GUIDELINES TO STREAM TWITTER DATA USING FLUME

To login into Hadoop cluster:

1. First login into the cluster with your username and password.
2. For windows users Login with your user name (CSX user name) to hadoopnn001.cs.okstate.edu from putty.
3. MAC/LINUX users type ssh user@hadoop-nn001.cs.okstate.edu from your terminal.
4. You will have access to user/your user_name directory in HDFS **To start streaming twitter data:**

1. Create an account in twitter and login with the credentials.
2. Navigate to **https://apps.twitter.com/** and create a new app.
3. Give the Name and the Description on what are you the data, and the website URL is https://twitter.com/
4. Get the consumer secret, consumer token, access token and access token secret for your application. To get the consumer secret, consumer token:
   - Go to Keys and Access Tokens get the consumer key and consumer secret, and then go to Your Access tokens and get the access tokens and access token secret.
4. There are 3 components for a twitter agent namely source, sink and channel.
5. The flume source connects to Twitter API and receives data in JSON format which in turn stored into HDFS.
6. Create a path in hdfs where data should be collected(same path should be given in conf file)

Eg: hdfs dfs -mkdir /user/ankaush/JoeBiden_data/

7. Now, create a configuration file for the flume agent by specifying the consumer key, consumer secret, access token, access token secret, keywords and HDFS path.

Sample Config file:

```
TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS


TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel
```

```
TwitterAgent.sources.Twitter.consumerKey = ****
TwitterAgent.sources.Twitter.consumerSecret = ****
TwitterAgent.sources.Twitter.accessToken = ****
TwitterAgent.sources.Twitter.accessTokenSecret = ****
TwitterAgent.sources.Twitter.keywords = biden
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop-
nn001.cs.okstate.edu:9000/sisheri/JoeBiden_data/%Y/%m/%d/%H
TwitterAgent.sinks.HDFS.hdfs.useLocalTimeStamp = true
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwtterAgent.sinks.HDFS.hdfs.rollCount = 0
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 10000
```

8. Command to start the flume agent:

nohup $FLUME_HOME/bin/flume-ng agent -n TwitterAgent –conf $FLUME_HOME/conf -f Configuration_File_Path &

Example: nohup $FLUME_HOME/bin/flume-ng agent -n TwitterAgent --conf $FLUME_HOME/conf -f /home/ankaush/joebiden.conf &

nohup will make sure the data collection process runs continuously at the backend. nohup.out file is the log file that will be created as we start the process. The data collected will be in JSON format.