination which includes optional headers to identify payload.

1. **Flume Agent-**

A Flume agent is an independent daemon process, which receives data (events) from clients and forwards it to the next destination. A Flume may have more than one agent.

A Flume agent contains three components-

1. **Source-** It is an inner component of agent which receives data from data generators and transfers it to channels in the form of Flume events.
2. **Channel-** A channel is a store which receives the events from the source and buffers them till they are consumed by sinks. Acts as a bridge between sources and sinks. These channels are fully transactional and can work with any number of sources and sinks.
3. **Sink-** A sink is the third component of agent, stores data into centralized stores like HBase and HDFS. It consumes the data from channels and delivers it to the destination, destination can be another agent or central stores.