Assignment-12

1. **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

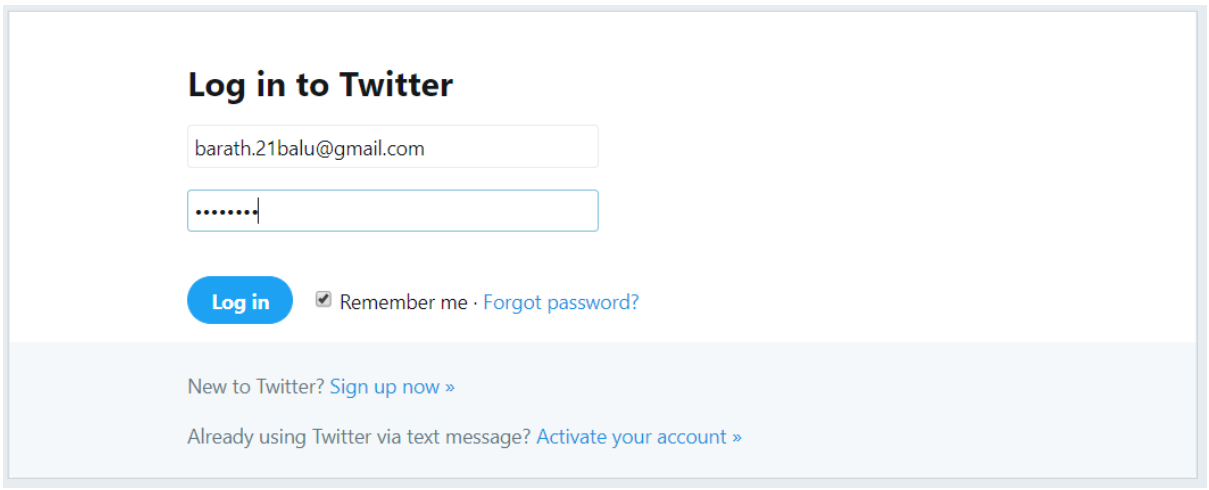
Make sure you have below jars placed in your $**FLUME_HOME/lib/conf directory**:

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

```
[acadgild@localhost lib]$ ls -l | grep twitter
-rw-r--r--. 1 acadgild acadgild   14733 May 11  2015 flume-twitter-source-1.6.0.jar
-rw-r--r--. 1 acadgild acadgild  284077 Aug 23  2014 twitter4j-core-3.0.3.jar
-rw-r--r--. 1 acadgild acadgild   27698 Aug 26  2014 twitter4j-media-support-3.0.3.jar
-rw-r--r--. 1 acadgild acadgild   56307 Aug 23  2014 twitter4j-stream-3.0.3.jar
[acadgild@localhost lib]$
```

Step-1:

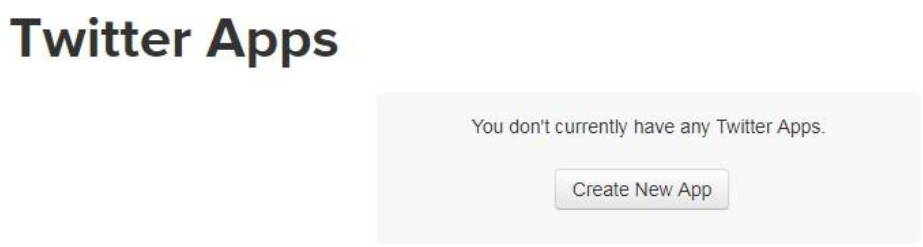
Login to twitter account.



Step 2:

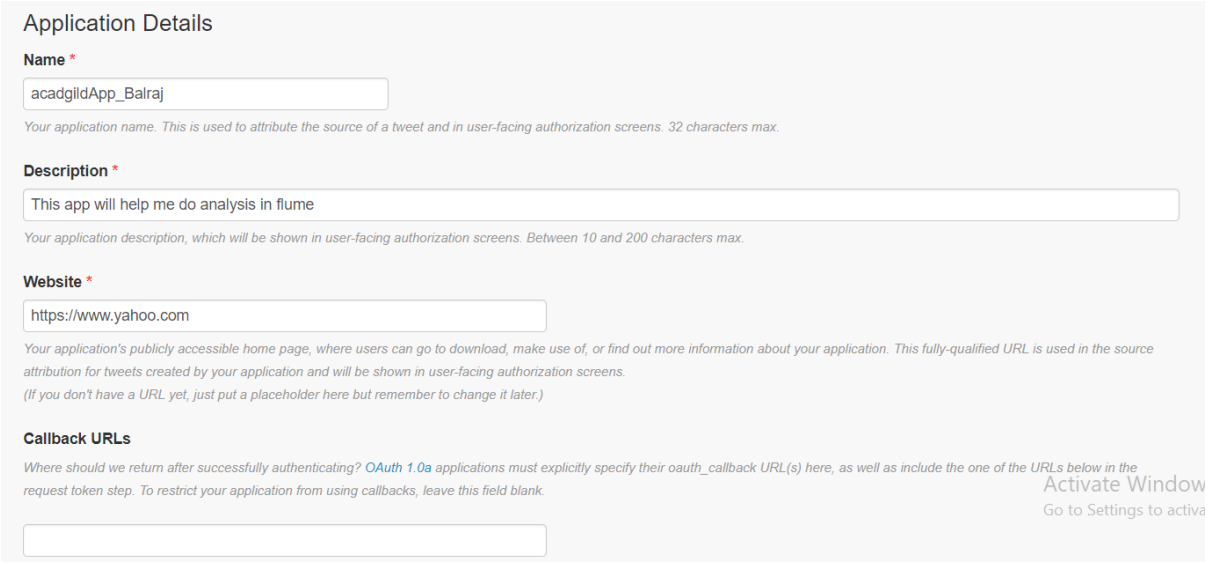
Go to the following link and click the ‘create new app’ button.

https://apps.twitter.com/app

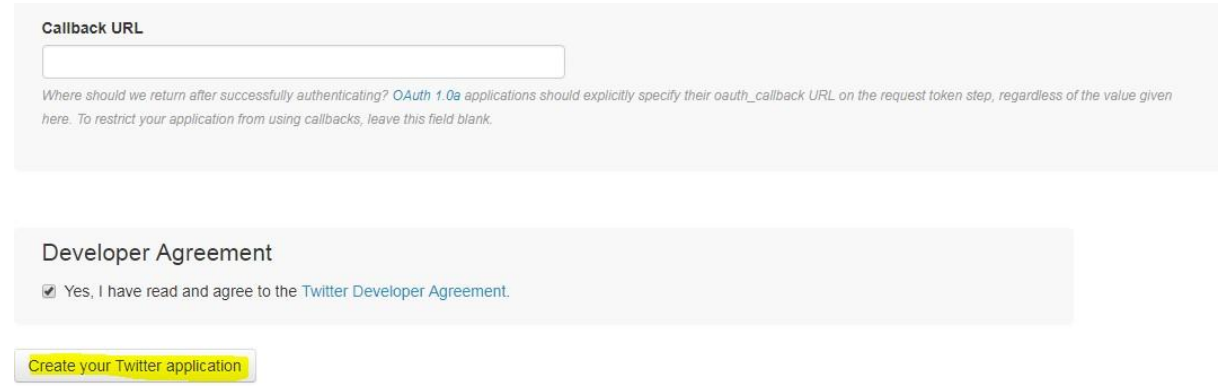


Step 3:

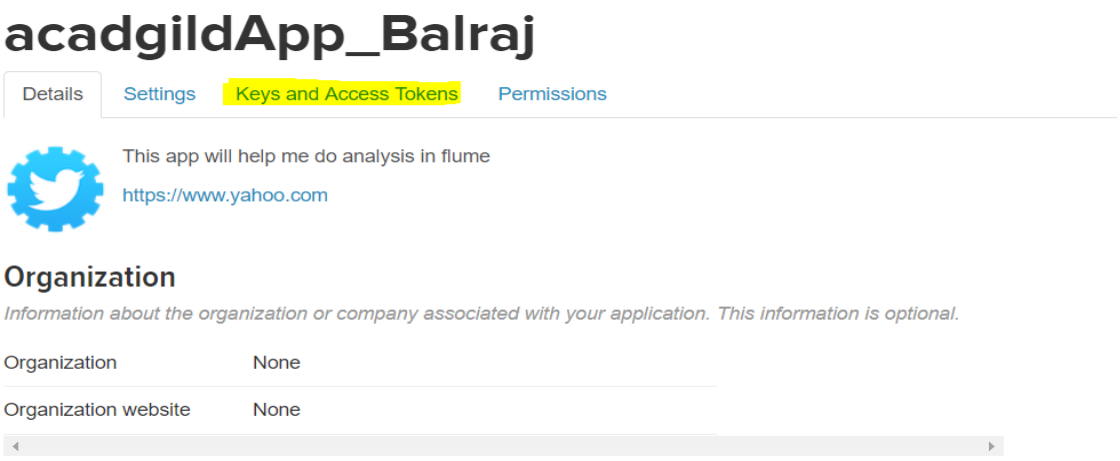
Provide the necessary details



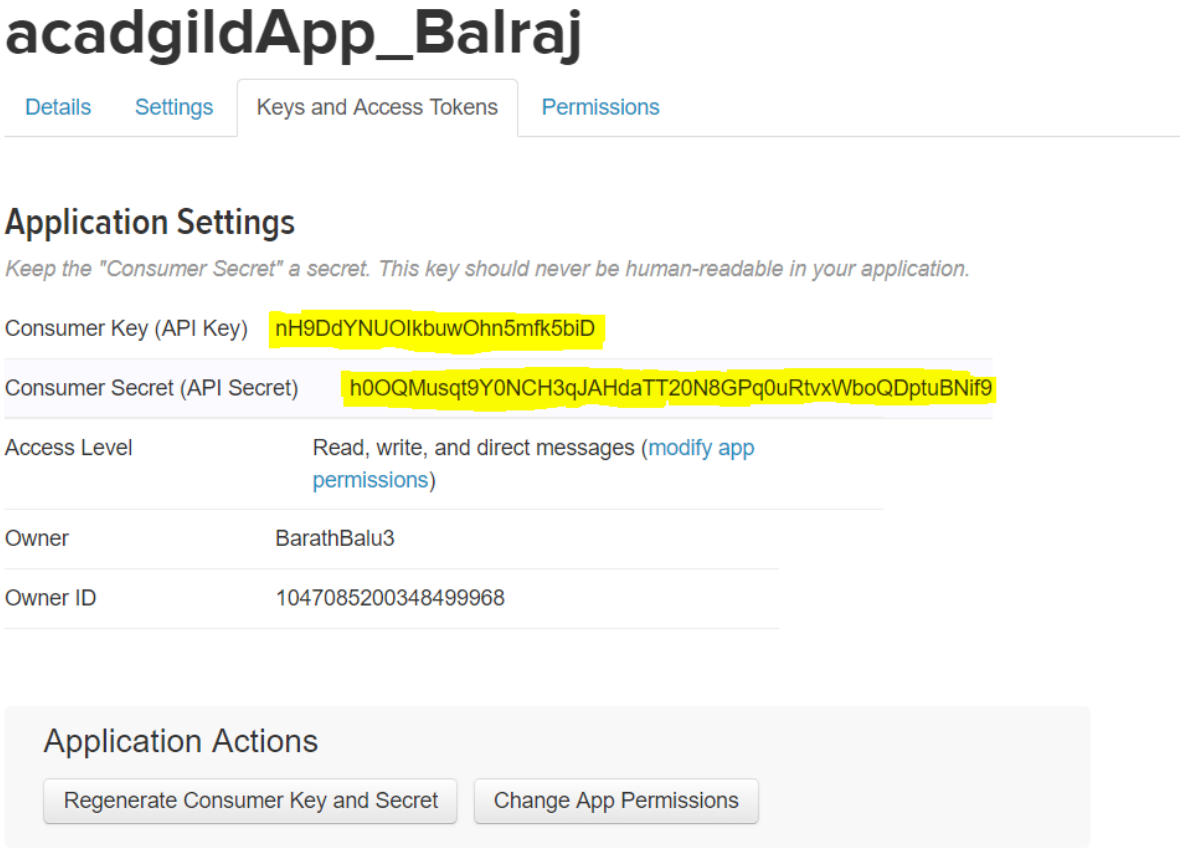
Accept the developer agreement and select the ‘create your Twitter application’ button’.

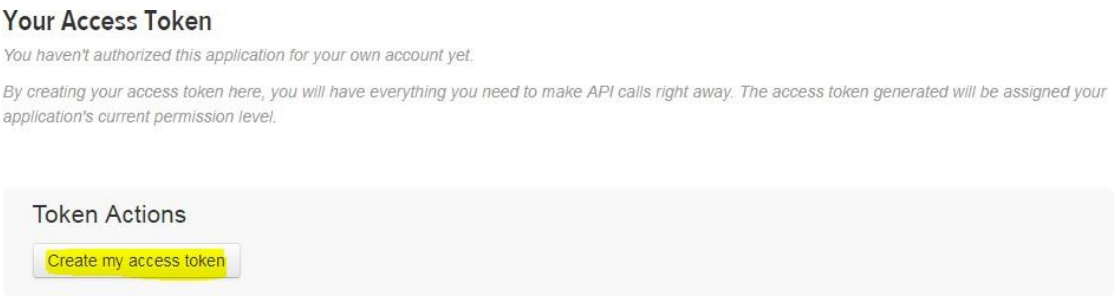


Select the ‘Keys and Access Token’ tab.

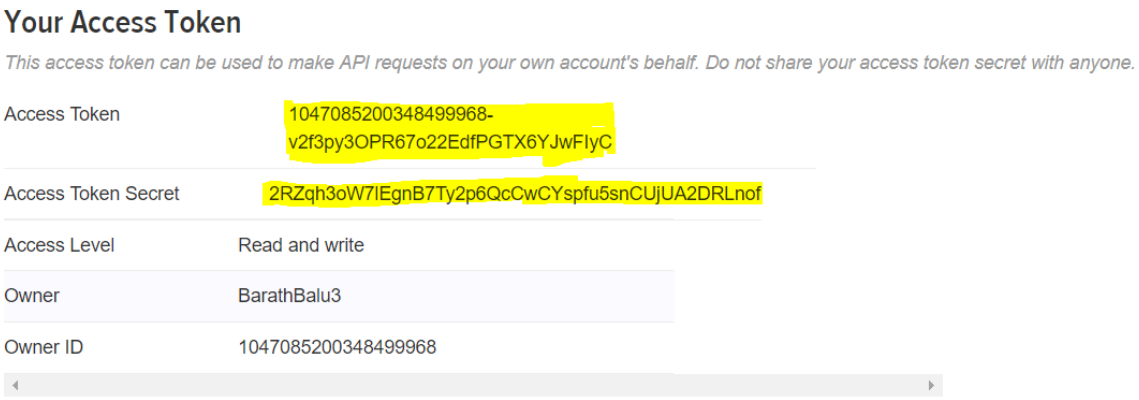


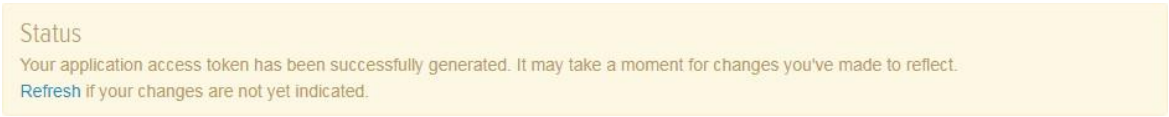
Copy the consumer key and the consumer secret code, Scroll down further and select the 'create my access token' button.





Now, you will get 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) nH9DdYNUOIkbuwOhn5mfk5biD

Consumer Secret (API Secret) h0OQMusqt9Y0NCH3qJAHdaTT20N8GPq0uRtvxWboQDptuBNif9

Access Token 1047085200348499968-v2f3py3OPR67o22EdfPGTX6YJwFIyC

Access Token Secret 2RZqh3oW7lEgnB7Ty2p6QcCwCYspfu5snCUjUA2DRLnof

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.**

flume_twitter.conf

```
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=nH9DdYNUOIkbuwOhn5mfk5biD
TwitterAgent.sources.Twitter.consumerSecret=h0OQMusqt9Y0NCH3qJAHdaTT20N8GPq0uRtvxWboQDptuBNif9
TwitterAgent.sources.Twitter.accessToken=1047085200348499968-v2f3py3OPR67o22EdfPGTX6YJwFIyC
TwitterAgent.sources.Twitter.accessTokenSecret=2RZqh3oW7lEgnB7Ty2p6QcCwCYspfu5snCUjUA2DRLnof
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink

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

Step 4:

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]$
```

Step 5:

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

```
[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
```

Step 6:

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

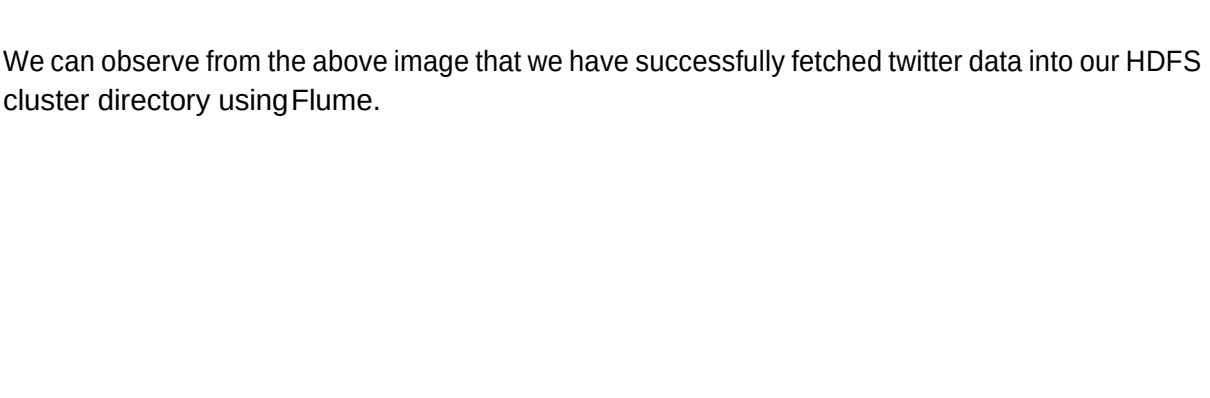
```
6172 Jps
[acadgild@localhost lib]$ flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.6.0-bin/conf/flume_twitter.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/hadoop-2.7.2/bin/hadoop) for HDFS access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

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
```

Step 7:

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



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