

## Problem Statement

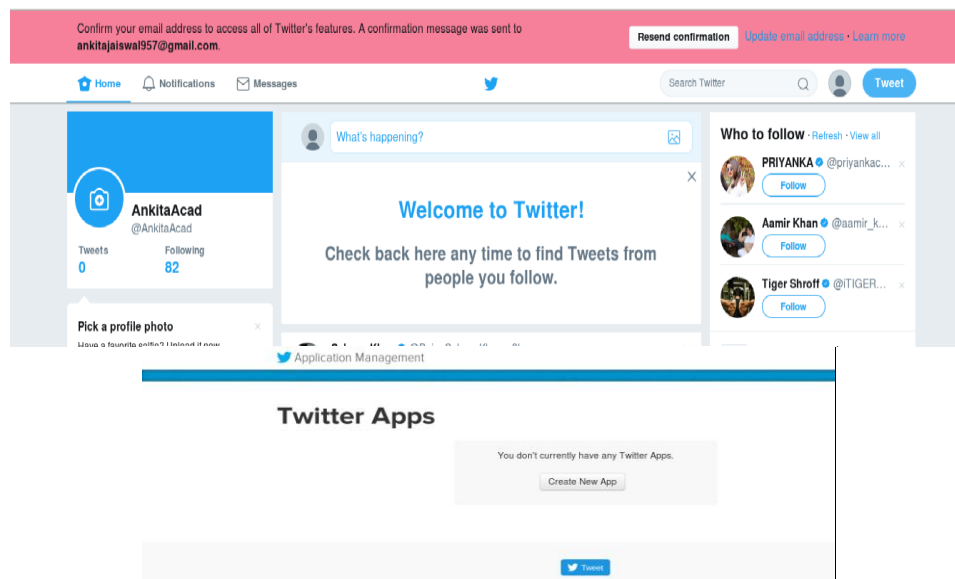
Create a flume agent that streams data from Twitter and stores in the HDFS.

## Solution:

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

- **Twitter account**
- **Hadoop cluster**

**Step 1:** Login to Twitter Account and click the 'create new app' button at <https://apps.twitter.com/app>

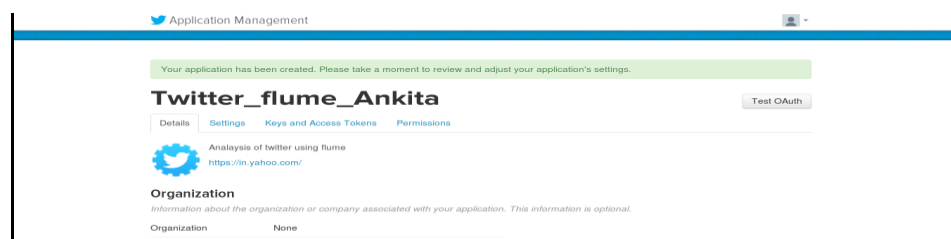


**Step 2:** Create a Twitter application by entering the appropriate details.

## Create an application

The image shows a screenshot of the 'Create New App' form on Twitter. The form is titled 'Application Details' and has three main sections: 'Name', 'Description', and 'Website'. The 'Name' field contains 'Twitter\_flume\_Ankita'. The 'Description' field contains 'Analysis of twitter using flume'. The 'Website' field contains 'https://in.yahoo.com'. There are also links for 'Resend confirmation', 'Update email address', and 'Learn more' at the top right of the form.

**Step 3:** Copy the **consumer key** and the **consumer secret code** and create and copy the **access tokens**.



## Application Settings

Your application's Consumer Key and Secret are used to [authenticate](#) requests to the Twitter Platform.

Access level	Read and write ( <a href="#">modify app permissions</a> )
Consumer Key (API Key)	xRWvRIgVgdsd8ca3b10Fk0LTc ( <a href="#">manage keys and access tokens</a> )
Callback URL	None
Callback URL Locked	No
Sign in with Twitter	Yes
App-only authentication	https://api.twitter.com/oauth2/token
Request token URL	https://api.twitter.com/oauth/request_token
Authorize URL	https://api.twitter.com/oauth/authorize
Access token URL	https://api.twitter.com/oauth/access_token

### Status

Your application access token has been successfully generated. It may take a moment for changes you've made to reflect. [Refresh](#) if your changes are not yet indicated.

## Twitter\_flume\_Ankita

[Test OAuth](#)[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

### Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	xRWvRIgVgdsd8ca3b10Fk0LTc
Consumer Secret (API Secret)	lyDZh3U36ErGvC0ad7VVT2oiPqTGC8VQgu3vzzurH5SbeVYolj
Access Level	Read and write ( <a href="#">modify app permissions</a> )
Owner	AnkitaAcad

### Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

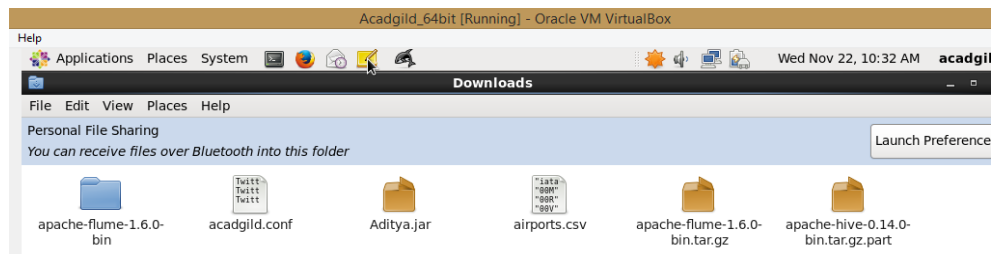
Access Token	932907924732526592-ABtkl1j4m96c8wawjBXlseLW6FIzi1L
Access Token Secret	3RjNP4bkXkj0hq1UEGJN0N25bklwvdSJsGLwSUW2Y0sBh
Access Level	Read and write
Owner	AnkitaAcad
Owner ID	932907924732526592

### Token Actions

[Regenerate My Access Token and Token Secret](#)[Revoke Token Access](#)

## Step 4:

- Add Apache Flume to the AcadGild VM and extracting the .tar file as below:



- Now we have to edit the path of the **.bashrc** file. This will update the path of the extracted flume directory. To do that we use the below command. This will open the **.bashrc** file for us to edit.

```
acadmild@localhost:~  
File Edit View Search Terminal Help  
[acadmild@localhost ~]$ pwd  
/home/acadmild  
[acadmild@localhost ~]$ sudo gedit .bashrc  
[sudo] password for acadmild: █
```

- Now we edit the FLUME\_HOME path

```

*.bashrc (/home/acadgild) - gedit
File Edit View Search Tools Documents Help
*.bashrc
export JAVA_HOME=/usr/local/java
export PATH=$PATH:$JAVA_HOME/bin

export HADOOP_HOME=/usr/local/hadoop-2.6.0
export PATH=$PATH:$HADOOP_HOME/bin

export FLUME_HOME=/home/acadgild/Downloads/apache-flume-1.6.0-bin
export PATH=$PATH:$FLUME_HOME/bin

export PIG_INSTALL=/usr/local/pig

export HIVE_HOME=/usr/local/hive
export PATH=$HIVE_HOME/bin:$PATH
export HADOOP_USER_CLASSPATH_FIRST=true

export HBASE_HOME=/usr/local/hbase

export SQOOP_HOME=/usr/local/sqoop

export PATH=$PATH:$FLUME_HOME/bin:$PIG_INSTALL/bin:$HIVE_HOME/bin:$HBASE_HOME/bin:$SQOOP_HOME/bin:

```

- After saving and closing the .bashrc file, we can update the file by using the below command

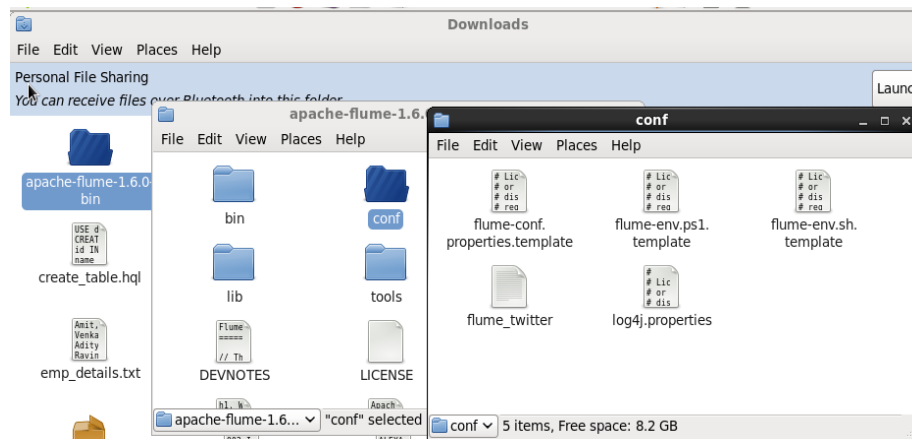
```

acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ sudo gedit .bashrc
[sudo] password for acadgild:
[acadgild@localhost ~]$ source .bashrc
[acadgild@localhost ~]$

```

## Step 5:

- Now, we have to create a new file in the **conf** directory of the apache flume extracted tar file I have named it **flume\_twitter**

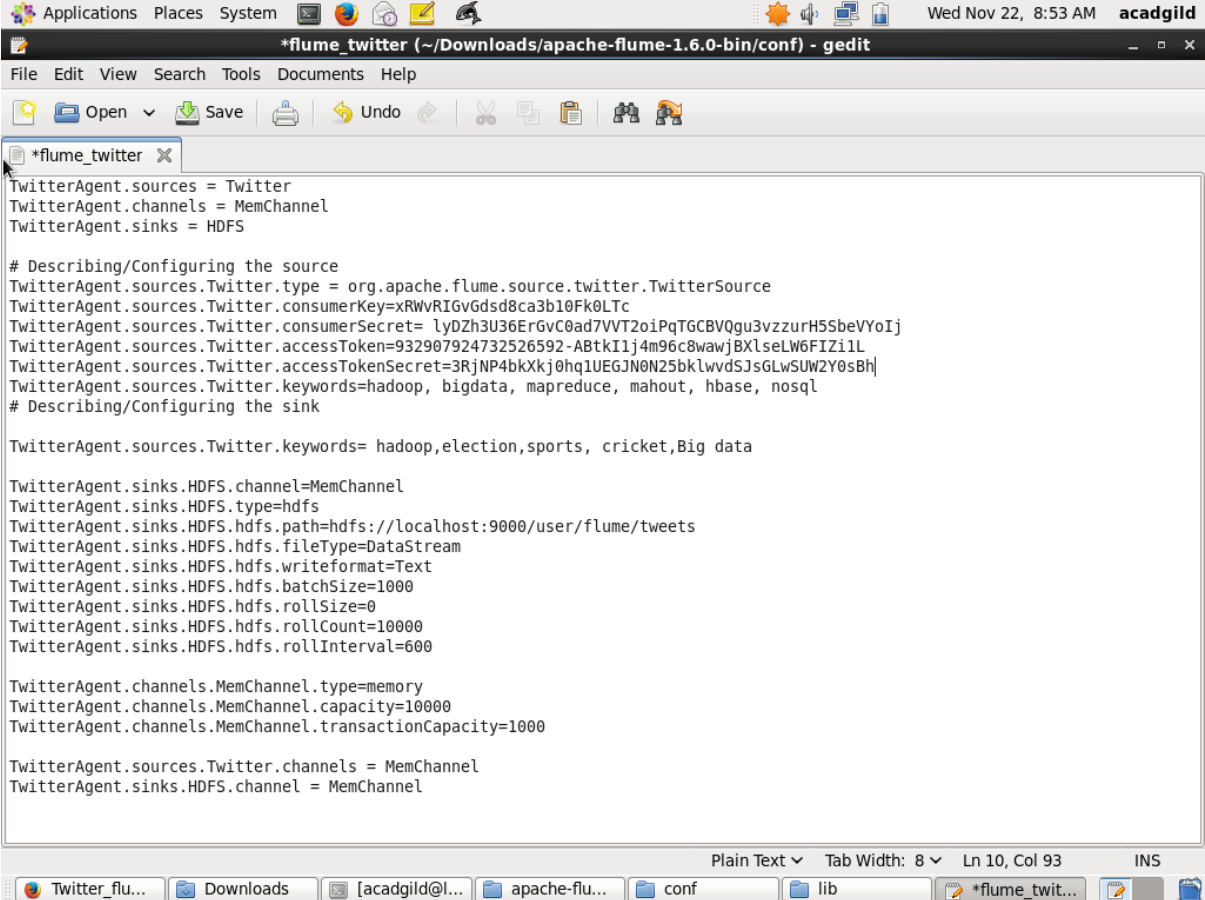


- We have to make sure that the below jars placed in the **lib** directory of the extracted Apache Flume jar:
  - twitter4j-core-X.XX.jar
  - twitter4j-stream-X.X.X.jar
  - twitter4j-media-support-X.X.X.jar



## Step 6:

- Now we have to write the configuration file for the Twitter Streaming. We use the newly created file `flume_twitter`. Below is a snapshot of the file.



```
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=xRWvRIGvGdsd8ca3b10Fk0LTc
TwitterAgent.sources.Twitter.consumerSecret= lyDZh3U36ErGvC0ad7VVT2oiPqTGCbVQgu3vzzurH5SbeVYoIj
TwitterAgent.sources.Twitter.accessToken=932907924732526592-ABtkI1j4m96c8wawjBXLseLW6FIZi1L
TwitterAgent.sources.Twitter.accessTokenSecret=3RjNP4bkXkj0hq1UEGJN0N25bklwvdSJsGLwSUW2Y0sBh
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink

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

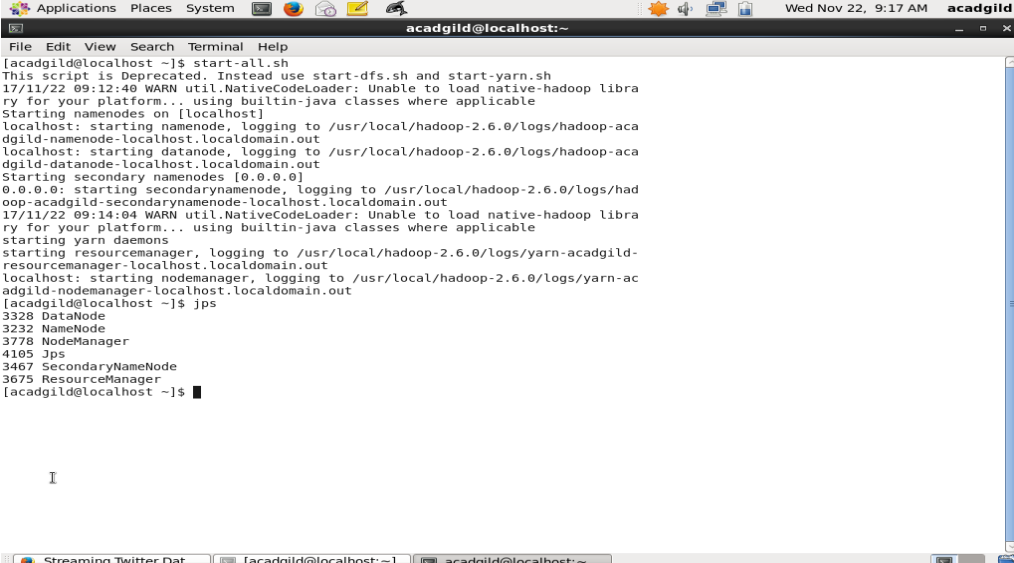
TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/user/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
TwitterAgent.sinks.HDFS.hdfs.rollSize=0
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
```

## Step 7:

- Now we have to execute the flume agent and configuration file by the following steps:
  - Start the terminal and start all hadoop daemons and run jps command.



```
acadmild@localhost:~
File Edit View Search Terminal Help
[acadmild@localhost ~]$ start-all.sh
This script is deprecated. Instead use start-dfs.sh and start-yarn.sh
17/11/22 09:12:40 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgmild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgmild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/had
oop-acadmild-secondarynamenode-localhost.localdomain.out
17/11/22 09:14:04 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadmild-
resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-ac
admild-nodemanager-localhost.localdomain.out
[acadmild@localhost ~]$ jps
3328 DataNode
3232 NameNode
3778 NodeManager
4105 Jps
3467 SecondaryNameNode
3675 ResourceManager
[acadmild@localhost ~]$
```

- My directories path is: `/user/flume/tweets/`

- Running the flume twitter agent using the below command and specifying the path of the flume configuration file: `flume-twitter`. This will start streaming the twitter data into the given HDFS path. To stop the streaming process we use '`Ctrl+C`' command

```

acagdild@localhost:~
File Edit View Search Terminal Help
  at java.io.BufferedReader.readLine(BufferedReader.java:324)
  at java.io.BufferedReader.readLine(BufferedReader.java:389)
  at twitter4j.StatusStreamBase.handleNextElement(StatusStreamBase.java:85)
  ... 2 more
17/11/22 09:35:21 INFO twitter4j.TwitterStreamImpl: Waiting for 250 milliseconds
17/11/22 09:35:21 INFO twitter4j.TwitterStreamImpl: Establishing connection.
17/11/22 09:35:24 INFO twitter4j.TwitterStreamImpl: Connection established.
17/11/22 09:35:24 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
17/11/22 09:35:25 INFO twitter.TwitterSource: Processed 6,300 docs
17/11/22 09:35:28 INFO twitter.TwitterSource: Processed 6,400 docs
17/11/22 09:35:30 INFO twitter.TwitterSource: Processed 6,500 docs
17/11/22 09:35:33 INFO twitter.TwitterSource: Processed 6,600 docs
17/11/22 09:35:37 INFO twitter.TwitterSource: Processed 6,700 docs
17/11/22 09:35:41 INFO twitter.TwitterSource: Processed 6,800 docs
17/11/22 09:35:42 INFO twitter.TwitterSource: Processed 6,900 docs
17/11/22 09:35:45 INFO twitter.TwitterSource: Processed 7,000 docs
17/11/22 09:35:45 INFO twitter.TwitterSource: Total docs indexed: 7,000, total skipped docs: 0
17/11/22 09:35:45 INFO twitter.TwitterSource: 33 docs/second
17/11/22 09:35:45 INFO twitter.TwitterSource: Run took 211 seconds and processed:
17/11/22 09:35:45 INFO twitter.TwitterSource: 0.009 MB/sec sent to index
17/11/22 09:35:45 INFO twitter.TwitterSource: 1.857 MB text sent to index
17/11/22 09:35:45 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/11/22 09:35:48 INFO twitter.TwitterSource: Processed 7,100 docs
17/11/22 09:35:51 INFO twitter.TwitterSource: Processed 7,200 docs
17/11/22 09:35:54 INFO twitter.TwitterSource: Processed 7,300 docs
17/11/22 09:35:56 INFO twitter.TwitterSource: Processed 7,400 docs
17/11/22 09:35:59 INFO twitter.TwitterSource: Processed 7,500 docs
17/11/22 09:36:02 INFO twitter.TwitterSource: Processed 7,600 docs
17/11/22 09:36:05 INFO twitter.TwitterSource: Processed 7,700 docs
17/11/22 09:36:09 INFO twitter.TwitterSource: Processed 7,800 docs
17/11/22 09:36:11 INFO twitter.TwitterSource: Processed 7,900 docs
17/11/22 09:36:14 INFO twitter.TwitterSource: Processed 8,000 docs
17/11/22 09:36:14 INFO twitter.TwitterSource: Total docs indexed: 8,000, total skipped docs: 0
17/11/22 09:36:14 INFO twitter.TwitterSource: 33 docs/second
17/11/22 09:36:14 INFO twitter.TwitterSource: Run took 239 seconds and processed:
17/11/22 09:36:14 INFO twitter.TwitterSource: 0.009 MB/sec sent to index
17/11/22 09:36:14 INFO twitter.TwitterSource: 2.123 MB text sent to index
17/11/22 09:36:14 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/11/22 09:36:17 INFO twitter.TwitterSource: Processed 8,100 docs

```

```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
17/11/22 09:38:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r-- 1 acadgild supergroup 2730522 2017-11-22 09:32 /user/flume/tweets/FlumeData.151132339763.tmp
[acadgild@localhost ~]$
```