Streaming twitter data using Flume

Steps on twitter :

1.Login to twiiter account

2.Go to https://apps.twitter.com/apps and click on ‘create new app’ button

3.Give your app a name, description and enter a website

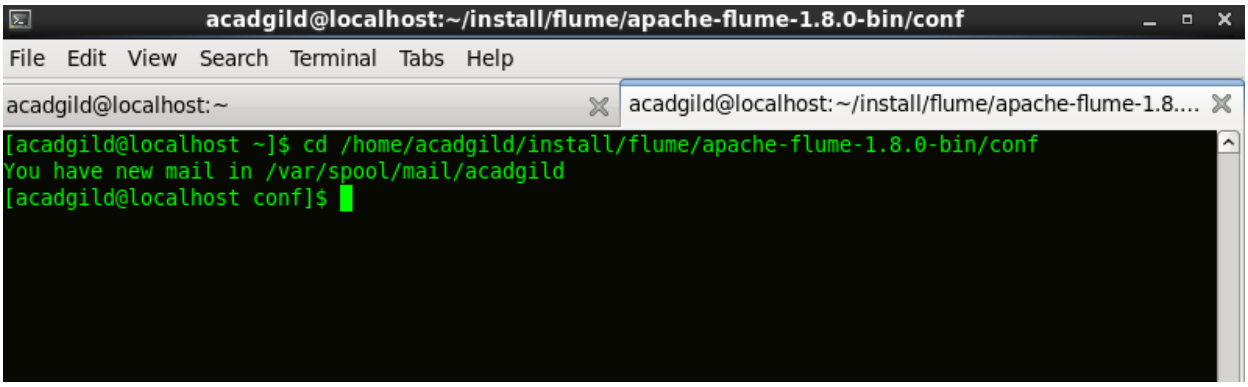
4. Accept the developer agreement and select the ‘create your Twitter application’ button.

5. Select the ‘Keys and Access Token’ tab. Copy consumer and consumer secret code and keep it.

6.Scroll down further to find ‘create access token’ button. Copy access token and access token secret code

Steps on acadgild VM :

Go to the conf directory inside flume extracted directory, In my case its /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf



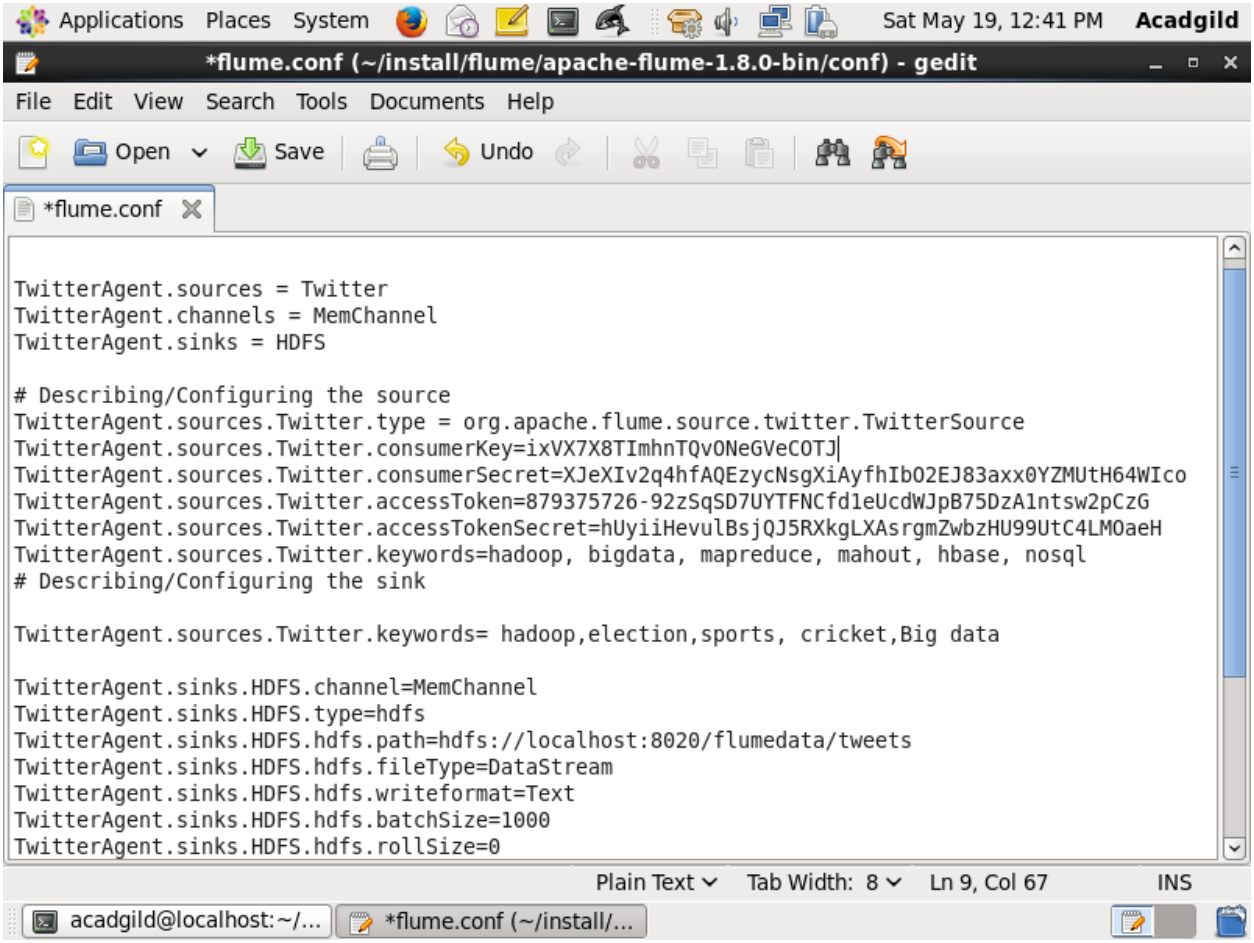
Note: Make sure you have below jars placed in your lib directory:

1. twitter4j-core-X.XX.jar
2. twitter4j-stream-X.X.X.jar
3. twitter4j-media-support-X.X.X.jar

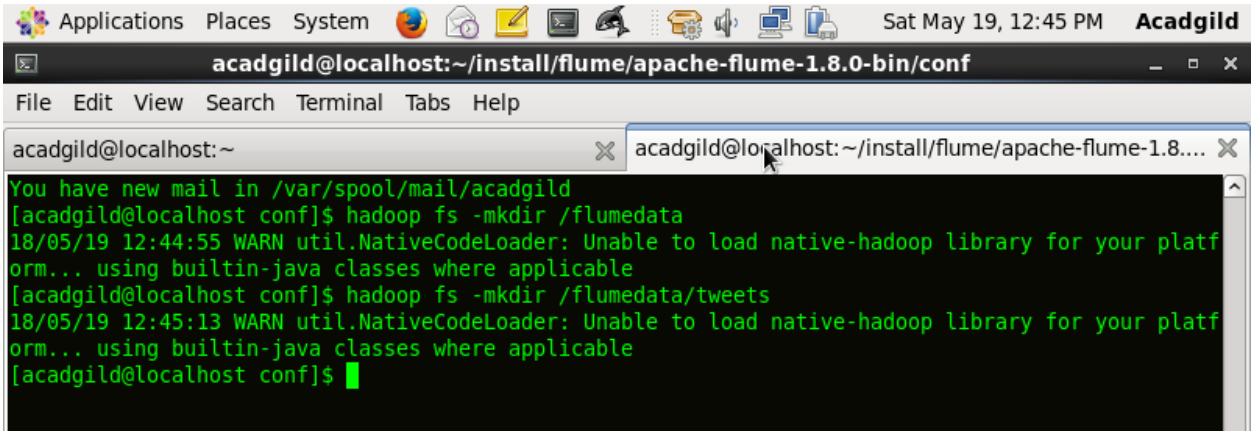
Run the command gedit flume.conf to open up the editor

Copy the flume configuration code and replace it with your own consumer key , consumer secret key, accesstoken and secret accesstoken. Also enter your hdfs path where data from twitter will be dumped

```
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=uX0TWqkx0okYEjjqLzxIx6mD6
TwitterAgent.sources.Twitter.consumerSecret=rzHIs3TMJnADbZNvdGU7LQUo0kPxPISq3
RGSLfqcBip39X5END
TwitterAgent.sources.Twitter.accessToken=559516596-
yDA9xqOljo4CV32wSnqsx2BXh4RBIRKFxZGSZrPC
TwitterAgent.sources.Twitter.accessTokenSecret=zDxePILZitS5tIWBhre0GWqps0FIj9
OadX8RZb6w8ZCwz
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout,
hbase, nosql
# Describing/Configuring the sink

TwitterAgent.sources.Twitter.keywords= hadoop,election,sports, cricket,Big
data

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/user/flume/tweets
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
TwitterAgent.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=1000

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChan
```



Note my hdfs path: hdfs ://localhost:8020/flumedata/tweets
Creating flumedata directory inside hdfs and tweets directory in flumedata

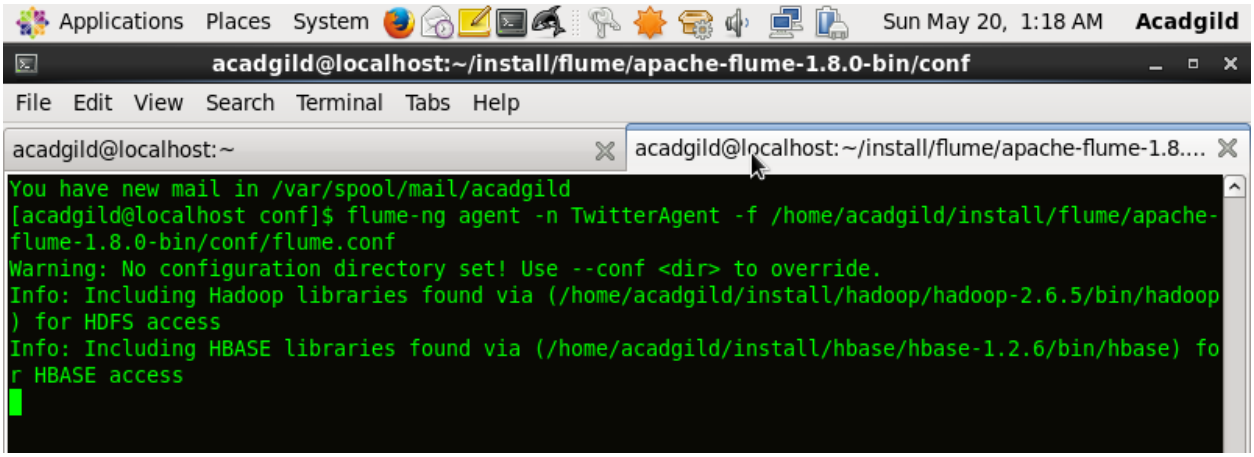


Start all the hadoop daemons run jps to check if all are running

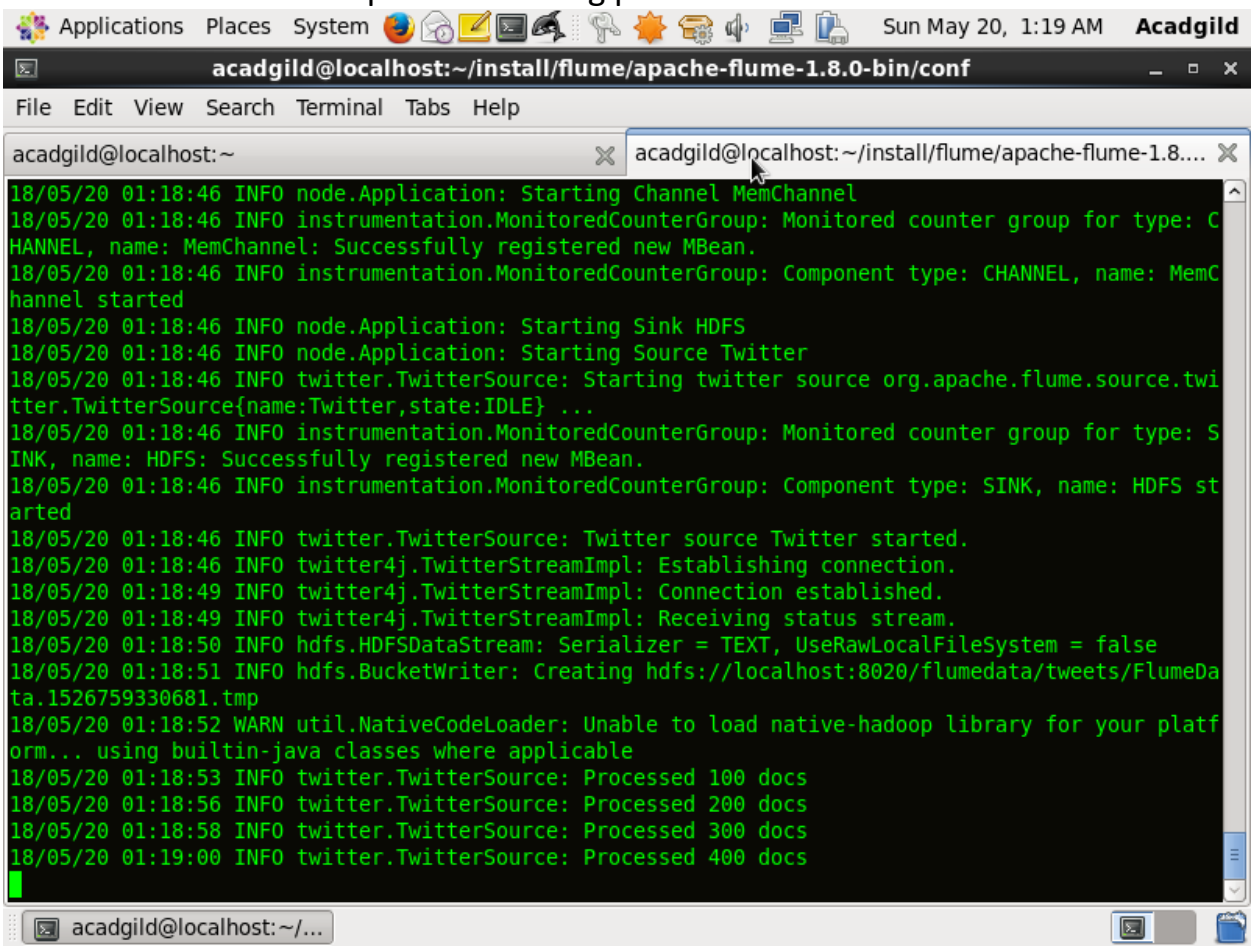
```
[acadgild@localhost conf]$ jps
28643 Jps
9109 NameNode
9544 ResourceManager
9401 SecondaryNameNode
9210 DataNode
9646 NodeManager
29343 HQuorumPeer
10079 JobHistoryServer
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost conf]$
```

For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

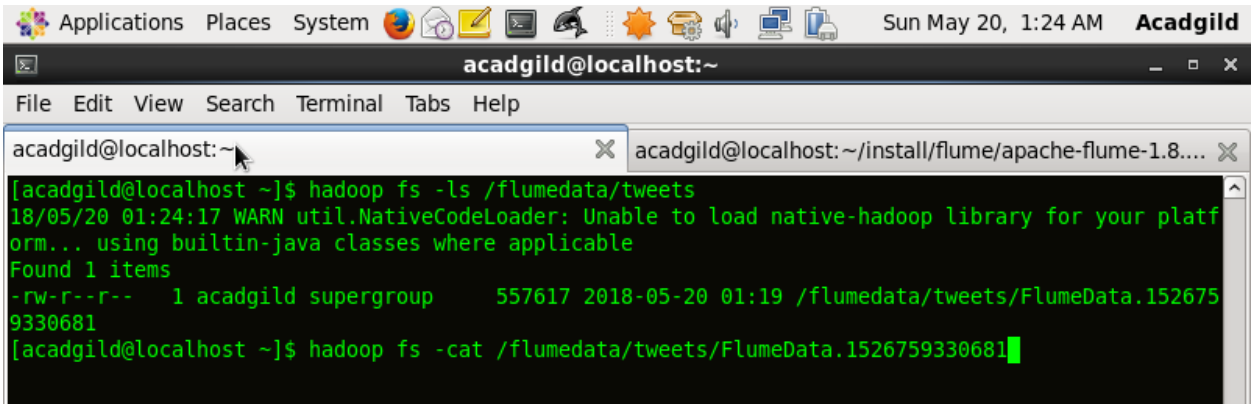
flume-ng agent -n TwitterAgent -f <location of created/edited conf file>



Once, the tweet data started streaming it into the given HDFS path we can use ‘Ctrl+c’ command to stop the streaming process.



Lets check the content of /flumedata/tweets



Screenshot of output of some data which has been dumped