**Session 12**

**Assignment 3**

**Part 1:** What is meant by Flume NG?

* 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).
* Where **Flume NG** is work related to new major revision of Flume.
* At a high-level, Flume NG uses a single-hop message delivery guarantee semantics to provide end-to-end reliability for the system.
* To accomplish this, certain new concepts have been incorporated into its design, while certain other existing concepts have been either redefined, reused or dropped completely.

**Part 2:** Can Flume provides 100 % reliability to the data flow?

* Yes.
* Apache Flume provides end-to-end reliability of the flow because by default it uses a transactional approach in the data flow.
* Source and sink encapsulate in a transactional repository provides by the channels.
* This channels responsible to pass reliably from end to end flow. So it provides 100% reliability to the data flow.

**Part 3:** Can Flume distribute data to multiple destinations?

* Yes.
* Apache Flume supports multiplexing flow.
* In this the event flows from one source to multiple channel and multiple destinations through multiple agents. It is achieved by defining a flow multiplexer.

**Part 4:** Explain about the different channel types in Flume. And which channel type is faster?

**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.

1. Different Types of Channels in Flume:

There are three different built in channel types available in Flume. They are-

* **MEMORY Channel** - Events are read from the source into memory and passed to the sink.
* **JDBC Channel** - JDBC Channel stores the events in an embedded Derby database.
* **FILE Channel** - File Channel writes the contents to a file on the file system after reading the event from a source. The file is deleted only after the contents are successfully delivered to the sink.

1. Fastest between the Channels:

* **MEMORY Channel** is the fastest channel among the three however has the risk of data loss.
* The channel that you choose completely depends on the nature of the big data application and the value of each event.
* In memory channel the event is simply stored in an in-memory queue, which is faster but any events still left in the memory channel when an agent process dies can’t be recovered and this condition is called data loss.