Introduction

In this assignment, we are going to streams data from twitter and stores into HDFS and the screen shots are shared.

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

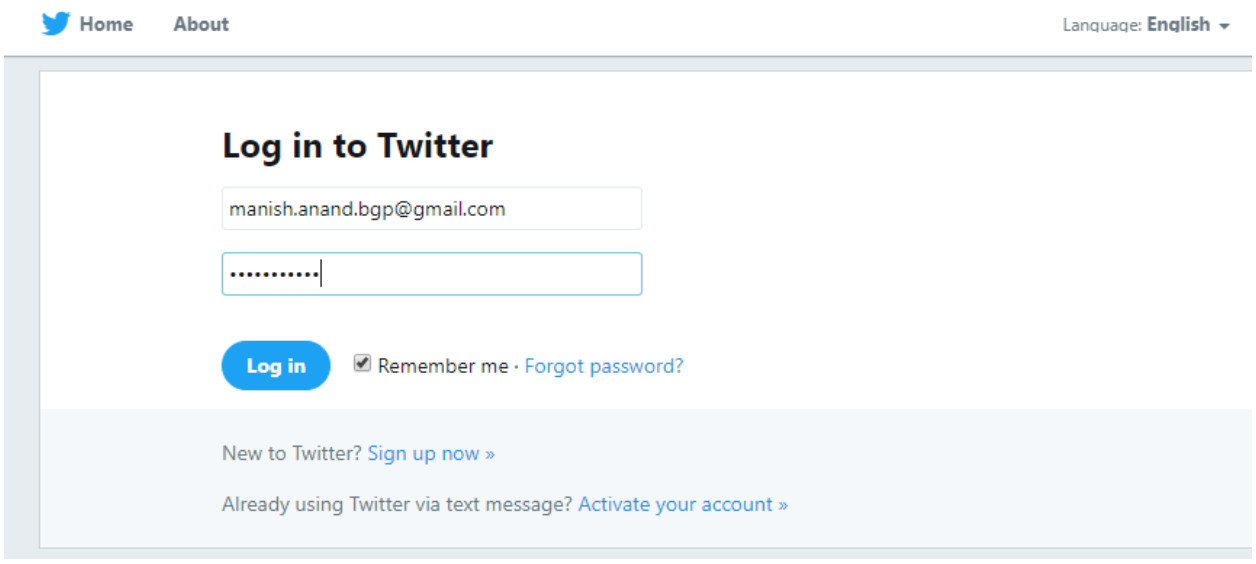
Make sure you have below jars placed in your $**FLUME_HOME/lib/conf directory**:

- 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:

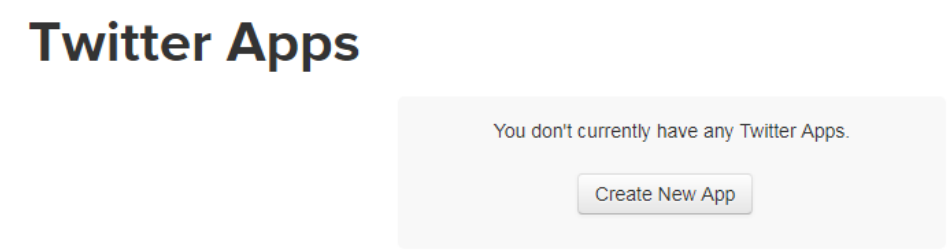
Login to the twitter account,



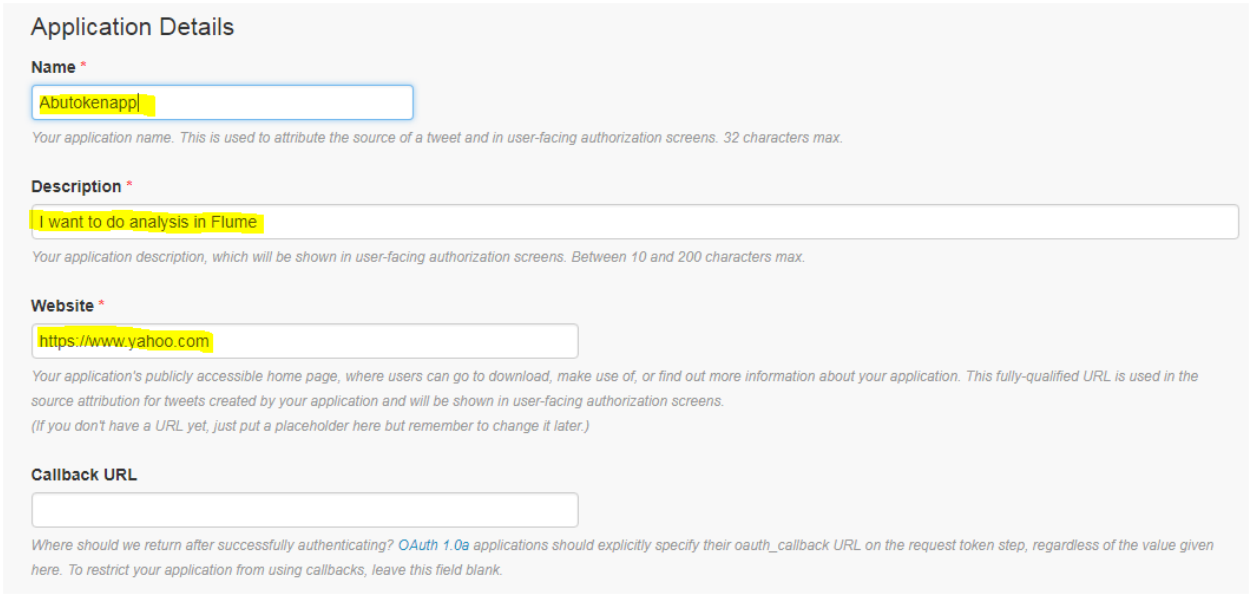
Step2:

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

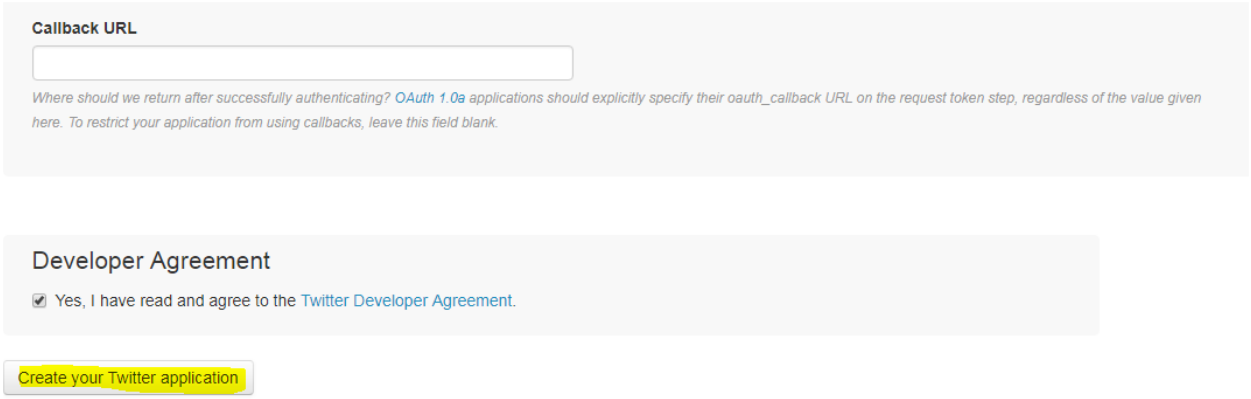
https://apps.twitter.com/app



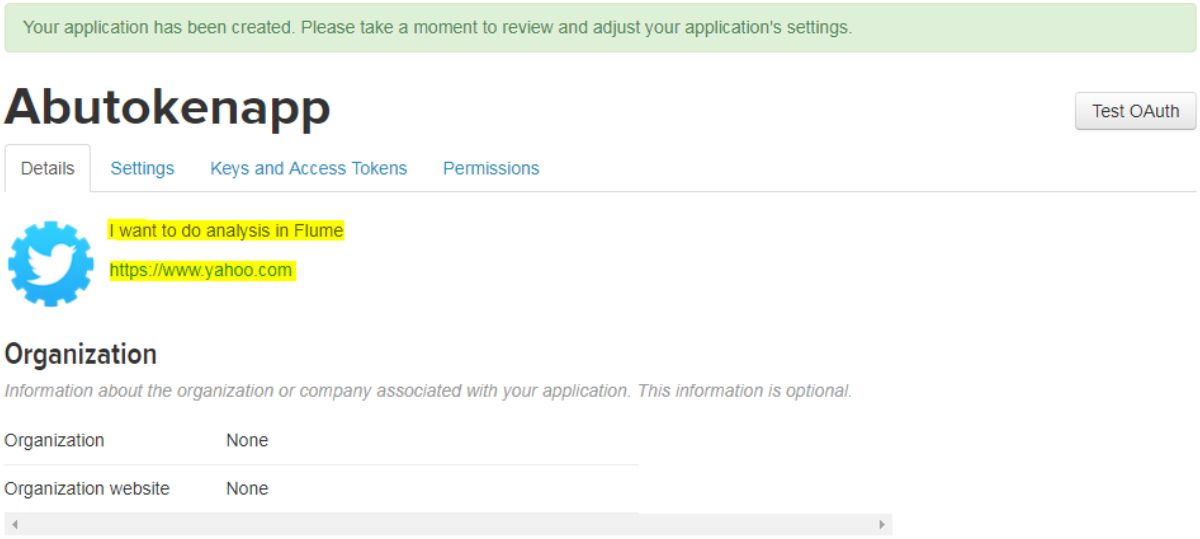
Providing necessary details,



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

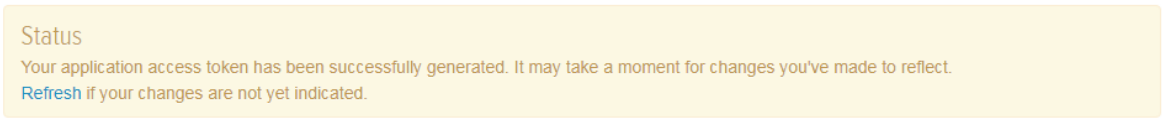


Select the 'Keys and Access Token' tab.



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 Secretohm8hds3X1d2S0JWsOaAu3HlpTjYvSsaI4In3lNVTAJJU

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.**

```
# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=DCjUjRSucocyREIvZQa6VJ5AP
TwitterAgent.sources.Twitter.consumerSecret=x1D1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj
TwitterAgent.sources.Twitter.accessToken=797943092-wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYg1ETi2sOA3
TwitterAgent.sources.Twitter.accessTokenSecret=ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsaI4In3lNVTAJJU
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink

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

Step4:

4.1 start all Hadoop daemons

```
[acadgild@localhost lib]$ jps
3234 NodeManager
2819 DataNode
3125 ResourceManager
4661 Main
2712 NameNode
4315 HMaster
4107 RunJar
6172 Jps
[acadgild@localhost lib]$
```

Step5:

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

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

```
[acadgild@localhost lib]$ hadoop dfs -mkdir /user/acadgild/hadoop/tweets
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
e stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/11/30 10:03:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
[acadgild@localhost lib]$
[acadgild@localhost lib]$
[acadgild@localhost lib]$ hadoop fs -ls /user/acadgild/hadoop
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which m
e stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/11/30 10:04:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
Found 14 items
drwxr-xr-x   - acadgild supergroup          0 2017-10-12 21:15 /user/acadgild/hadoop/InvalidDataMR
drwxr-xr-x   - acadgild supergroup          0 2017-10-12 18:46 /user/acadgild/hadoop/InvalidRecord2
drwxr-xr-x   - acadgild supergroup          0 2017-10-12 17:44 /user/acadgild/hadoop/InvalidRecordsoutput
drwxr-xr-x   - acadgild supergroup          0 2017-10-31 17:26 /user/acadgild/hadoop/OnidaTV
drwxr-xr-x   - acadgild supergroup          0 2017-10-31 17:41 /user/acadgild/hadoop/TV
drwxr-xr-x   - acadgild supergroup          0 2017-10-31 17:49 /user/acadgild/hadoop/TV1
-rw-r--r--   1 acadgild supergroup       1958 2017-10-13 18:56 /user/acadgild/hadoop/WordCount.txt
-rw-r-----   1 acadgild supergroup        237 2017-09-25 11:10 /user/acadgild/hadoop/max-temp.txt
drwxr-xr-x   - acadgild supergroup          0 2017-09-24 14:31 /user/acadgild/hadoop/maxout
-rw-r--r--   1 acadgild supergroup      21007 2017-09-24 14:25 /user/acadgild/hadoop/sample_temperature_dataset.csv
-rw-r--r--   1 acadgild supergroup      26204 2017-11-26 02:06 /user/acadgild/hadoop/student.txt
-rw-r--r--   1 acadgild supergroup       2938 2017-10-31 17:47 /user/acadgild/hadoop/television.txt
drwxr-xr-x   - acadgild supergroup          0 2017-11-30 10:03 /user/acadgild/hadoop/tweets
-rw-r--r--   1 acadgild supergroup        300 2017-09-24 14:16 /user/acadgild/hadoop/word-count.txt
[acadgild@localhost lib]$
```

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

```
6172 Jps
[acadgild@localhost lib]$ flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.6.0-bin/conf/flume_twitter.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/hadoop-2.7.2/bin/hadoop) for HDFS access
```

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

```
17/11/30 10:12:30 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
17/11/30 10:12:30 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/acadgild/hadoop/tweets/FlumeData.1512016950366.tmp
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop-2.7.2/lib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix th
e stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>', or link it with '-z noexecstack'.
17/11/30 10:12:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/30 10:12:33 INFO twitter.TwitterSource: Processed 100 docs
17/11/30 10:12:35 INFO twitter.TwitterSource: Processed 200 docs
17/11/30 10:12:39 INFO twitter.TwitterSource: Processed 300 docs
17/11/30 10:12:42 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:44 INFO twitter.TwitterSource: Processed 500 docs
17/11/30 10:12:47 INFO twitter.TwitterSource: Processed 600 docs
17/11/30 10:12:50 INFO twitter.TwitterSource: Processed 700 docs
17/11/30 10:12:53 INFO twitter.TwitterSource: Processed 800 docs
17/11/30 10:12:56 INFO twitter.TwitterSource: Processed 900 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed 1,000 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
17/11/30 10:13:00 INFO twitter.TwitterSource:     31 docs/second
17/11/30 10:13:00 INFO twitter.TwitterSource: Run took 32 seconds and processed:
17/11/30 10:13:00 INFO twitter.TwitterSource:     0.008 MB/sec sent to index
17/11/30 10:13:00 INFO twitter.TwitterSource:     0.259 MB text sent to index
17/11/30 10:13:00 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/11/30 10:13:03 INFO twitter.TwitterSource: Processed 1,100 docs
17/11/30 10:13:06 INFO twitter.TwitterSource: Processed 1,200 docs
17/11/30 10:13:08 INFO twitter.TwitterSource: Processed 1,300 docs
17/11/30 10:13:12 INFO twitter.TwitterSource: Processed 1,400 docs
17/11/30 10:13:15 INFO twitter.TwitterSource: Processed 1,500 docs
```

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.