Assignment-12.3:

**What is meant by Flume NG?**

 Flume NG is a refactoring of Flume. To solve certain known issues and limitations, Flume requires a refactoring of some core classes and systems.

The following known issues are specifically to be addressed:

* Code Complexity
* Core component lifecycle standardization and control code-(example-anything that can be started or stopped, sources, sinks)
* Drastic simplification of common data paths
* Renaming packages to org.apache.flume
* Heartbeat and master Rearchitecture.

**Can Flume provides 100 % reliability to the data flow?**

Yes, it provide end-to-end reliability of the flow. Sources and sinks encapsulated in a transactional repository provides by the channels.

The most important thing in flume is that all the transfers of the events are in transactional in nature. Hence the loss of the data is totally avoided. i.e every time the receiver gets the data it commits the received chunk and also sends the acknowledgement.

**Can Flume can distributes data to multiple destinations?**

Yes, it support multiplexing flow. The event flows from one sources to multiple channels and multiple destinations. It’s achieved by defining a flow multiplexer.

As the flume agent can declare more than one source, channels and sink. Hence they can take inputs from many sources and buffer in multiple channels as well as give them to multiple sinks (Destinations). For doing this it has to change the configuration of the flume and allocate the respective sinks.

Example:

Setting channels for sinks:

agent\_foo.sinks.hdfs-ca-sink.channel = mem-channel-ca

agent\_foo.sinks.hdfs-ny-sink.channel = mem-channel-ny

**Explain about the different channel types in Flume. And which channel type is faster?**

Channels are buffers that sit in between sources and sinks. As such, channels allow sources and sinks to operate at different rates. Channels are key to Flume’s guarantees of not losing data. Sources write data to one or more channels, which are read by one or more sinks

Different Channels in Flume are

1. Memory Channel

The Memory Channel is an in-memory channel that stores events written to it on the heap. The Memory Channel supports very high throughput, as it holds all data in memory. The Memory Channel is an in-memory queue—the sources write to its tail and sinks read off its head.

* The events are stored in an in-memory queue with configurable max size. It’s ideal for flows that need higher throughput and are prepared to lose the staged data in the event of agent failures.
* The maximum number of events stored in the channel -100.
* The maximum number of events the channel will take from a source or give to a sink per transaction-100.
* Timeout in seconds for adding or removing an event-3sec

1. File Channel:

The File Channel is Flume’s persistent channel. It writes out all events to disk and thus does not lose data on process or machine shutdown or crash. The File Channel ensures that any events committed into the channel are removed from the channel only when a sink takes the events and commits the transaction, even if the machine or agent crashed and was restarted. It is designed to be highly concurrent and to handle several sources and sinks at the same time

* The maximum size of transaction supported by the channel-100.
* Amount of time (in millis) between checkpoints – 30000 mili secs.
* Maximum capacity of the channel- 1000000
* Amount of time (in sec) to wait for a put operation- 3.