**Session 4**

**Assignment 4.5**

Student Name: Karthik K

Course: Big Data Hadoop & Spark Training

# Problem Statement

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

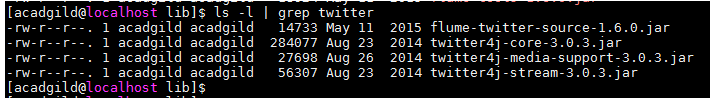
# Prerequisite

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

* Twitter account
* Hadoop cluster

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

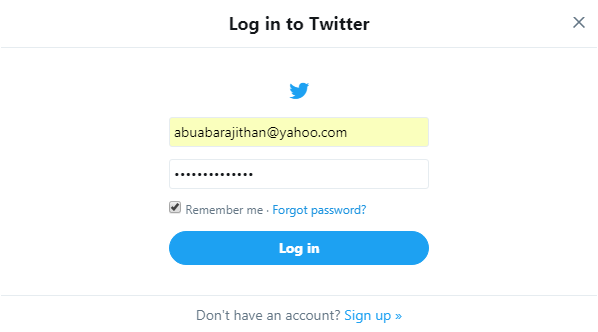
* twitter4j-core-X.XX.jar
* twitter4j-stream-X.X.X.jar
* twitter4j-media-support-X.X.X.jar



If the above prerequisites are available we can move to our further step.

# Step1:

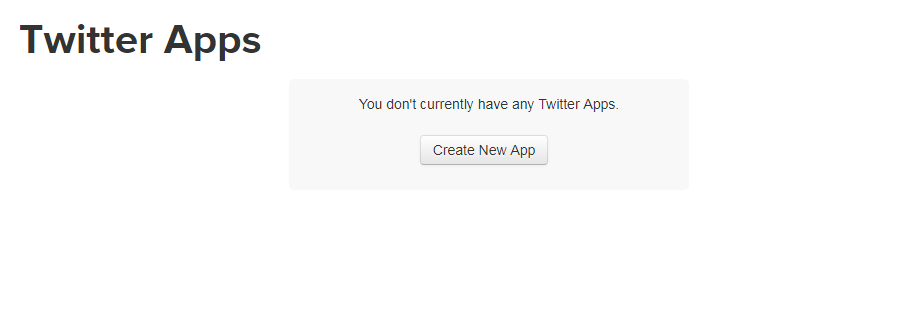
Login to the twitter account,



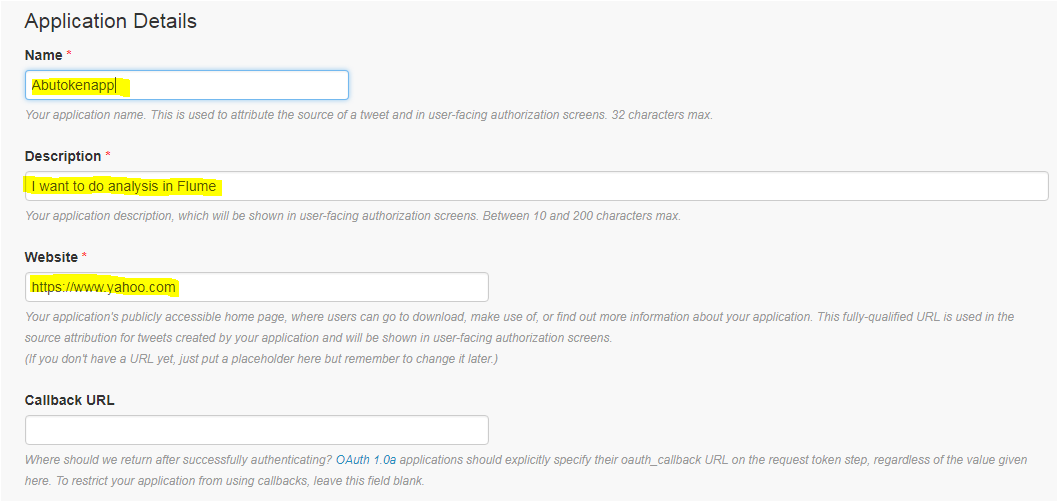
# Step2:

Go to the following link and click the ‘create new app’ button.

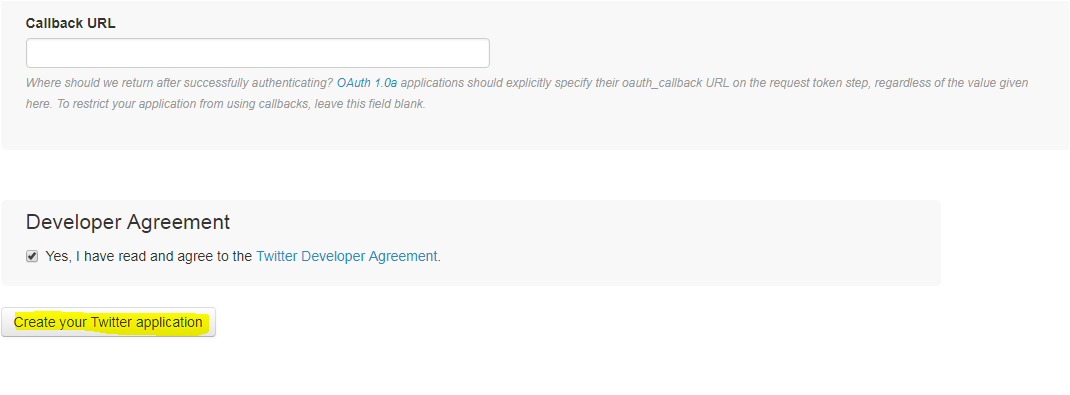
<https://apps.twitter.com/app>



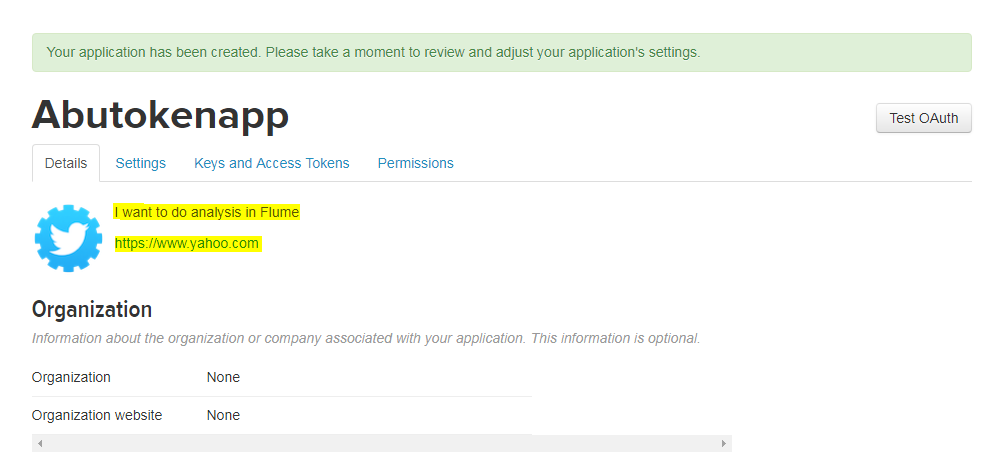
Providing necessary details,



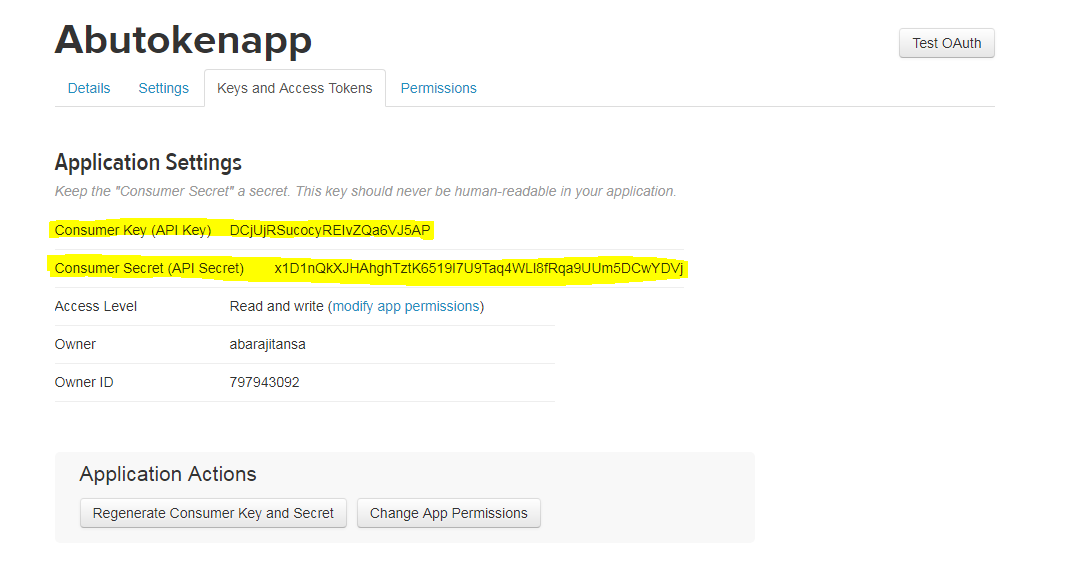
Accept the developer agreement and select the ‘create your Twitter application’ button’

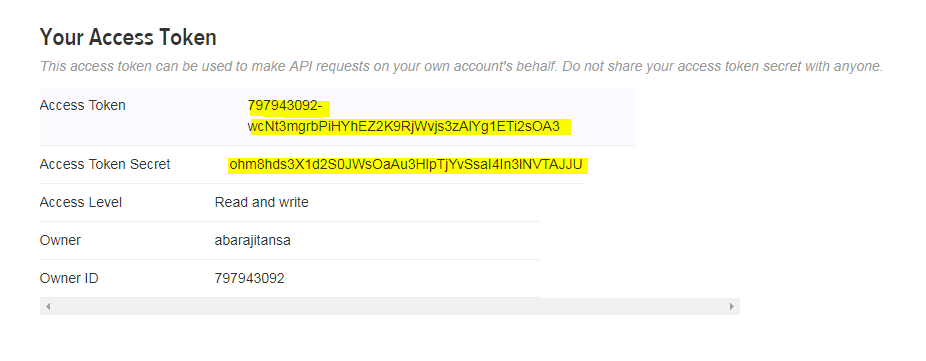
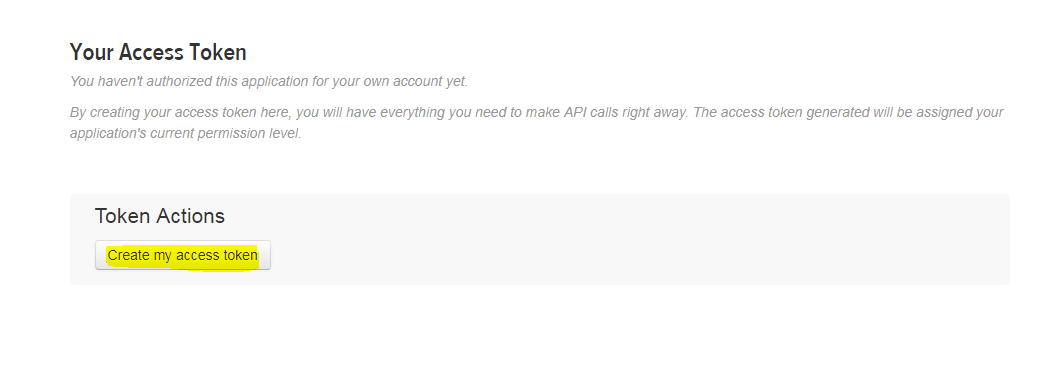


Select the ‘Keys and Access Token’ tab.

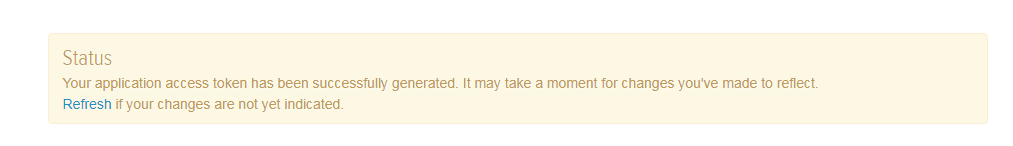


Copy the consumer key and the consumer secret code, Scroll down further and select the ‘create my access token’ button.





Now, you will receive a message stating “that you have successfully generated your application access token”.



Copy the Access Token and Access token Secret code.

Consumer Key (API Key) DCjUjRSucocyREIvZQa6VJ5AP

Consumer Secret (API Secret) x1D1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj

Access Token 797943092-wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYg1ETi2sOA3

Access Token Secret ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsaI4In3lNVTAJJU

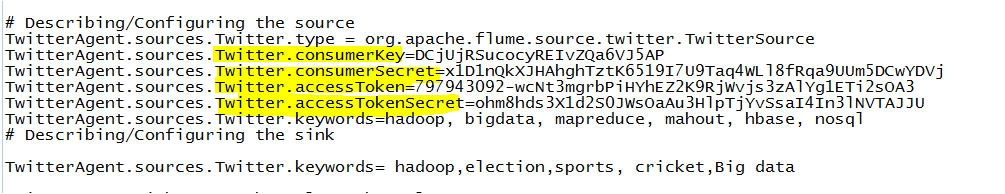
# Step 3:

Copy the Flume configuration code from the below link and paste it in the newly created file in the location,

***/home/acadgild/apache-flume-1.6.0-bin/conf/flume\_twitter.conf***

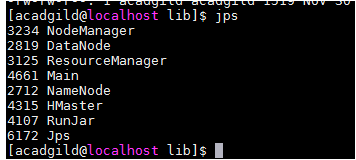
<https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk>

Update the newly created file with twitter **api** keys like consumer key, Consumer token, Access token and the access token secret code and with the **key words.**



# Step4:

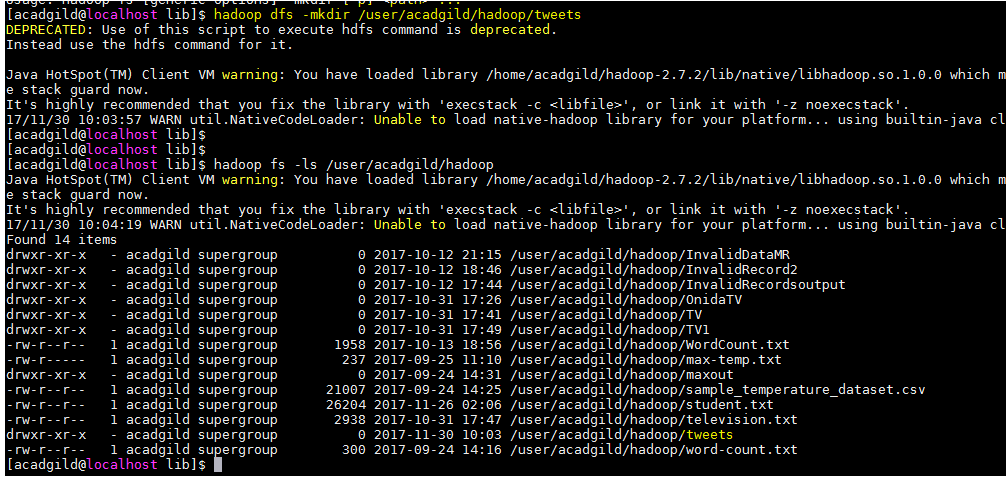
4.1 start all Hadoop daemons



# Step5:

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

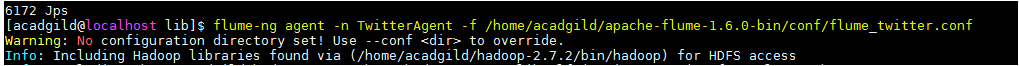
**Hadoop dfs –mkdir /user/acadgild/hadoop/tweets**



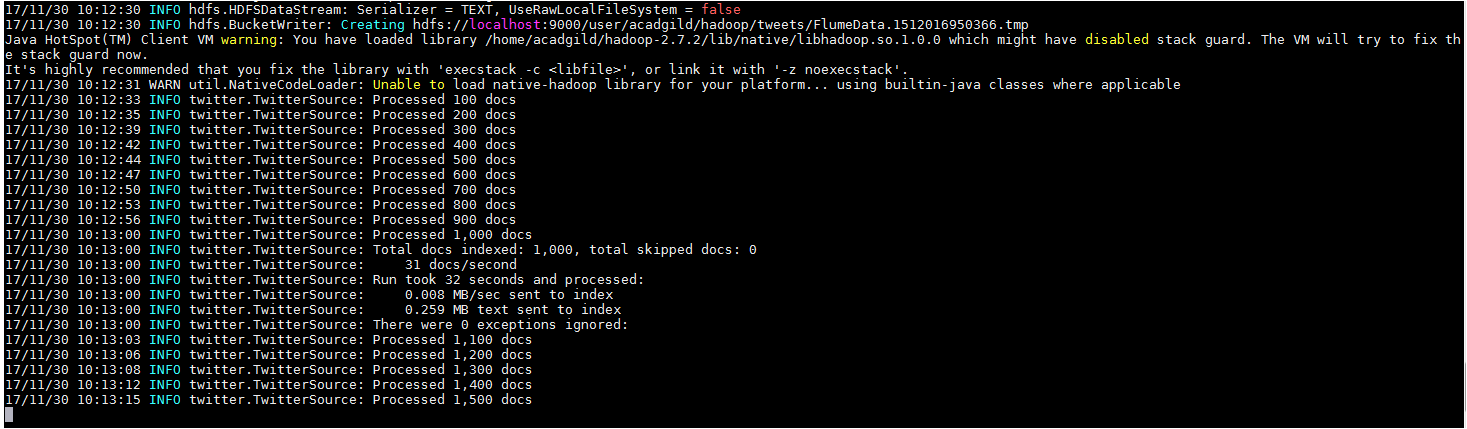
# Step6:

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 /home/acadgild/apache-flume-1.6.0-bin/conf/flume\_twitter.conf***



The above command will start fetching data from Twitter and steams it into the HDFS given path.



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

# Step7:

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

**hadoop fs -cat /user/acadgild/hadoop/tweets/FlumeData.1512016950366**



We can observe from the above image that we have successfully fetched twitter data into our HDFS cluster directory using Flume.