Session 11: Sqoop Flume

Assignment 3

**Problem Statement**

Create a flume agent that streams data from Twitter and stores in the HDFS.

**Solution:-**

**Step 1 – Creating an app in twitter**

To create a Twitter application, click on the <https://apps.twitter.com/>. Sign in to your Twitter account.

Click on the **Create New App** button. You will be redirected to a window where you will get an application form in which you have to fill in your details in order to create the App. While filling the website address, give the complete URL pattern, for example, [http://example.com](http://example.com/)





Under **keys and Access Tokens** tab at the bottom of the page, you can observe a button named **Create my access token**. Click on it to generate the access token.



Finally, click on the **Test OAuth** button which is on the right side top of the page. This will lead to a page which displays your **Consumer key, Consumer secret, Access token,** and **Access token secret**. Copy these details. These are useful to configure the agent in Flume.

**Step 2: Create a Directory in HDFS**

In Hadoop DFS, you can create directories using the command **mkdir**. Browse through it and create a directory with the name **twitter\_data** in the required path as shown below.

Hadoop fs –mkdir /twitter\_data



**Step 3- Create configuration file**

Create file twitter.conf and put in $FLUME\_HOME/conf



Contents of twitter.conf

# Naming the components on the current agent.

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

# Describing/Configuring the source

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.consumerKey = -----------------------

TwitterAgent.sources.Twitter.consumerSecret = ------------------------

TwitterAgent.sources.Twitter.accessToken = --------------------------------------------------

TwitterAgent.sources.Twitter.accessTokenSecret = --------------------------------------------

TwitterAgent.sources.Twitter.keywords = tutorials point,java, bigdata, mapreduce, mahout, hbase, nosql, hi, hello, welcome, the, this

# Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/twitter\_data/

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

# Describing/Configuring the channel

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel

**Step 4 – Execution**

Go to FLUME\_HOME (/usr/local/flume in acadgild VM)

Run flume agent using command

**flume-ng agent --conf ./conf/ -f conf/twitter.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent**



Checking output directory HDFS:/twitter\_data

Command –

**hadoop fs -ls /twitter\_data**



Upon examining the files we could see that Flume agent has dumped all twitter feed data in the files

Commands –

**hadoop fs -tail -f /twitter\_data/FlumeData.1508951689563**

**hadoop fs -tail -f /twitter\_data/FlumeData.1508951749122**

