Session 12 - Oozie and Flume

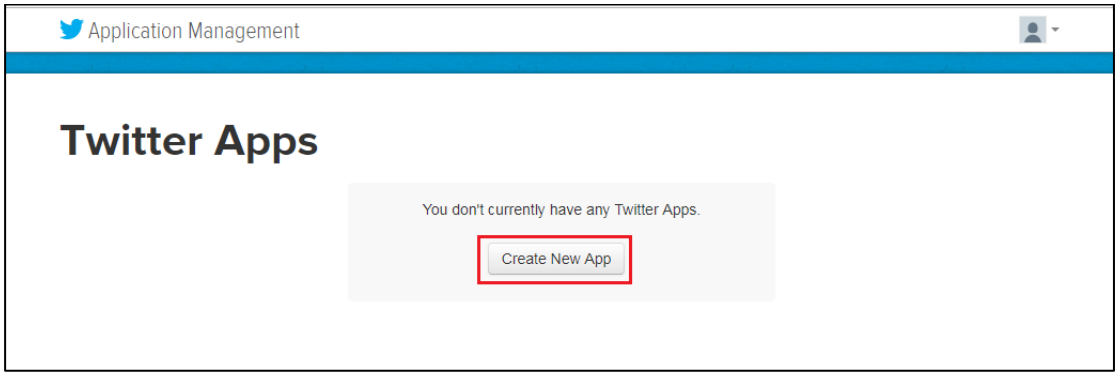
Assignment 1

Task 1:

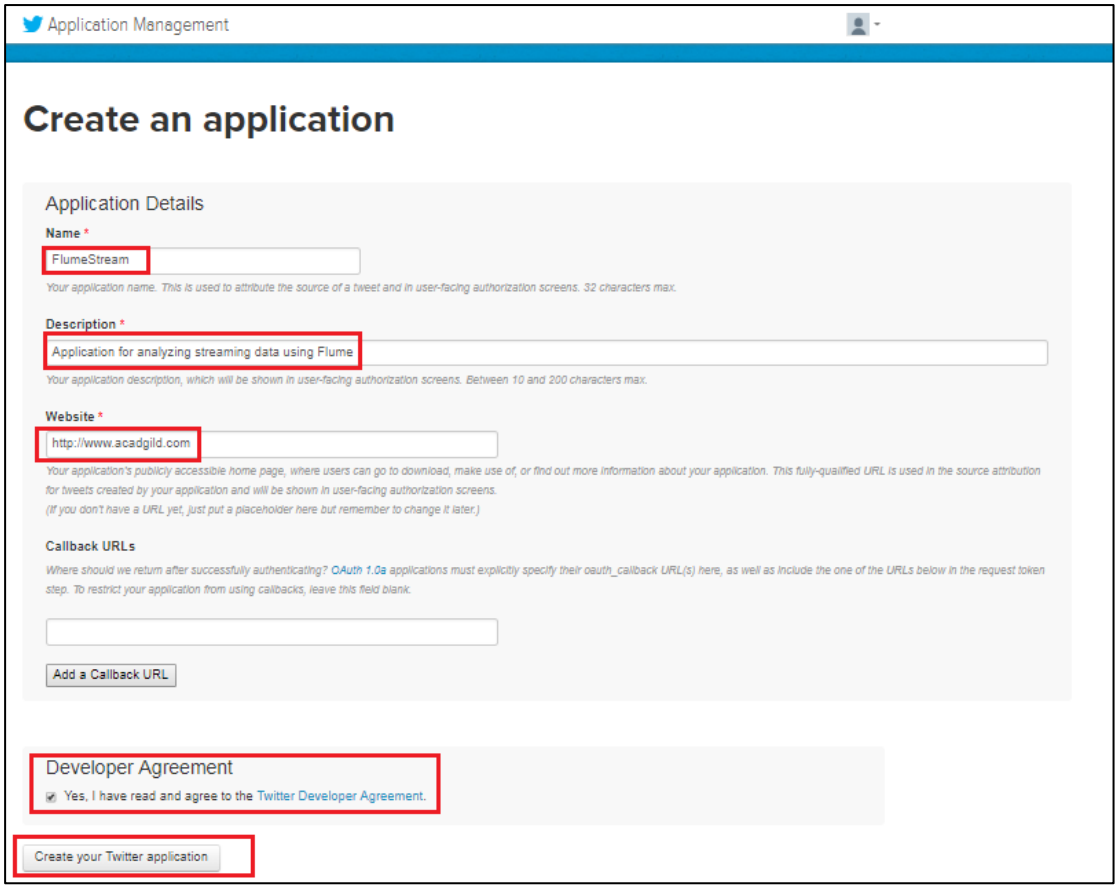
Create a flume agent that streams data from Twitter and stores in the HDFS.

PART 1 : Getting access keys from Twitter

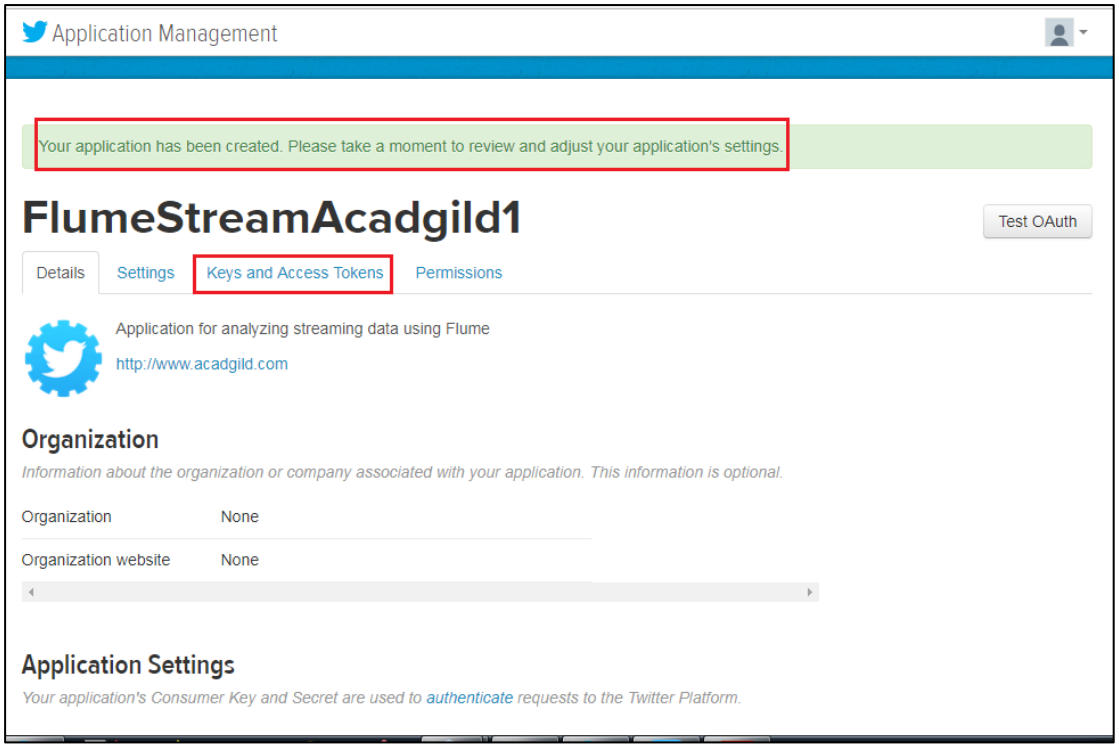
Step 1 : Login to Twitter account and then go to '*https://apps.twitter.com/app*' to go to the Twitter developer site and click on 'Create New App'.



Step 2 : Then enter all the necessary details, click on the 'I Agree' button for Developer Agreement and then continue by clicking in the 'Create Twitter Application' button.



Step 3 : Once the application is created, a confirmation message is sent. Then go to the 'Keys and Access Tokens' tab of the application.



Step 4 : From the 'Keys and Access Tokens' tab, copy the Consumer Key (API Key), Consumer Secret (API Secret) values and save them.

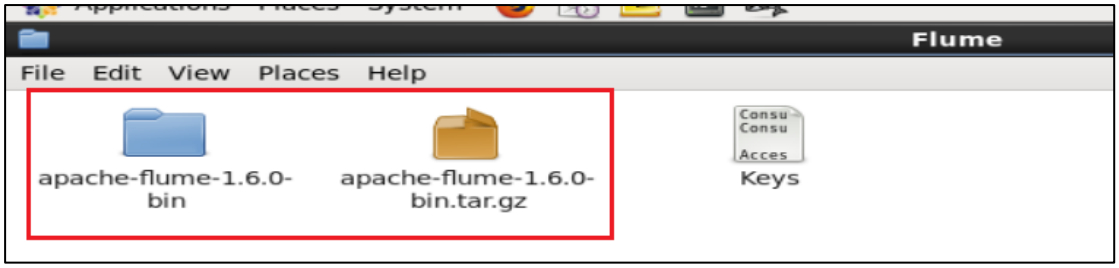
Then, click on the 'Generate Consumer Key and Secret' button below.Then, copy and save these values.

These values will be used to programmatically access Twitter from the Flume agent.

These values can be used to access the twitter account and do any activities from the account. So do not use these values publicly.

PART 2 : Setting Up a Flume Agent for streaming Twitter data

Step 1 : Install Flume by downloading the flume jar file and unzipping it in the installing directory.

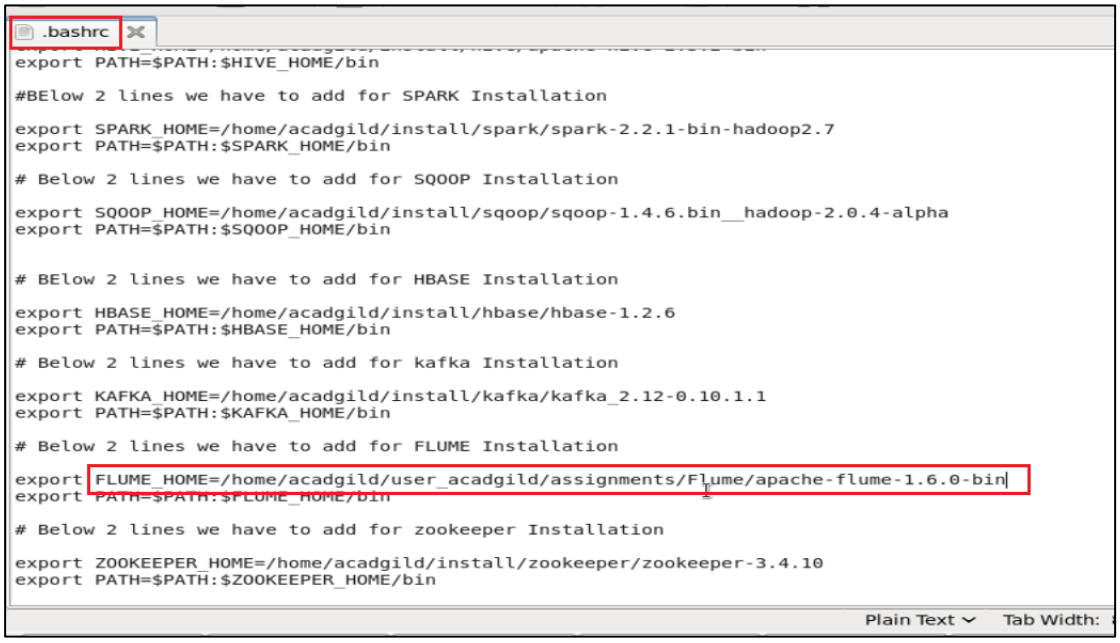


Step 2 : Setting the path of the installed directory in .bashrc

Open .bashrc file : *sudo gedit .bashrc*

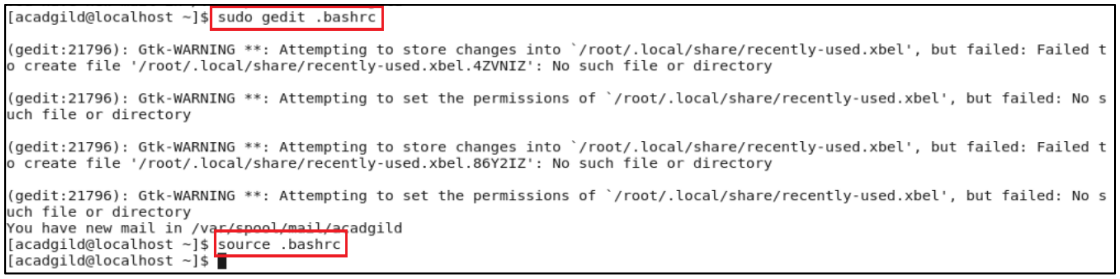
It will ask for the password of the current user.

Once the file is opened, go to the FLUME_HOME and change it's value to the directory that Flume was just installed.



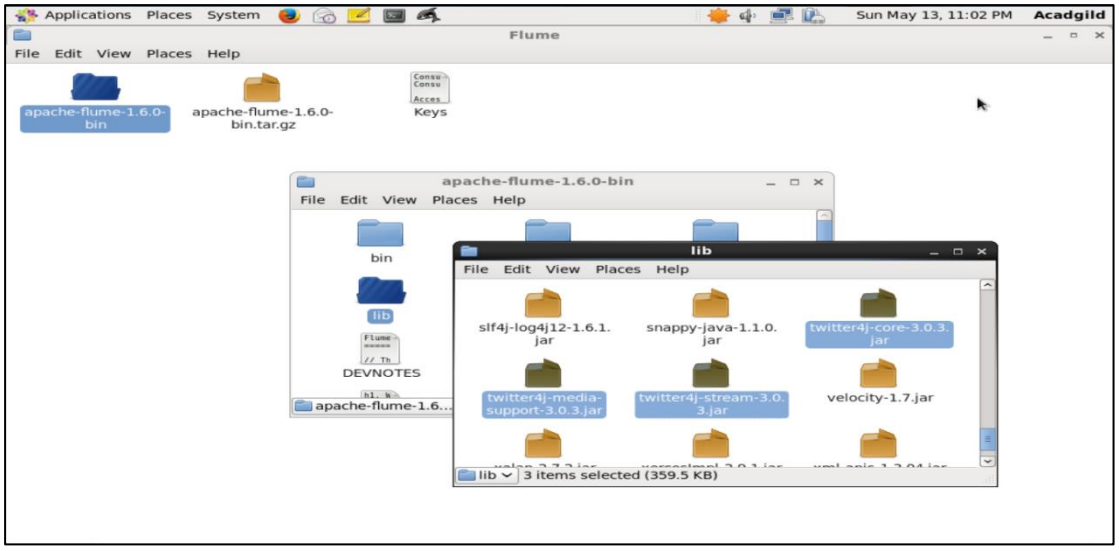
Step 3 : Save the .bashrc file and run the following command to update the .bashrc file.

source .bashrc

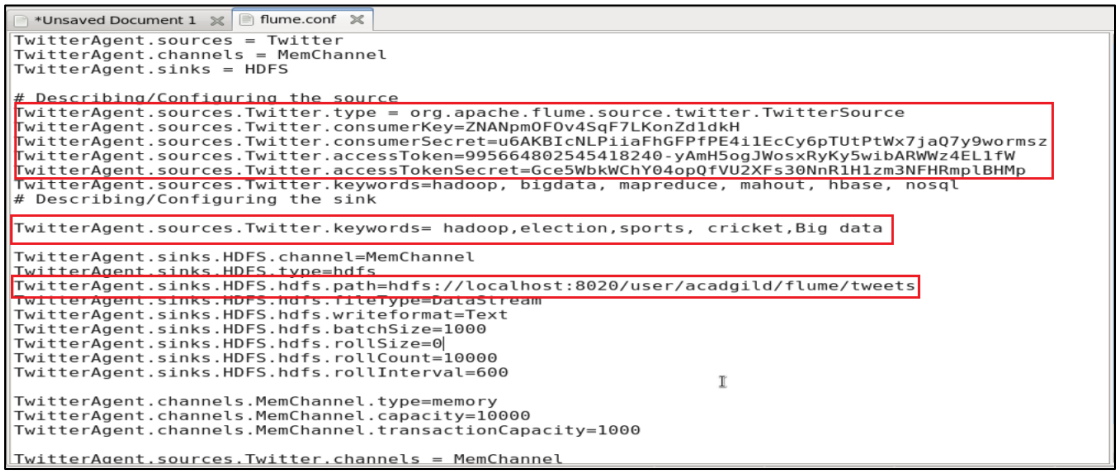


Step 4 : Make sure you have below jars placed in your $FLUME_HOME/lib directory:

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



Step 5 : Create an empty file in the apache-flume-1.6.0-bin/conf folder named 'flume.conf' and copy paste the configuration parameters in the file.



```
*Unsaved Document 1    flume.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=ZNANpmOFOv4SqF7LKonZd1dkH
TwitterAgent.sources.Twitter.consumerSecret=u6AKBIcNLPiiaFhGFPfPE4i1EcCy6pTUtPtWx7jaQ7y9wormsz
TwitterAgent.sources.Twitter.accessToken=995664802545418240-yAmH5ogJWosxRyKy5wibARWWz4EL1fW
TwitterAgent.sources.Twitter.accessTokenSecret=Gce5WbkWChY04opQfVU2XFs30NnR1H1zm3NFHRmplBHMp
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:8020/user/acadgild/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
```

Replace the existing values of consumer Key, Consumer Secret, Access Token, Access Token Secret with the values obtained from twitter w=in the earlier part.

'TwitterAgent.sources.Twitter.keywords' is the command to set the list of keywords that are to be searched for in the tweets.

‘TwitterAgent.sinks.HDFS.hdfs.path’ : set the HDFS path where the streaming feed will be stored on HDFS. Give the port (8020 or 9000) that is specified in the ‘core-site.xml’ of your hadoop installation.

Step 6 : Make sure the hadoop components are up by running a jps command. If not up, run start-all.sh to get them up and running.

Also, create the target HDFS folders specified in the flume.conf file.

```
[acadgild@localhost ~]$ jps
3104 NameNode
3380 SecondaryNameNode
3205 DataNode
26197 HQuorumPeer
3542 ResourceManager
22743 Jps
5258 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
3645 NodeManager
[acadgild@localhost ~]$ hadoop dfs -mkdir /user/acadgild/flume
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

18/05/13 23:22:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost ~]$ hadoop dfs -mkdir /user/acadgild/flume/tweets
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

18/05/13 23:22:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost ~]$
```

Step 7 : Run the flume agent to get the twitter feed stream.

flume-ng agent -n TwitterAgent -f <location of created/edited conf file>

flume-ng agent -n TwitterAgent -f /home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin/conf/flume.conf

```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-b
in/conf/flume.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop) for HDFS access
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.2.6/bin/hbase) for HBASE access
Info: Excluding /home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-api-1.7.7.jar from classpath
Info: Excluding /home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/home/acadgild/install/hive/apache-hive-2.3.2-bin) for Hive access
+ exec /usr/java/jdk1.8.0_151/bin/java -Xmx20m -cp '/home/acadgild/user_acadgild/assignments/Flume/apache-flume-1.6.0-bin/lib
/*:/home/acadgild/install/hadoop/hadoop-2.6.5/etc/hadoop/:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/
activation-1.1.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadg
ild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/install/hadoop/h
```

The connection with Twitter is established and the tweet stream is read and written to HDFS.

```
18/05/14 00:31:31 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfull
y registered new MBean.
18/05/14 00:31:31 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/05/14 00:31:31 INFO twitter.TwitterSource: Twitter source Twitter started.
18/05/14 00:31:31 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/05/14 00:31:34 INFO twitter4j.TwitterStreamImpl: Connection established.
18/05/14 00:31:34 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/05/14 00:31:35 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/05/14 00:31:36 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/acadgild/flume/tweets/FlumeData.1526238095572.t
mp
18/05/14 00:31:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/05/14 00:31:37 INFO twitter.TwitterSource: Processed 100 docs
18/05/14 00:31:40 INFO twitter.TwitterSource: Processed 200 docs
18/05/14 00:31:42 INFO twitter.TwitterSource: Processed 300 docs
18/05/14 00:31:45 INFO twitter.TwitterSource: Processed 400 docs
18/05/14 00:31:48 INFO twitter.TwitterSource: Processed 500 docs
18/05/14 00:31:50 INFO twitter.TwitterSource: Processed 600 docs
18/05/14 00:31:53 INFO twitter.TwitterSource: Processed 700 docs
^C18/05/14 00:31:53 INFO lifecycle.LifecycleSupervisor: Stopping lifecycle supervisor 10
```

The stream will be continuously read. To stop the process press ctrl+c.

Step 8 : Then, the stream will be read into the HDFS location specified.

hadoop dfs –ls /user/acadgild/flume/tweets

To see the contents of the file, do a cat on the file

hadoop dfs -cat /user/acadgild/flume/tweets/FlumeData.1526237236072

```
File  Edit  View  Search  Terminal  Help
a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>995737437488975874chennai, Ind
ia@Proud Indian, Tax Payer , Die Hard Fan of Modi Sarkar, Believes In Digitial Economy.@Arun KumarArun_KumaR2712(2018-05
-14T00:17:59ZRT @amitmalviya: Blaming the RSS for 26/11 was a well crafted strategy of the Congress party to create a spect
re of Hindu terror with an ey…a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>995
737437629056O
夜无语eIa67ZkemeDFRwI(2018-05-14T00:17:59Z@スマホRPGは今これをやってるよ。今はこのイベントが開催中！
→ https://t.co/zPvL7B4BL8 https://t.co/fkEHjGfuJka href="http://granbluefantasy.jp/" rel="nofollow">グランブルー ファ
ンタジー</a>https://pbs.twimg.com/media/DdGR6_9VQAAIdKT.jpghttps://twitter.com/eIa67ZkemeDFRwI/status/995737437476290560/
photo/1$995737437493190656Bangtan's wife, joking around.Iae biased or Yoongi biased? Seven amazing dorks (@BTS_twt).Bi
tch, the Queen is here. Fan Account.h2r|| tear. ||[ s/h ]Bangtansbae7(2018-05-14T00:17:59ZRMay 13th.
8.47pm. https://t.co/sJ7R1qHPoS a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
https://pbs.twimg.com/media/DdGR6gCW0AAOWq2.jpghttps://twitter.com/bangtansbae7/status/995737437493190656/photo/1$9957374
37497372672amet Özenametze74729906(2018-05-14T00:17:59Zh" TürkGençliği MHPdiyor " https://t.co/dXjdqtsUHs a hr
ef="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>https://pbs.twimg.com/media/DdGR6okXcAA9y5p.
jpghttps://twitter.com/Sametze74729906/status/995737437497372672/photo/1$995737441666387970eepali Rawatawa
t_deepali(2018-05-14T00:18:00ZRT @VORdotcom: @ pallavighosh @TwitterIndia I am not a troll... I am angry like so many other
s like me who have lost their employment, had…a href="https://mobile.twitter.com" rel="nofollow">Twitter Lite</a>995
737441687388160RomalNotizie dal mondo dei Giochi di Ruolo online italiani!OR-online.comdronline(2018-05-14T00:18:00
ZRecensione Bordertown di daethel: Recensione Bordertown di daethel https://t.co/cRRQ7rlpVW<a href="https://dlvrit.com/"
rel="nofollow">dlvr.it</a>995737441674801153Nitin Middhanitin_middha(2018-05-14T00:18:00ZRT @saini_guljhari: @N
rhmSangh @VasundharaBJP @BJP4India Manniya CM Sahiba se karbadh nivedan hi NHM management karmiko rqulzied in medical…a
href="https://mobile.twitter.com" rel="nofollow">Twitter Lite</a>995737441699938304Noida, IndiaNews State covers bre
aking news on politics, world, business, sports and cinema. Follow us for latest updates from India.Zews StateNewsStat
eHindi(2018-05-14T00:18:00Z
@manoharparrikar

https://t.co/CyrfkvoOwy a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>995737441678970880日本 熊
本Phone用アカウントです。熊本担当で、未央・響Pです。猫派です。

.ビデ(モバイル版)Vbz8cFo(2018-05-14T00:18:00ZRT @imas_columbia: Serendipity NIGHTありがとうございました！そしてニコ生でも発
表の通りSSA公演のBlu-rayが8/29に発売決定！明日から予約スタートです。各公演のBlu-rayも引き続きご予約受付中！発売をお楽しみに！
```

The contents of the stream of tweets containing the keywords specified in the flume.conf file can be seen in the output.