1. Explain the need of Flume.

Apache Flume is a tool ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events etc. From various sources to a centralized data store. Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data, log data from various web servers to HDFS.

Assume an e-commerce web application wants to analyze the customer behavior from a particular region. To do so, they would need to move the available log data in to Hadoop for analysis. Here, Apache Flume comes to our use. Flume is used to move the log data generated by application servers into HDFS at a higher speed.

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

* **Event:** A byte payload with optional string headers that represent the unit of data that Flume can transport from it’s point of origination to it’s final destination.
* **Flow:** Movement of events from the point of origin to their final destination is considered a data flow, or simply flow. This is not a rigorous definition and is used only at a high level for description purposes.
* **Client:** An interface implementation that operates at the point of origin of events and delivers them to a Flume agent. Clients typically operate in the process space of the application they are consuming data from. For example, *Flume Log4j Appender* is a client.
* **Agent:** An independent process that hosts flume components such as sources, channels and sinks, and thus has the ability to receive, store and forward events to their next-hop destination.
* **Source:** An interface implementation that can consume events delivered to it via a specific mechanism. For example, an Avro source is a source implementation that can be used to receive Avro events from clients or other agents in the flow. When a source receives an event, it hands it over to one or more channels.
* **Channel:** A transient store for events, where events are delivered to the channel via sources operating within the agent. An event put in a channel stays in that channel until a sink removes it for further transport. An example of channel is the JDBC channel that uses a file-system backed embedded database to persist the events until they are removed by a sink. Channels play an important role in ensuring durability of the flows.
* **Sink:** An interface implementation that can remove events from a channel and transmit them to the next agent in the flow, or to the event’s final destination. Sinks that transmit the event to it’s final destination are also known as terminal sinks. The *Flume HDFS sink* is an example of a terminal sink. Whereas the Flume Avro sink is an example of a regular sink that can transmit messages to other agents that are running an Avro source.