**Assignment 11.3**

**Problem Statement**

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

**To stream data to our database from twitter we should have the following pre-requisites.**

Twitter account

Hadoop cluster

**If both prerequisites are available we can move to our further step.**

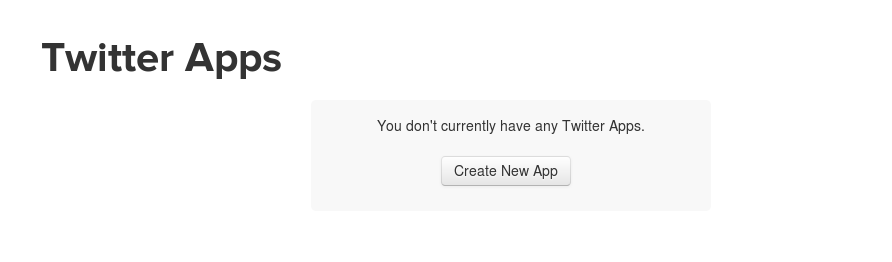
**Step 1:**

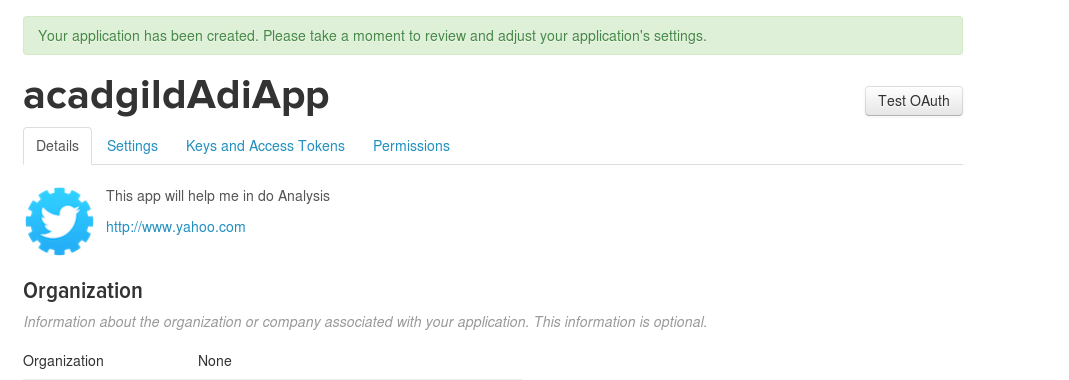
Login to twitter account

**STEP 2:**

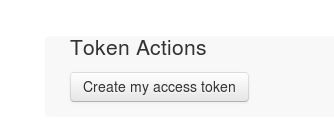
**Go to this link and Create New Application**

[**https://apps.twitter.com/**](https://apps.twitter.com/)





**Now go to Keys and Access and create token**



**Consumer Key (API Key):** g3LJ1l7eTscLbERCeRI1VX2AX

**Consumer Secret (API Secret):** bxHmWYmODV4wr4ksgIbIR5SXHmiy4KoD1eKP8DkxDKevgzACC6

**Access Token :** 482791704-wF0JQTnagjhCB5Z1xYsIHLl2Hs8EVVhAgIBjcGoQ

**Access Token Secret:** xUKRNvVX0BIcBcdXoeCJzXFGgd5BXbPcXmA6M6K8FPQko

**Create a new file inside the conf directory inside the Flume-extracted directory.**

**File :** /home/cloudera/Desktop/Acadgild/Assignment\_11\_3/flume.conf

**Note: Make sure you have below jars placed in your $FLUME\_HOME/lib directory:**

**twitter4j-core-X.XX.jar**

**twitter4j-stream-X.X.X.jar**

**twitter4j-media-support-X.X.X.jar**

**Copy the Flume configuration code from the below link and paste it in the newly created file.**



[**https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk**](https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk)

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=g3LJ1l7eTscLbERCeRI1VX2AX

TwitterAgent.sources.Twitter.consumerSecret=bxHmWYmODV4wr4ksgIbIR5SXHmiy4KoD1eKP8DkxDKevgzACC6

TwitterAgent.sources.Twitter.accessToken=482791704-wF0JQTnagjhCB5Z1xYsIHLl2Hs8EVVhAgIBjcGoQ

TwitterAgent.sources.Twitter.accessTokenSecret=xUKRNvVX0BIcBcdXoeCJzXFGgd5BXbPcXmA6M6K8FPQko

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

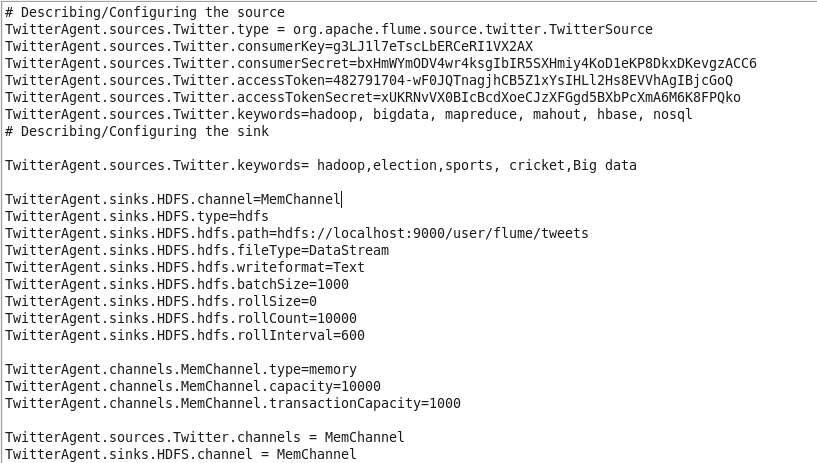
TwitterAgent.channels.MemChannel.type=memory

TwitterAgent.channels.MemChannel.capacity=10000

TwitterAgent.channels.MemChannel.transactionCapacity=1000

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sinks.HDFS.channel = MemChannel



**We have to decide which keywords tweet data to be collected from the twitter application. So, you can change the keywords in the TwitterAgent.sources.Twitter.keywords command.**

In our example, we are fetching tweet data related to Hadoop, election, sports, cricket and Big data.

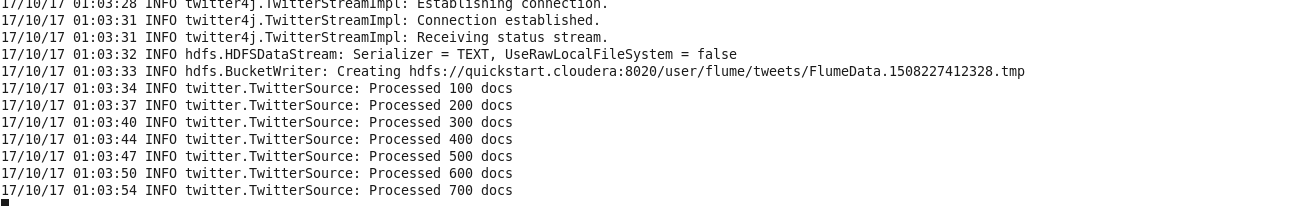
**Create a new directory inside HDFS path, where the Twitter tweet data should be stored.**



**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>

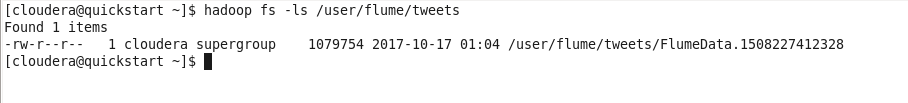




**Once, the tweet data started streaming it into the given HDFS path we can use ‘Ctrl+c’ command to stop the streaming process.**

**To check the contents of the tweet data we can use the following command:**

hadoop dfs –ls /user/flume/tweets



**We can use the ‘cat’ command to display the tweet data inside the** /user/flume/tweets/FlumeData.15\* path.

