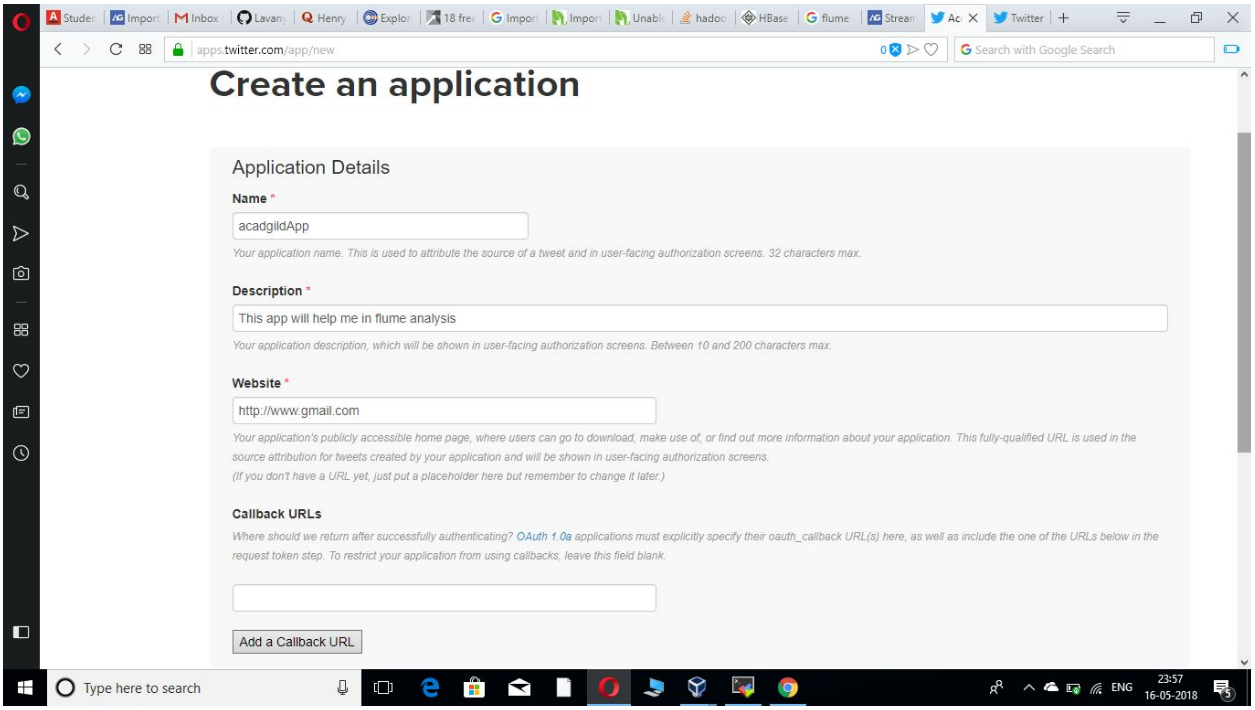
Assignment 12:

Steps to stream data using flume:

1. Go to the following link and click the 'create new app' button. https://apps.twitter.com/app
2. Create an application and provide name,description, website as given in the below snapshot:



Create an application

Application Details

Name *

acadgildApp

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

This app will help me in flume analysis

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

http://www.gmail.com

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

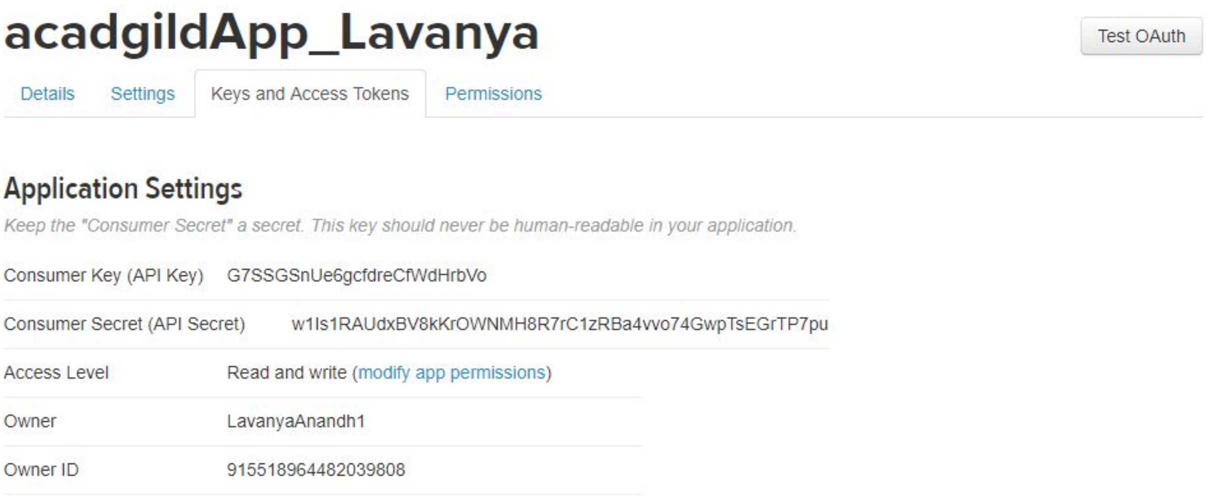
Callback URLs

Where should we return after successfully authenticating? OAuth 1.0a applications must explicitly specify their oauth_callback URL(s) here, as well as include the one of the URLs below in the request token step. To restrict your application from using callbacks, leave this field blank.

Add a Callback URL

3. Accept the developer agreement and select the 'create your Twitter application' button.

4. Select the 'Keys and Access Token' tab as below:



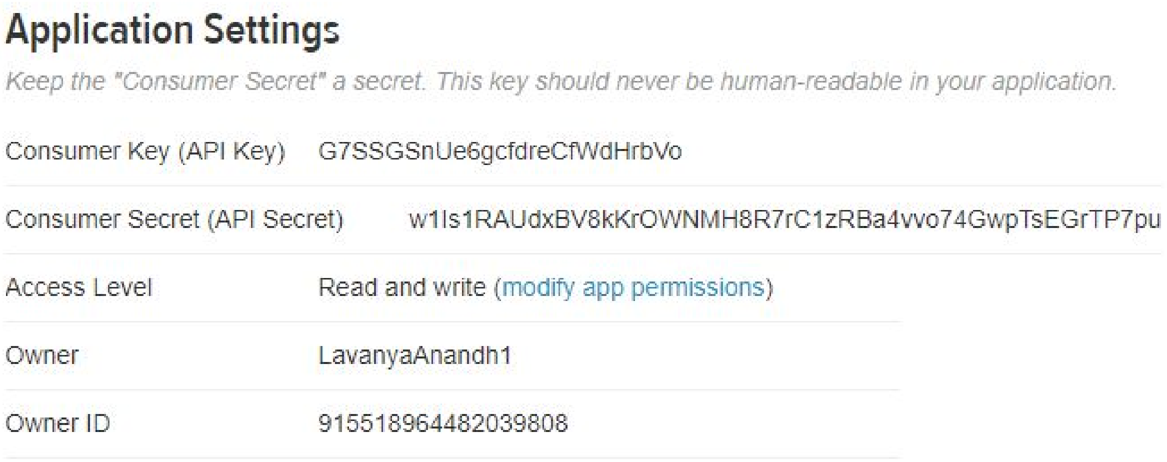
acadgildApp_Lavanya

Application Settings

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

Consumer Key (API Key)	G7SSGSnUe6gcfdreCfWdHrbVo
Consumer Secret (API Secret)	w1Is1RAUdxBV8kKrOWNMH8R7rC1zRBa4vvo74GwpTsEGrTP7pu
Access Level	Read and write (modify app permissions)
Owner	LavanyaAnandh1
Owner ID	915518964482039808

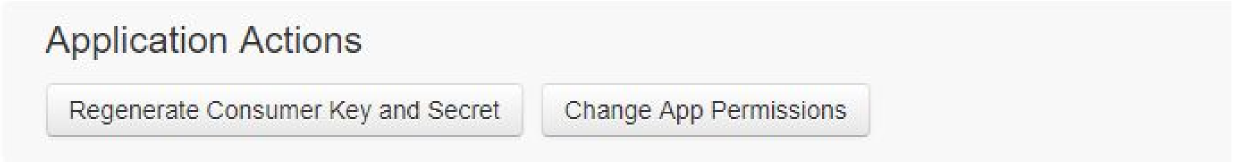
5. Copy the consumer key and the consumer secret code as below:
Consumer Key (API Key)G7SSGSnUe6gcfdreCfWdHrbVo
Consumer Secret (API Secret)w1Is1RAUdxBV8kKrOWNMH8R7rC1zRBa4vvo74GwpTsEGrTP7pu



Application Settings

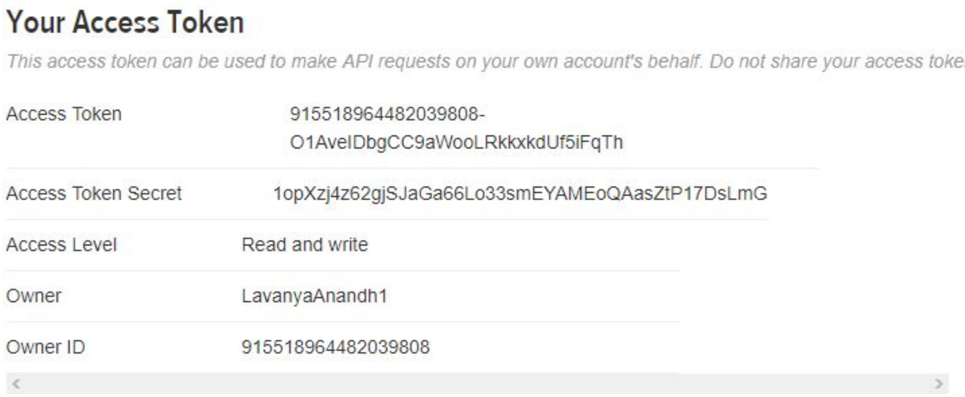
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	G7SSGSnUe6gcfdreCfWdHrbVo
Consumer Secret (API Secret)	w1Is1RAUdxBV8kKrOWNMH8R7rC1zRBa4vvo74GwpTsEGrTP7pu
Access Level	Read and write (modify app permissions)
Owner	LavanyaAnandh1
Owner ID	915518964482039808



Application Actions

6. Scroll down further and select the ‘create my access token’ button
7. you will receive a message stating that you have successfully generated your application access token.
8. Copy the Access Token and Access token Secret code as below:
Access Token915518964482039808-O1AvelDbgCC9aWooLRkkxkdUf5iFqTh
Access Token Secret1opXzj4z62gjSJaGa66Lo33smEYAMEoQAasZtP17DsLmG

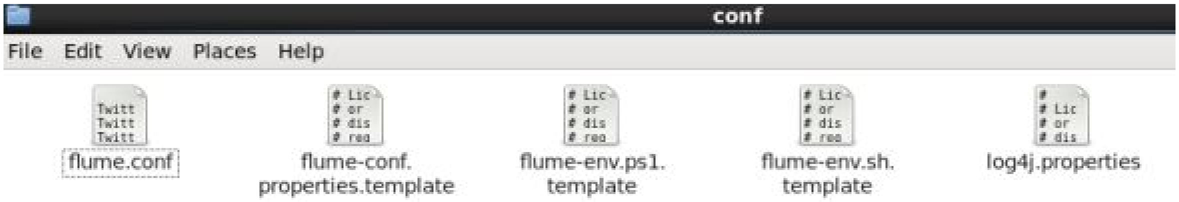


Your Access Token

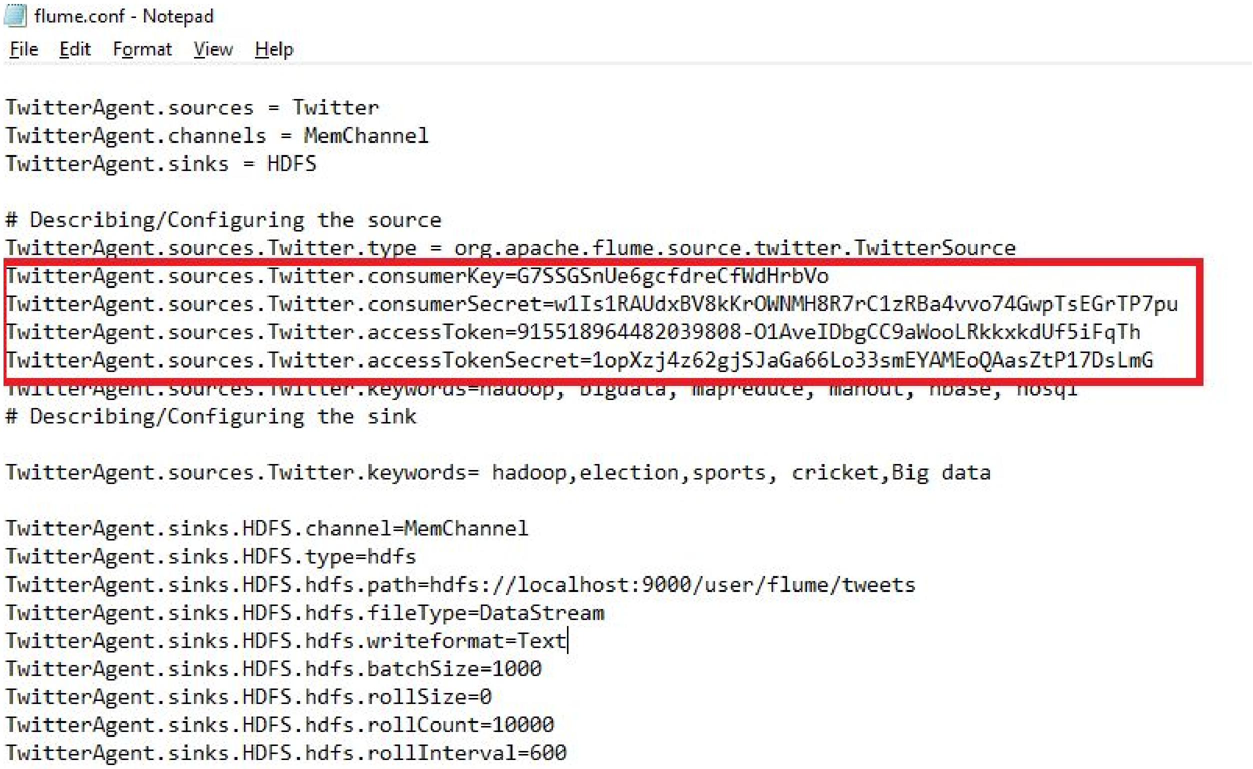
This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	915518964482039808- O1AvelDbgCC9aWooLRkkxkdUf5iFqTh
Access Token Secret	1opXzj4z62gjSJaGa66Lo33smEYAMEoQAasZtP17DsLmG
Access Level	Read and write
Owner	LavanyaAnandh1
Owner ID	915518964482039808

9. Create a new file inside the conf directory under **flume(/home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/flume.conf)**.



10. Update the file with below contents and copy the saved consumerKey, consumerSecret, accessToken, accessTokenSecret as highlighted in red below.



```
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=G7SSGSnUe6gcfdreCfWdHrbVo
TwitterAgent.sources.Twitter.consumerSecret=w1Is1RAUdxBV8kKrOWNMH8R7rC1zRBa4vvo74GwpTsEGrTP7pu
TwitterAgent.sources.Twitter.accessToken=915518964482039808-O1AveIDbgCC9aWooLRkkxkdUf5iFqTh
TwitterAgent.sources.Twitter.accessTokenSecret=1opXzj4z62gjSJaGa66Lo33smEYAMEoQAasZtP17DsLmG
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
```

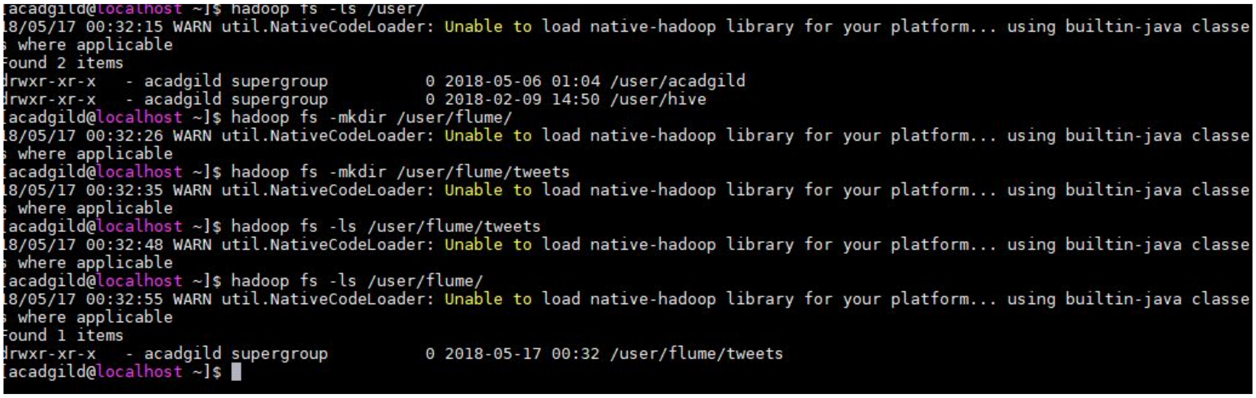
11. Make sure you have below jars placed in your $FLUME_HOME/lib directory:

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

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.

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.

14. Create a new directory inside HDFS path, where the Twitter tweet data should be stored as below. **Hadoop dfs –mkdir –p /user/flume/tweets**

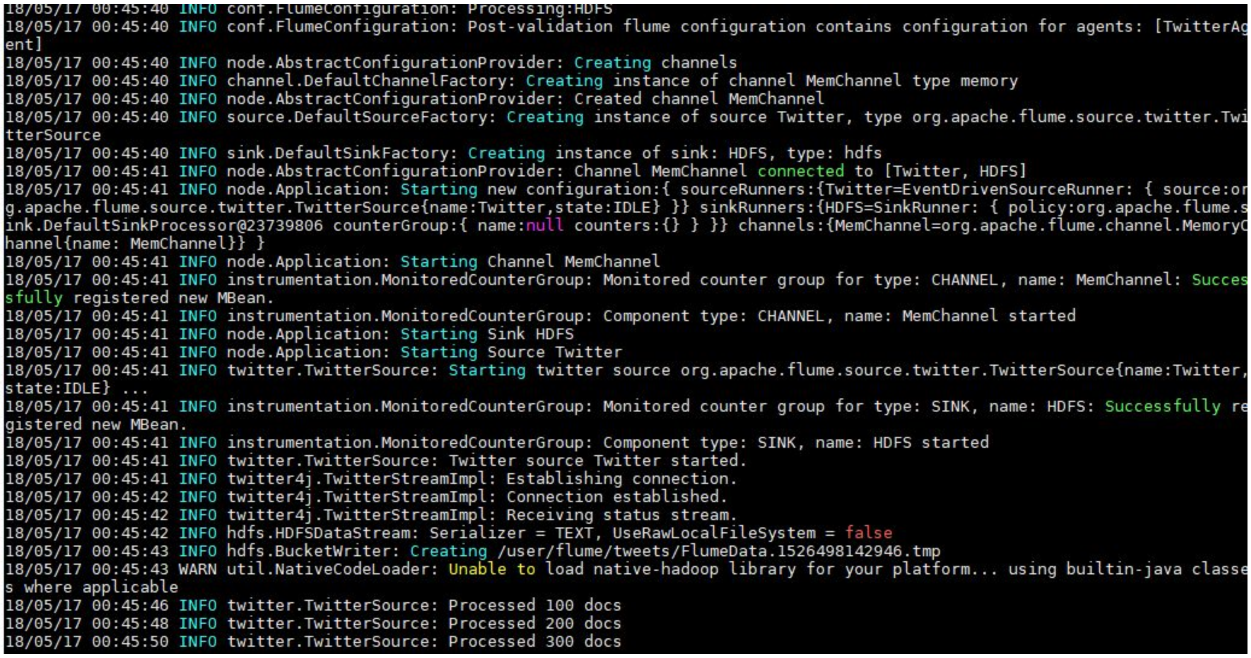


```
[acadgild@localhost ~]$ hadoop fs -ls /user/
18/05/17 00:32:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 2 items
drwxr-xr-x   - acadgild supergroup          0 2018-05-06 01:04 /user/acadgild
drwxr-xr-x   - acadgild supergroup          0 2018-02-09 14:50 /user/hive
[acadgild@localhost ~]$ hadoop fs -mkdir /user/flume/
18/05/17 00:32:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
[acadgild@localhost ~]$ hadoop fs -mkdir /user/flume/tweets
18/05/17 00:32:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
18/05/17 00:32:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/
18/05/17 00:32:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 1 items
drwxr-xr-x   - acadgild supergroup          0 2018-05-17 00:32 /user/flume/tweets
[acadgild@localhost ~]$ 
```

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/install/flume/apache-flume-1.8.0-bin/conf/flume.conf

16. The below snapshot shows the processing of records that are getting streamed.
17. To stop the streaming press ctrl+c so that streaming is stopped.

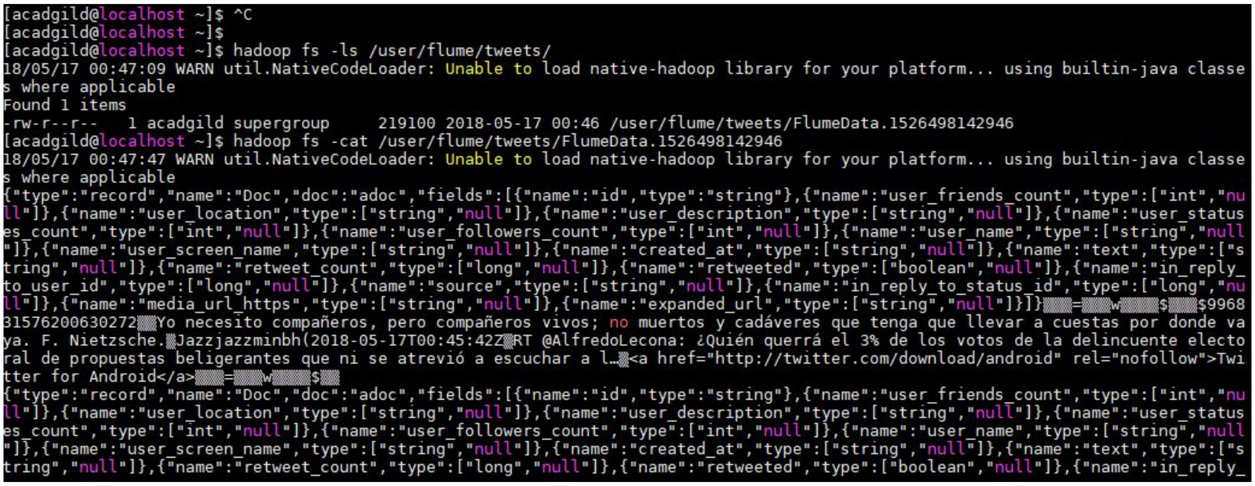


```
18/05/17 00:45:40 INFO conf.FlumeConfiguration: Processing:HDFS
18/05/17 00:45:40 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAg
ent]
18/05/17 00:45:40 INFO node.AbstractConfigurationProvider: Creating channels
18/05/17 00:45:40 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
18/05/17 00:45:40 INFO node.AbstractConfigurationProvider: Created channel MemChannel
18/05/17 00:45:40 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.Twi
tterSource
18/05/17 00:45:40 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
18/05/17 00:45:41 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
18/05/17 00:45:41 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { source:or
g.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache.flume.s
ink.DefaultSinkProcessor@23739806 counterGroup:{ name:null counters:{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryC
hannel{name: MemChannel}} }
18/05/17 00:45:41 INFO node.Application: Starting Channel MemChannel
18/05/17 00:45:41 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Succes
sfully registered new MBean.
18/05/17 00:45:41 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
18/05/17 00:45:41 INFO node.Application: Starting Sink HDFS
18/05/17 00:45:41 INFO node.Application: Starting Source Twitter
18/05/17 00:45:41 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,
state:IDLE} ...
18/05/17 00:45:41 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully re
gistered new MBean.
18/05/17 00:45:41 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/05/17 00:45:41 INFO twitter.TwitterSource: Twitter source Twitter started.
18/05/17 00:45:41 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/05/17 00:45:42 INFO twitter4j.TwitterStreamImpl: Connection established.
18/05/17 00:45:42 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/05/17 00:45:42 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/05/17 00:45:43 INFO hdfs.BucketWriter: Creating /user/flume/tweets/FlumeData.1526498142946.tmp
18/05/17 00:45:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
18/05/17 00:45:46 INFO twitter.TwitterSource: Processed 100 docs
18/05/17 00:45:48 INFO twitter.TwitterSource: Processed 200 docs
18/05/17 00:45:50 INFO twitter.TwitterSource: Processed 300 docs
```

18. We can use the *‘cat’* command to display the tweet data inside the /user/flume/tweets/ path.

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

hadoop fs -cat /user/flume/tweets/FlumeData.1526498142946



```
[acadgild@localhost ~]$ ^C
[acadgild@localhost ~]$
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets/
18/05/17 00:47:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup     219100 2018-05-17 00:46 /user/flume/tweets/FlumeData.1526498142946
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.1526498142946
18/05/17 00:47:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classe
s where applicable
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int","nu
ll"]},{"name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"user_status
es_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":["s
tring","null"]},{"name":"retweet_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]},{"name":"in_reply_
to_user_id","type":["long","null"]},{"name":"source","type":["string","null"]},{"name":"in_reply_to_status_id","type":["long","nu
ll"]},{"name":"media_url_https","type":["string","null"]},{"name":"expanded_url","type":["string","null"]}]}▒▒=▒▒w▒▒▒$▒▒$9968
31576200630272▒Yo necesito compañeros, pero compañeros vivos; no muertos y cadáveres que tenga que llevar a cuestas por donde va
ya. F. Nietzsche.▒Jazzjazzminbh(2018-05-17T00:45:42Z▒RT @AlfredoLecona: ¿Quién querrá el 3% de los votos de la delincuente electo
ral de propuestas beligerantes que ni se atrevió a escuchar a l…▒<a href="http://twitter.com/download/android" rel="nofollow">Twi
tter for Android</a>▒▒=▒▒w▒▒▒$▒▒
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int","nu
ll"]},{"name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"user_status
es_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":["s
tring","null"]},{"name":"retweet_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]},{"name":"in_reply_
```