Assignment 11.3:

Problem Statement:

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

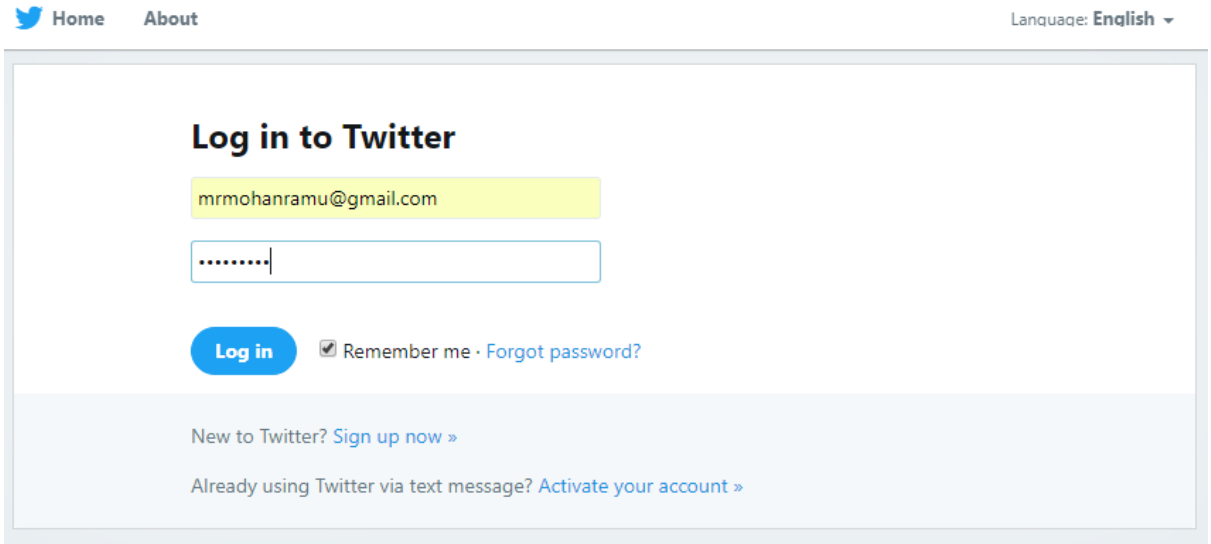
Solution/Steps:

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

- Twitter account
- Hadoop cluster

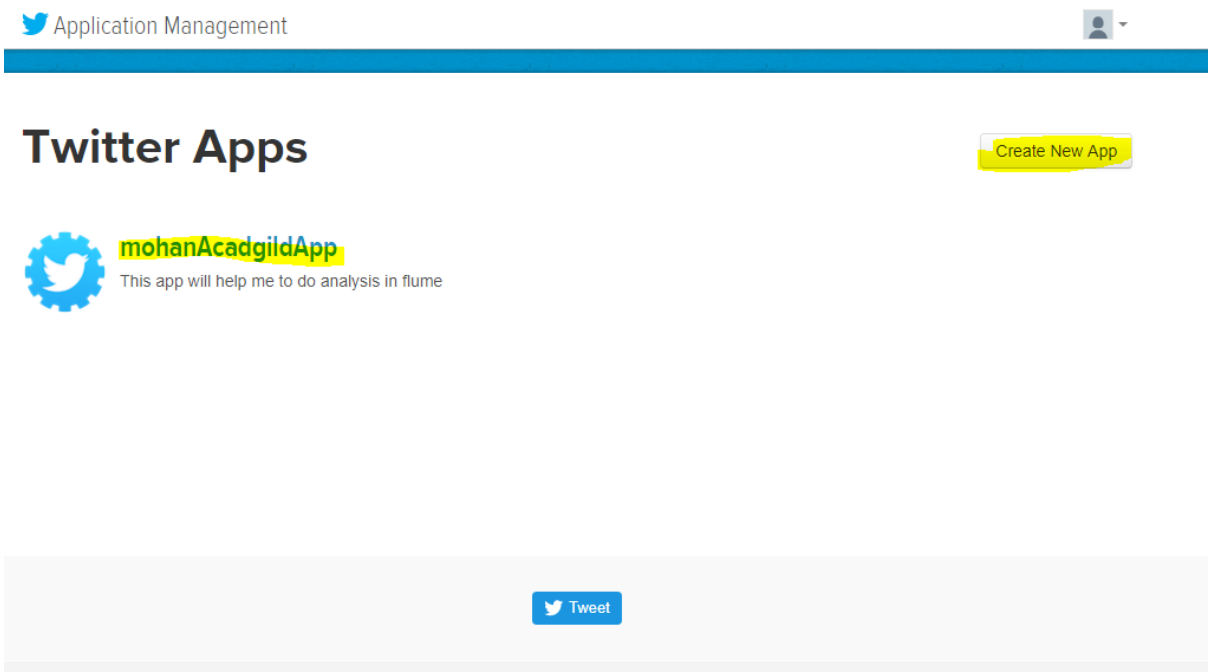
If both prerequisites are available we can move to our further step.

Step 1:
Login to the twitter account



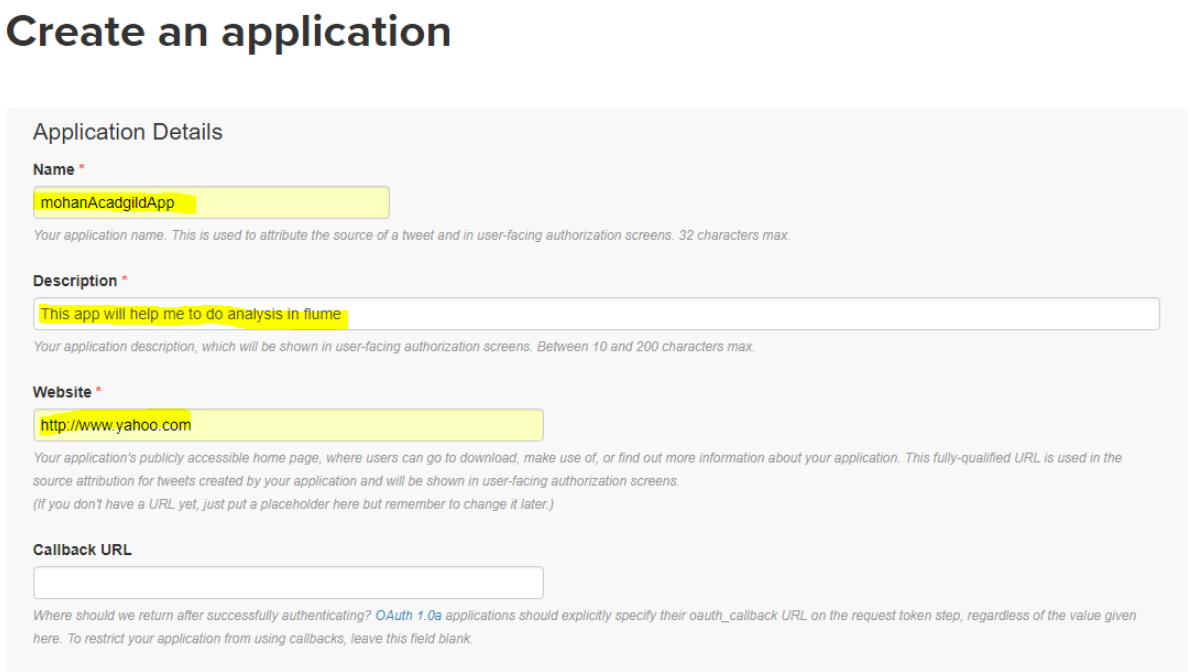
Step 2:
Go to the following link and click the 'create new app' button.

https://apps.twitter.com/app

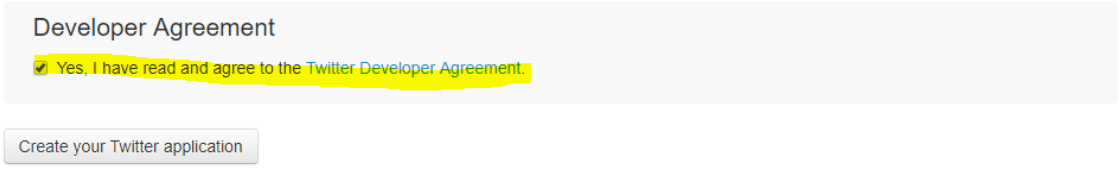


Step 3:

Enter the necessary details.

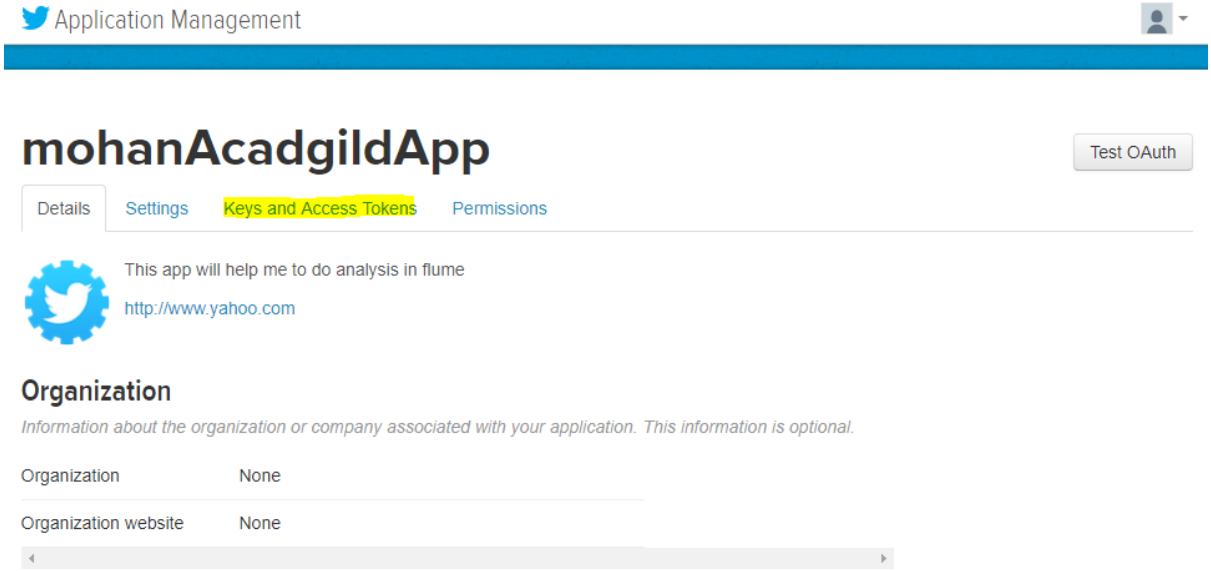


Step 4:
Accept the developer agreement and select the 'create your Twitter application' button.



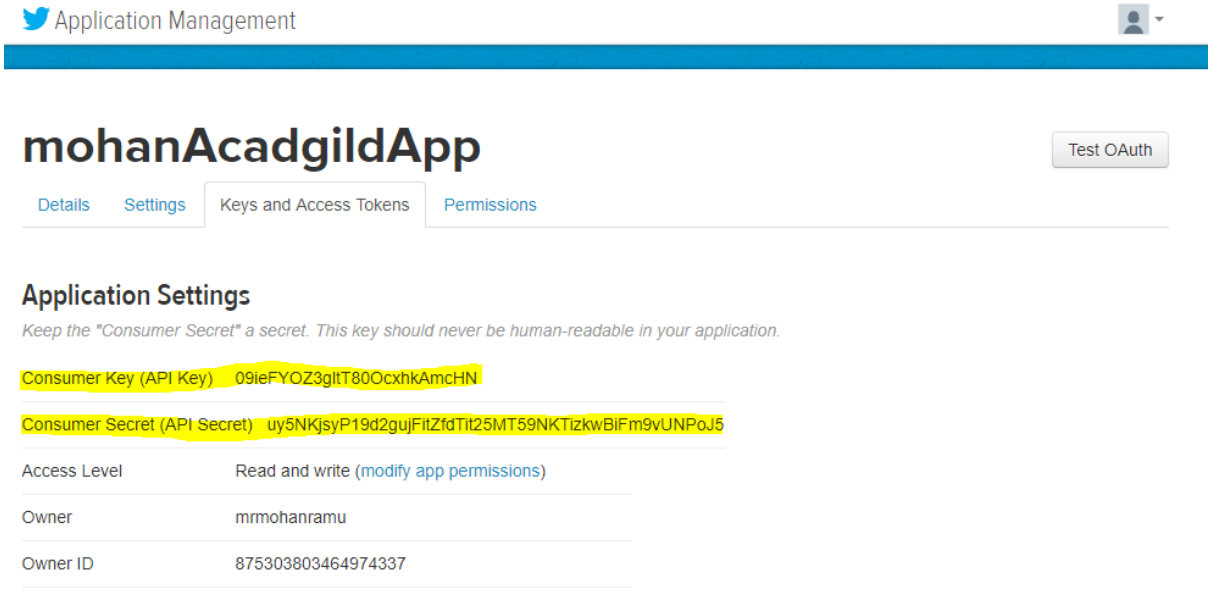
Step 5:

Select the 'Keys and Access Token' tab.



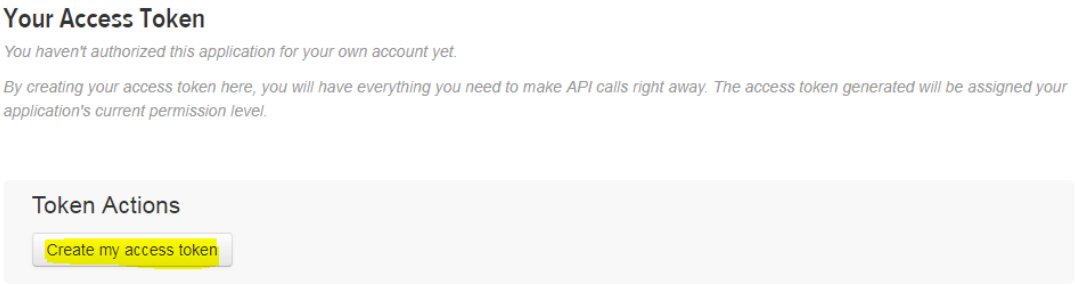
Step 6:

Copy the consumer key and the consumer secret code.

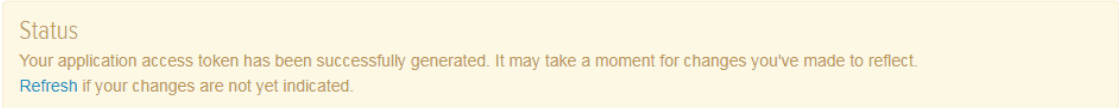


Step 7:

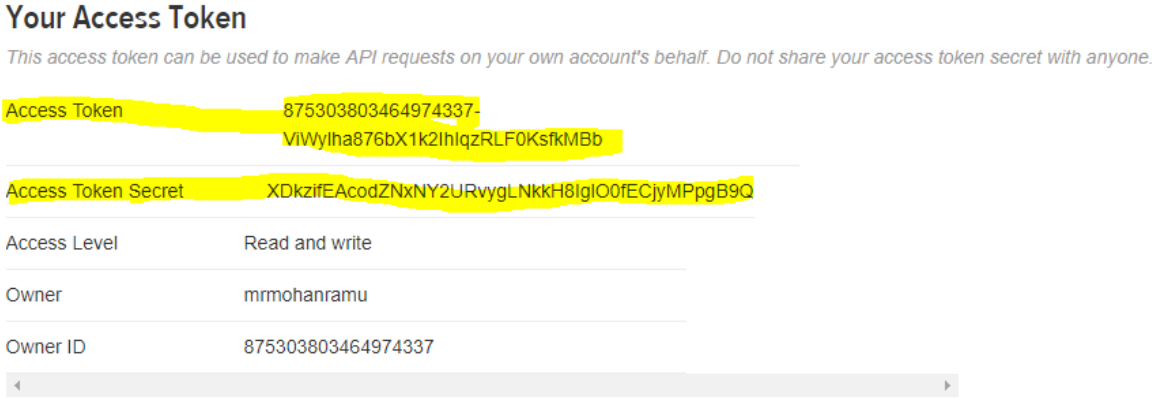
Scroll down further and select the 'create my access token' button.



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



Step 8:
Copy the Access Token and Access token Secret code.



Follow Step 9 and Step 10 to install Apache flume

Step 9: Download flume tar file from below link and extract it.

wget http://archive.apache.org/dist/flume/1.8.0/apache-flume-1.8.0-bin.tar.gz

```
[acadgild@localhost ~]$ wget http://archive.apache.org/dist/flume/1.8.0/apache-flume-1.8.0-bin.tar.gz
--2017-12-05 07:00:57--  http://archive.apache.org/dist/flume/1.8.0/apache-flume-1.8.0-bin.tar.gz
Resolving archive.apache.org... 163.172.17.199
Connecting to archive.apache.org|163.172.17.199|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 58688757 (56M) [application/x-gzip]
Saving to: “apache-flume-1.8.0-bin.tar.gz”

100%[===========================================================================================>] 58,688,757   293K/s   in 7m 15s

2017-12-05 07:08:27 (132 KB/s) - “apache-flume-1.8.0-bin.tar.gz” saved [58688757/58688757]

[acadgild@localhost ~]$ ls
apache-flume-1.8.0-bin.tar.gz                hadoop                          Public
Assignment3.2-0.0.1-SNAPSHOT.jar             hdfs:                           sample_temperature_dataset.csv
Assignment3.3-0.0.1-SNAPSHOT.jar             hive                            spark
assignmnet11.txt                             hive-site.xml                   spark-2.2.0-bin-hadoop2.7
company.java                                 mapreduce-0.0.1-SNAPSHOT.jar    spark-2.2.0-bin-hadoop2.7.tgz
customer.java                                max-temp.txt                    Sqoop_try.txt
derby.log                                    metastore_db                    sqoop.zip
Desktop                                      Music                           television.txt
Documents                                    mysql-connector-java-5.1.44.zip Templates
Downloads                                    Pictures                        Videos
eclipse                                      pig                             workspace
eclipse-jee-neon-M3-linux-gtk-x86_64.tar.gz  pig_1510851976774.log
employee.java                                pig_1511973662738.log
[acadgild@localhost ~]$ tar zxvf apache-flume-1.8.0-bin.tar.gz
apache-flume-1.8.0-bin/lib/flume-ng-configuration-1.8.0.jar
apache-flume-1.8.0-bin/lib/slf4j-api-1.6.1.jar
apache-flume-1.8.0-bin/lib/slf4j-log4j12-1.6.1.jar
apache-flume-1.8.0-bin/lib/log4j-1.2.17.jar
```

Update the path of extracted flume directory in the .bashrc file as mentioned in the below image.
NOTE: keep the path same as where the extracted file exists.

```
[acadgild@localhost ~]$ cat .bashrc
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# User specific aliases and functions

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=/usr/local/flume

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 OOZIE_HOME=/usr/local/oozie-4.1.0

export FLUME_HOME=/home/acadgild/apache-flume-1.8.0-bin

export PATH=$PATH:$FLUME_HOME/bin:$PIG_INSTALL/bin:$HIVE_HOME/bin:$HBASE_HOME/bin:$SQOOP_HOME/bin:$HADOOP_HOME/sbin:/$OOZIE_HOME/bin:/$FL
UME_HOME/bin
```

After setting the path of flume directory, save and close the .bashrc file. And then in the terminal type the below command to update the .bashrc file.

```
[acadgild@localhost ~]$ source .bashrc
```

Step 10:

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

1. twitter4j-core-X.XX.jar
2. twitter4j-stream-X.X.X.jar
3. twitter4j-media-support-X.X.X.jar

```
[acadgild@localhost apache-flume-1.8.0-bin]$ ls lib | grep twitter
flume-twitter-source-1.8.0.jar
twitter4j-core-3.0.3.jar
twitter4j-media-support-3.0.3.jar
twitter4j-stream-3.0.3.jar
```

Step 11:

Create a new file inside the conf directory inside the Flume-extracted directory.
Copy the Flumee configuration code from the below link and paste it in the newly created file.

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

```
[acadgild@localhost apache-flume-1.8.0-bin]$ cd conf/
[acadgild@localhost conf]$ pwd
/home/acadgild/apache-flume-1.8.0-bin/conf
[acadgild@localhost conf]$ ls
flume-conf.properties.template  flume-env.ps1.template  flume-env.sh.template  flume_twitter.conf  log4j.properties
[acadgild@localhost conf]$ cat 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=09ieFYOZ3gltT800cxhkAmcHN
TwitterAgent.sources.Twitter.consumerSecret=uy5NKjsyP19d2gujFitZfdTit25MT59NKTizkwBiFm9vUNPoJ5
TwitterAgent.sources.Twitter.accessToken=875303803464974337-ViWylha876bX1k2IhIqzRLF0KsfkMBb
TwitterAgent.sources.Twitter.accessTokenSecret=XDkzifEAcodZNxNY2URvyqLNkkH8Iql00fECjyMPpqB9Q
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[acadgild@localhost conf]$
```

Step 12:
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.

Step 13:

Open a new terminal and start all the Hadoop daemons, before running the flume command to fetch the twitter data.

Use the 'jps' command to see the running Hadoop daemons.

```
[acadgild@localhost ~]$ jps
3088 NameNode
3526 ResourceManager
3192 DataNode
3627 NodeManager
3388 SecondaryNameNode
4764 Jps
```

Step 14:

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

hadoop fs -mkdir -p /user/flume/tweets

```
[acadgild@localhost ~]$ hadoop fs -mkdir -p /user/flume/tweets
17/12/05 07:31:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
17/12/05 07:31:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
```

Step 15:

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>

flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.8.0-bin/conf/flume_twitter.conf

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

```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.8.0-bin/conf/flume_twitter.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop) for HDFS access
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/usr/local/hbase/bin/hbase) for HBASE access
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
17/12/05 07:33:56 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
17/12/05 07:33:56 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
17/12/05 07:33:56 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { source:org.apache
.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkP
rocessor@7fbe26af counterGroup:{ name:null counters:{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name: MemChannel}
} }
17/12/05 07:33:56 INFO node.Application: Starting Channel MemChannel
17/12/05 07:33:56 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully r
egistered new MBean.
17/12/05 07:33:56 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
17/12/05 07:33:56 INFO node.Application: Starting Sink HDFS
17/12/05 07:33:56 INFO node.Application: Starting Source Twitter
17/12/05 07:33:56 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:ID
LE} ...
17/12/05 07:33:56 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered
 new MBean.
17/12/05 07:33:56 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
17/12/05 07:33:56 INFO twitter.TwitterSource: Twitter source Twitter started.
17/12/05 07:33:56 INFO twitter4j.TwitterStreamImpl: Establishing connection.
17/12/05 07:34:14 INFO twitter4j.TwitterStreamImpl: Connection established.
17/12/05 07:34:14 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
17/12/05 07:34:18 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
17/12/05 07:34:19 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/flume/tweets/FlumeData.1512439458728.tmp
17/12/05 07:34:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
17/12/05 07:34:20 INFO twitter.TwitterSource: Processed 100 docs
17/12/05 07:34:28 INFO twitter.TwitterSource: Processed 200 docs
17/12/05 07:34:39 INFO twitter.TwitterSource: Processed 300 docs
17/12/05 07:34:48 INFO twitter.TwitterSource: Processed 400 docs
17/12/05 07:34:50 INFO twitter.TwitterSource: Processed 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.

Step 16:

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

hadoop dfs –ls /user/flume/tweets

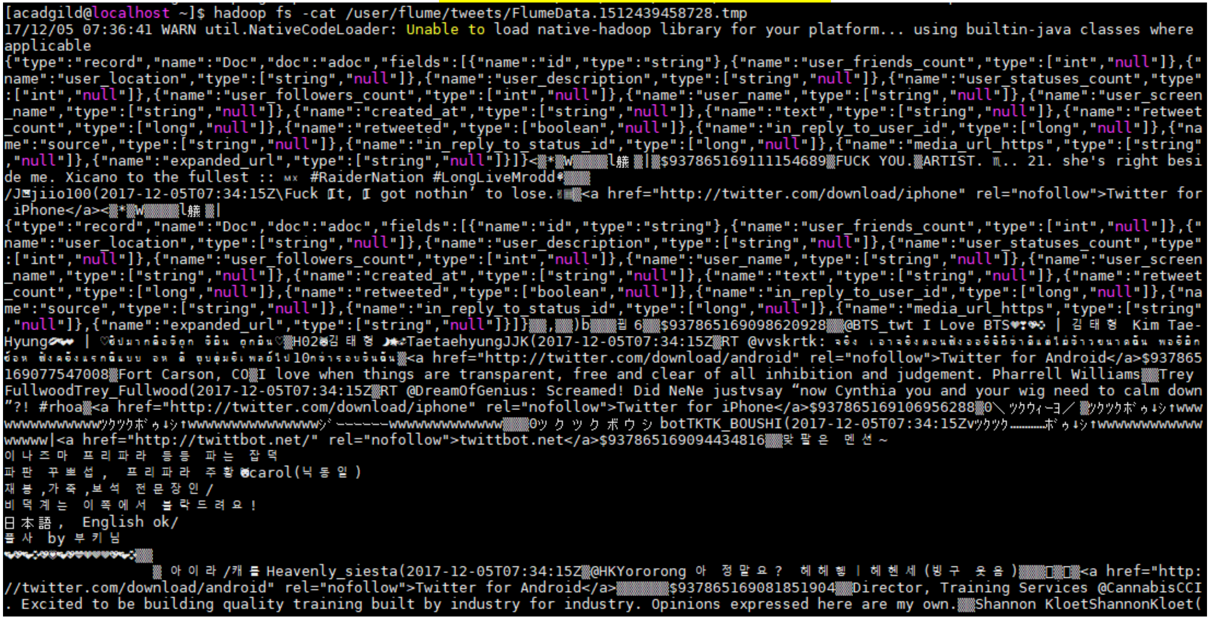
```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
17/12/05 07:36:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup     235283 2017-12-05 07:34 /user/flume/tweets/FlumeData.1512439458728.tmp
```

Step 17:

We can use the 'cat' command to display the tweet data inside the /user/flume/tweets/ path.

hadoop dfs –cat /us er/flume/tweets/<flumeData file name>

hadoop fs -cat /user/flume/tweets/FlumeData.1512439458728.tmp



```
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.1512439458728.tmp
17/12/05 07:36:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int","null"]},{"
name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"user_statuses_count","type"
:["int","null"]},{"name":"user_followers_count","type":["int","null"]},{"name":"user_name","type":["string","null"]},{"name":"user_screen
_name","type":["string","null"]},{"name":"created_at","type":["string","null"]},{"name":"text","type":["string","null"]},{"name":"retweet
_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]},{"name":"in_reply_to_user_id","type":["long","null"]},{"na
me":"source","type":["string","null"]},{"name":"in_reply_to_status_id","type":["long","null"]},{"name":"media_url_https","type":["string"
,"null"]},{"name":"expanded_url","type":["string","null"]}]}<▒*▒W▒▒▒l▒ ▒|▒$937865169111154689▒FUCK YOU.▒ARTIST. ⅲ.. 21. she's right besi
de me. Xicano to the fullest :: ᴹˣ #RaiderNation #LongLiveMrodd▒▒▒
/J▒jiio100(2017-12-05T07:34:15Z\Fuck It, I got nothin’ to lose.▒▒<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
 iPhone</a><▒*▒W▒▒▒l▒ ▒|
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int","null"]},{"
name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"user_statuses_count","type"
:["int","null"]},{"name":"user_followers_count","type":["int","null"]},{"name":"user_name","type":["string","null"]},{"name":"user_screen
_name","type":["string","null"]},{"name":"created_at","type":["string","null"]},{"name":"text","type":["string","null"]},{"name":"retweet
_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]},{"name":"in_reply_to_user_id","type":["long","null"]},{"na
me":"source","type":["string","null"]},{"name":"in_reply_to_status_id","type":["long","null"]},{"name":"media_url_https","type":["string"
,"null"]},{"name":"expanded_url","type":["string","null"]}]}▒,▒)b▒필 6▒$937865169098620928▒@BTS_twt I Love BTS♥︎♡ | 김 태 형  Kim Tae-
Hyung | ♡[illegible]♡▒H02▒김 태 형 ★TaetaehyungJJK(2017-12-05T07:34:15Z▒RT @vvskrtk: [illegible]
[illegible] 10[illegible]▒<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>$937865
169077547008▒Fort Carson, CO▒I love when things are transparent, free and clear of all inhibition and judgement. Pharrell Williams▒Trey
FullwoodTrey_Fullwood(2017-12-05T07:34:15Z▒RT @DreamOfGenius: Screamed! Did NeNe justvsay “now Cynthia you and your wig need to calm down
”?! #rhoa▒<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>$937865169106956288▒0＼ツクウィーヨ／ ▒ツクツクボゥ↓シ↑www
wwwwwwwwwwツクツクボゥ↓シ↑wwwwwwwwwwwwwwwwシ゛ーーーーーーwwwwwwwwwwwwww▒0ツ ク ツ ク ボ ウ シ botTKTK_BOUSHI(2017-12-05T07:34:15Zvツクツク………ボゥ↓シ↑wwwwwwwwwwwww
wwwww|<a href="http://twittbot.net/" rel="nofollow">twittbot.net</a>$937865169094434816▒맞 팔 은  멘 션 ~
이 나 즈 마  프 리 파 라  등 등  파 는  잡 덕
파 판  꾸 뽀 섭 ,  프 리 파 라  주 황 ▒carol(닉 동 일 )
재 봉 ,가 죽 ,보 석  전 문 장 인 /
비 덕 계 는  이 쪽 에 서  블 락 드 려 요 !
日 本 語 ,  English ok/
플 사  by 부 키 님
▒▒▒
                    ▒ 아 이 라 /캐 ▒ Heavenly_siesta(2017-12-05T07:34:15Z▒@HKYororong 아  정 말 요 ?  헤 헤 헹 | 헤 헨 세 (빙 구  웃 음 )▒□▒□▒<a href="http:
//twitter.com/download/android" rel="nofollow">Twitter for Android</a>▒▒▒$937865169081851904▒Director, Training Services @CannabisCCI
. Excited to be building quality training built by industry for industry. Opinions expressed here are my own.▒Shannon KloetShannonKloet(
```

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