Flume and Sqoop for Ingesting Big Data

Modality: On Demand

Duration: 2 Hours

About this course:

Import data: Flume and Sqoop have a crucial part to play in the Hadoop ecosystem. They have the responsibility of transferring the data from sources like local file systems, HTTP, MySQL and Twitter. These hold/produce data to data stores like HDFS, HBase and Hive. Both the tools have default functionality and have the ability of abstracting away the users from the complication of transferring data among these systems.

Flume: Flume Agents have the ability to transfer data created by a streaming application to data stores like HDFS and HBase.

Sqoop: Sqoop can be used to bulk import data from typical RDBMS to Hadoop storage structures like HDFS or Hive.

Learning Objectives:

Practical application for the various sources and data stores:

- Sources: Twitter, MySQL, Spooling Directory, HTTP
- Data stores: HDFS, HBase, Hive

Flume components:

- Flume Agents
- Flume Events
- Event bucketing
- Channel selectors
- Interceptors

Sqoop components:

- Sqoop import from MySQL
- Incremental imports using Sqoop Jobs

Audience:

This course will be highly useful for those engineers who have the responsibility designing an application with HDFS/HBase/Hive as the data store. This will also be suitable for those engineers who intend to port data from legacy data stores to HDFS.

Requirements:

The course has a mandatory requirement of having knowledge of HDFS. You should also be having fundamental understanding of HBase and Hive shells, as HBase and Hive examples require that. Additionally, you should also be having a working installation of HDFS, because it is required to run majority of the examples.

Course Outline:

- You, This Course and Us
- Why do we need Flume and Sqoop?
- Sqoop
- Flume