

11.3 Problem Statement

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

STEP 1:

Synchronize system clock and acadgild vm clock for correct system time.

Download link for apache flume:

Command:

```
[acadgild@quickstart ~]$ wget http://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz
```

STEP 2:

Extract file from flume tar file.

Command:

```
[acadgild@quickstart ~]$ tar -xvf apache-flume-1.6.0-bin.tar.gz
```

STEP 3:

apache-flume-1.6.0-bin will be saved in downloads directory

Command:

```
[acadgild@quickstart ~]$ ls /home/acadgild/Downloads/apache-flume-1.6.0-bin/
```

STEP 4:

Download below three jars for twitter streaming from below mentioned MAVEN repository:

twitter4j-core-3.0.3.jar

twitter4j-stream-3.0.3.jar

twitter4j-media-support-3.0.3.jar

<https://mvnrepository.com/artifact/org.twitter4j/twitter4j-media-support/3.0.3>

<https://mvnrepository.com/artifact/org.twitter4j/twitter4j-core/3.0.3>

<https://mvnrepository.com/artifact/org.twitter4j/twitter4j-stream/3.0.3>

STEP 5:

We need to remove protobuf-java-2.4.1.jar and guava-10.1.1.jar from lib directory of apache-flume-1.6.0-bin (when using hadoop-2.x)

Command:

```
[acadgild@quickstart ~]$ sudo rm /home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/protobuf-  
java-2.5.0.jar /home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/guava-11.0.2.jar
```

STEP 6:

We need to copy 3 twitter jars from Downloads directory to lib directory of apache-flume-1.6.0-bin.

Commands:

```
[acadgild@quickstart ~]$ cp /home/acadgild/Downloads/twitter4j-core-3.0.3.jar  
/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/
```

```
[acadgild@quickstart ~]$ cp /home/acadgild/Downloads/twitter4j-stream-3.0.3.jar  
/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/
```

```
[acadgild@quickstart ~]$ cp /home/acadgild/Downloads/twitter4j-media-support-3.0.3.jar  
/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/
```

STEP 7:

Use below link and download -> flume-sources-1.0-SNAPSHOTS.jar

files.acadgild.com/samples/flume-sources-1.0-SNAPSHOT.jar

STEP 8:

Copy the flume-sources-1.0-SNAPSHOT.jar file from Downloads directory to lib directory of apache flume:

Command:

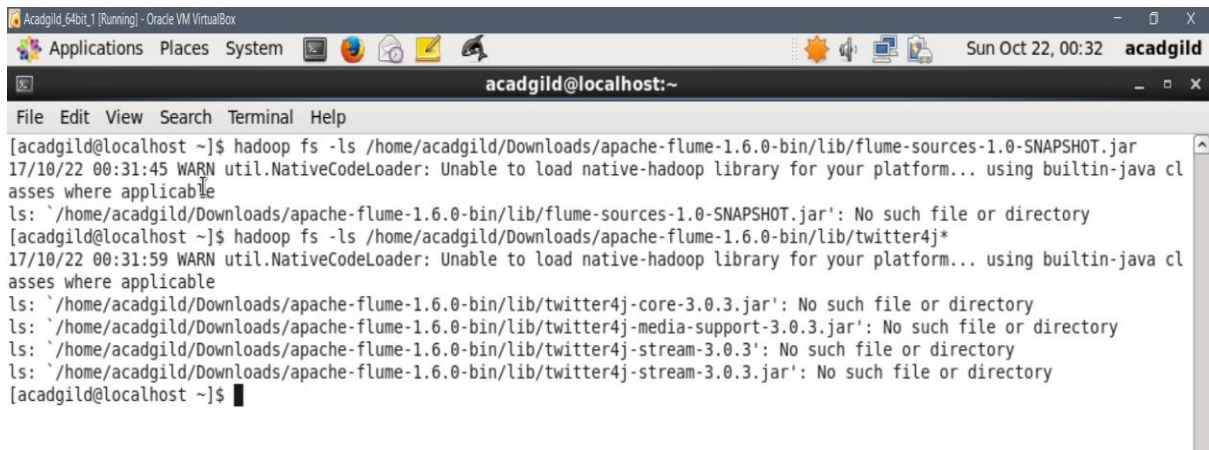
```
[acadgild@quickstart ~]$ cp /home/acadgild/Downloads/flume-sources-1.0-SNAPSHOT.jar  
/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/
```

STEP 9:

Check whether flume SNAPSHOT has moved to the lib folder of apache flume:

Command:

```
[acadgild@quickstart ~]$ ls /home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/flume-sources-  
1.0-SNAPSHOT.jar
```



The screenshot shows a terminal window titled 'Acadgild_64bit_1 [Running] - Oracle VM VirtualBox'. The window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The terminal output shows the user 'acadgild' at 'localhost' running several 'hadoop fs -ls' commands to check for the presence of various JAR files in the directory '/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/'. The files checked are 'flume-sources-1.0-SNAPSHOT.jar', 'twitter4j*', 'twitter4j-core-3.0.3.jar', 'twitter4j-media-support-3.0.3.jar', and 'twitter4j-stream-3.0.3.jar'. All files are reported as 'No such file or directory'. There are also two warning messages from 'util.NativeCodeLoader' about not being able to load the native-hadoop library, suggesting the use of built-in Java classes where applicable. The terminal ends with a prompt '[acadgild@localhost ~]\$'.

STEP 10:

Copy flume-env.sh.template content to flume-env.sh

Commands:

```
[acadgild@quickstart ~]$ cd /home/acadgild/Downloads/apache-flume-1.6.0-bin/conf
```

```
[acadgild@quickstart conf~]$ sudo cp flume-env.sh.template flume-env.sh
```

STEP 11:

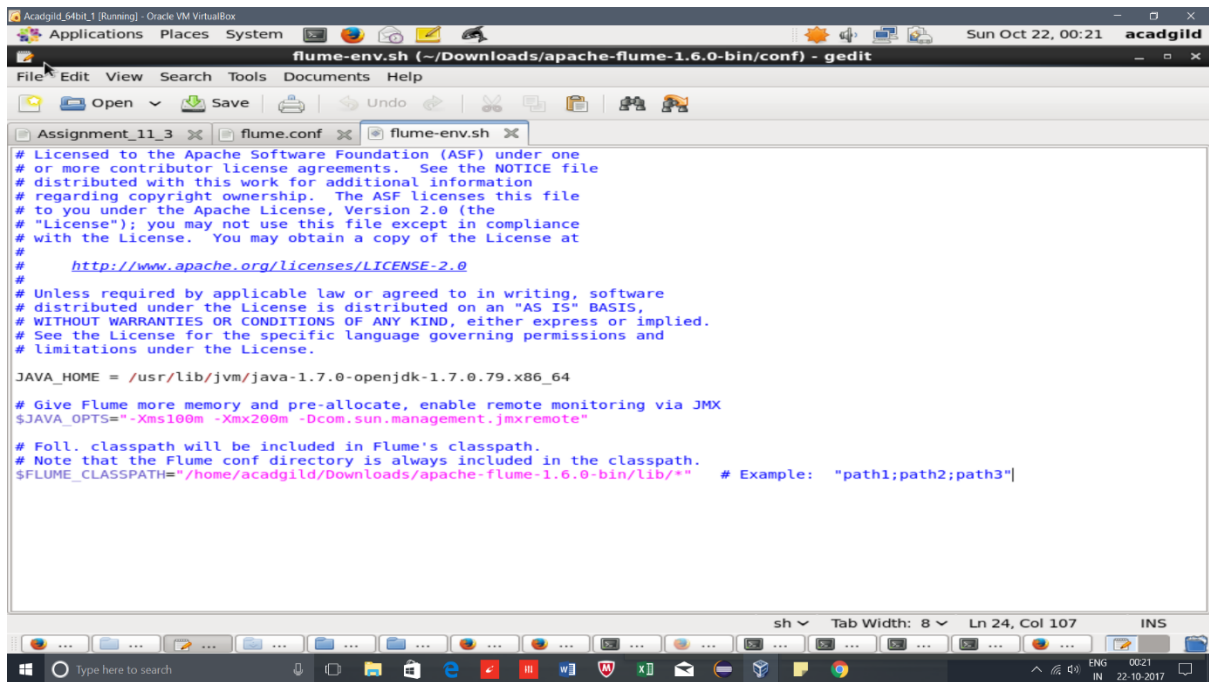
Edit flume-env.sh as mentioned in below snapshot for JAVA_HOME & FLUME_CLASSPATH.

Commands:

```
[acadgild@quickstart conf~]$ sudo gedit flume-env.sh
```

```
#JAVA_HOME=/usr/java/jdk1.7.0_67-acadgild
```

```
#FLUME_CLASSPATH="/home/acadgild/Downloads/apache-flume-1.6.0-bin/lib/*"
```



STEP 12:

Open a Browser and go to the below URL:

URL:https://twitter.com/

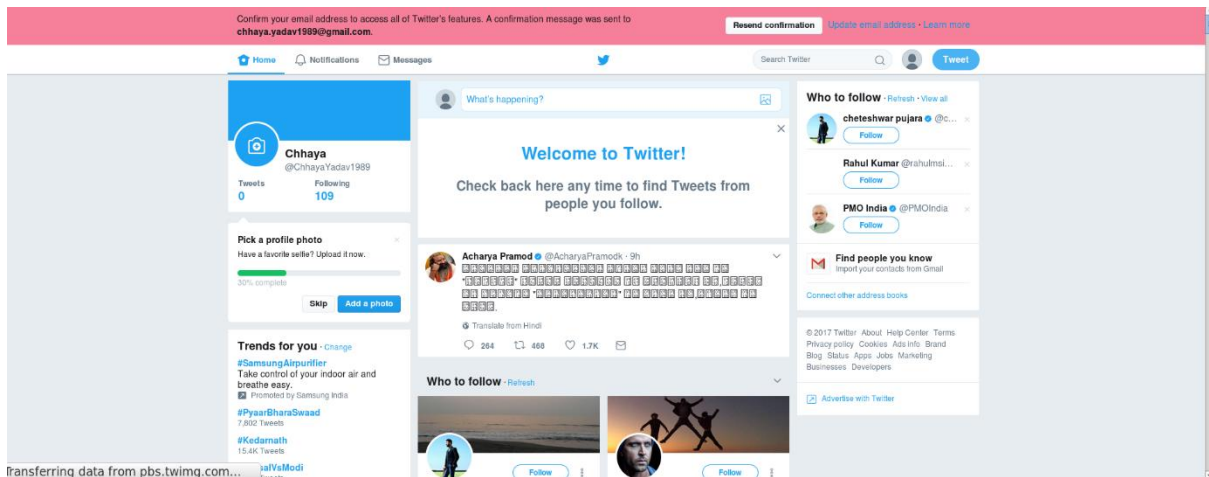
Sign up on Twitter

STEP 13:

Enter your Twitter account credentials and sign in:

STEP 14:

Your twitter home page will open:

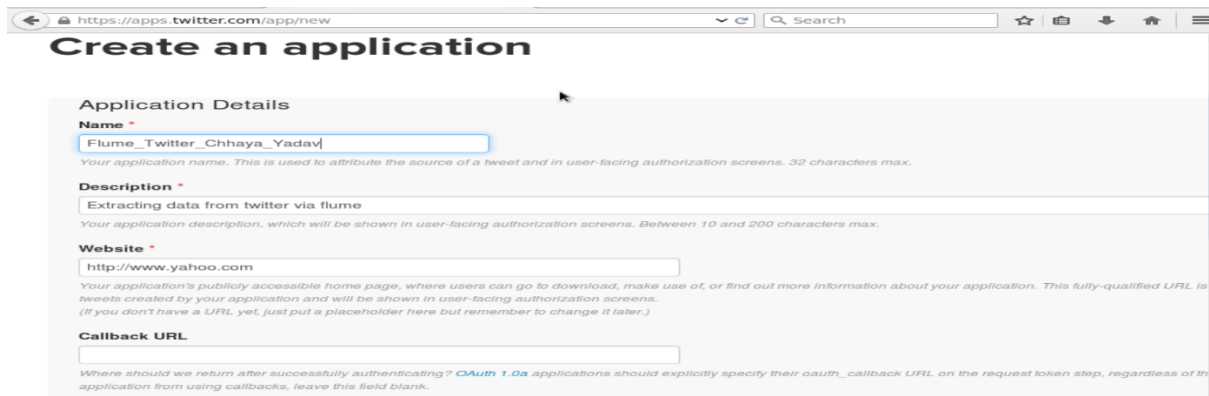


STEP 15:

Change the URL to <https://apps.twitter.com>

STEP 16:

Click on Create New App to create a new application and enter all the details in the application:

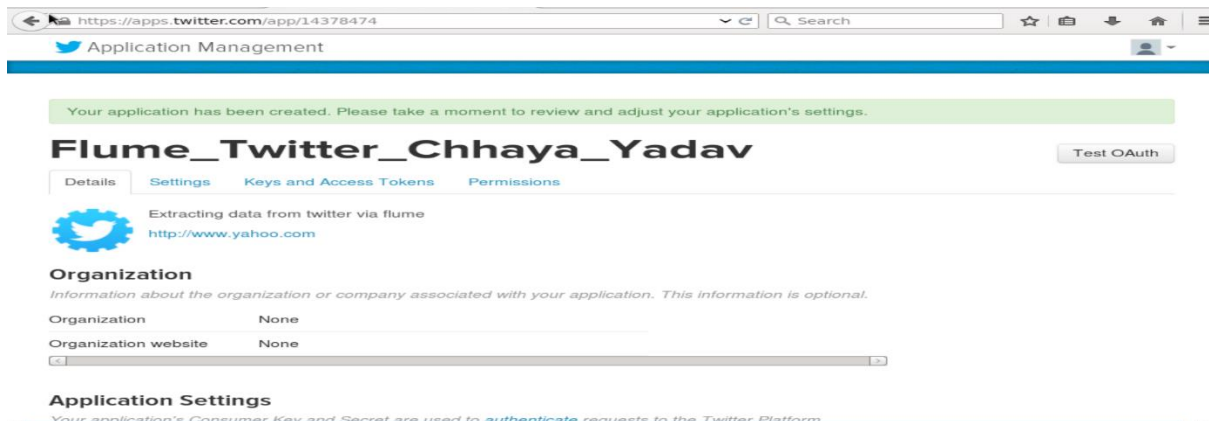


STEP 17:

Check Yes, I agree and click on Create your Twitter application:

STEP 18:

Your Application will be created:



STEP 19:

Click on Keys and Access Tokens, you will get Consumer Key and Consumer Secret.

STEP 20:

Scroll down and Click on Create my access token:

STEP 21:

Use below link to download flume.conf file

Download link:

<https://drive.google.com/file/d/0B-CI0fLnRozdIRuN3pPWEJ1RHc/view?usp=sharing>

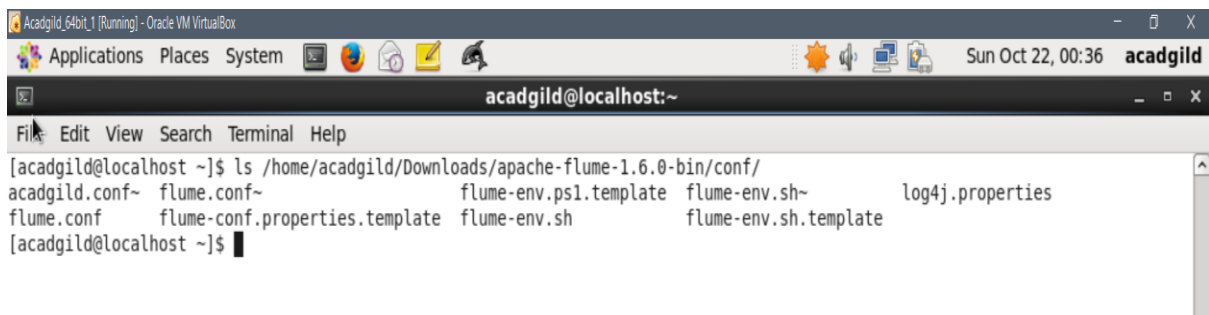
STEP 22:

Put the flume.conf in the conf directory of apache-flume-1.6.0-bin

Commands:

```
[acadgild@quickstart ~]$ sudo cp /home/acadgild/Downloads/flume.conf
```

```
/home/acadgild/Downloads/apache-flume-1.6.0-bin/conf/
```



STEP 23:

Create a hdfs directory to store twitter data.

Command:

```
[acadgild@quickstart ~]$ hadoop dfs -mkdir -p /user/flume/tweets/
```

STEP 24:

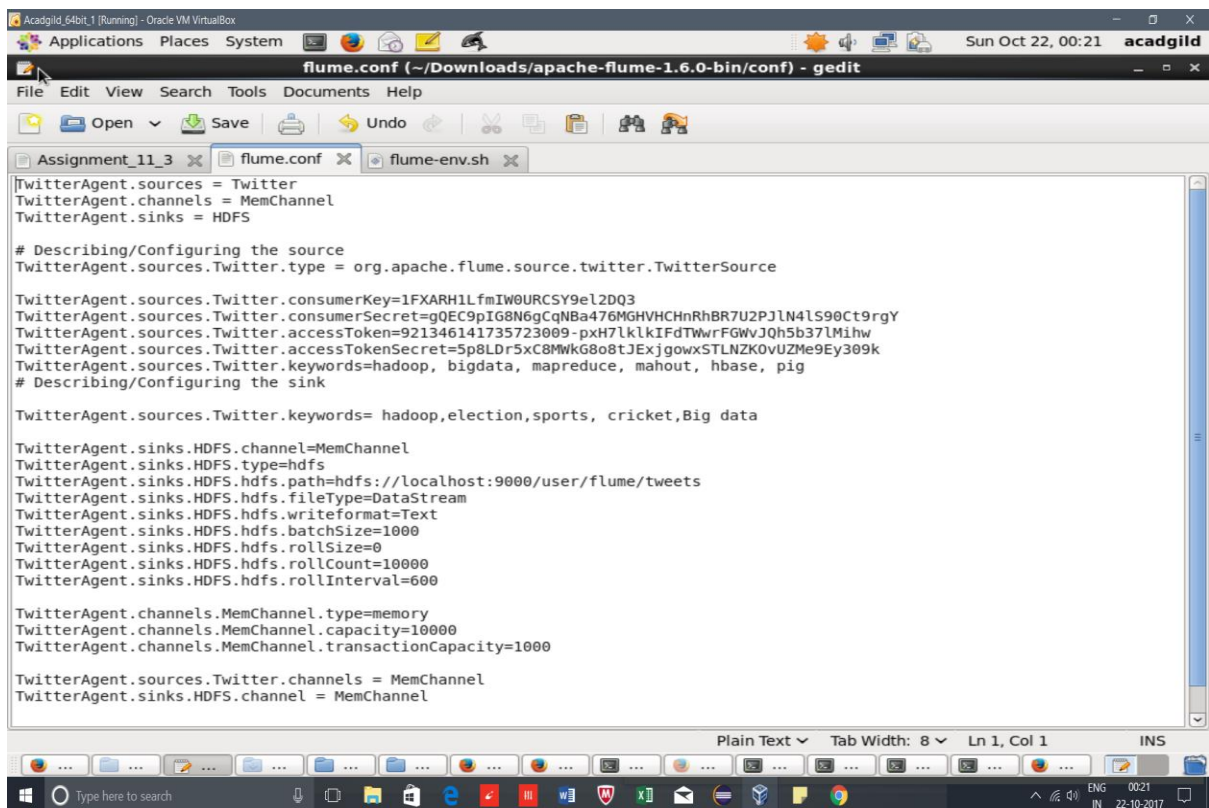
Edit flume.conf.

Replace all the below highlighted credentials in flume.conf with the credentials (Consumer Key, Consumer Secret, Access Token, Access Token Secret) you received after creating the application very carefully, rest all will remain same, save the file and close it.

Commands:

```
[acadgild@quickstart ~]$ cd /home/acadgild/Downloads/apache-flume-1.6.0-bin/
```

```
[acadgild@quickstart ~]$ sudo gedit /home/acadgild/Downloads/apache-flume-1.6.0-bin/conf/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=1FXARH1LfmIW0URCSY9eL2DQ3
TwitterAgent.sources.Twitter.consumerSecret=g0EC9pIG8N6gCqNBa476MGHVHCHnRhBR7U2PJlN4lS90Ct9rgY
TwitterAgent.sources.Twitter.accessToken=921346141735723009-pxH7lkIFdTwwrFGWvJQh5b37lMihw
TwitterAgent.sources.Twitter.accessTokenSecret=5p8LDr5xC8MwG8o8tJExjgowxSTLNZK0vUZMe9Ey309k
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, pig
# 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

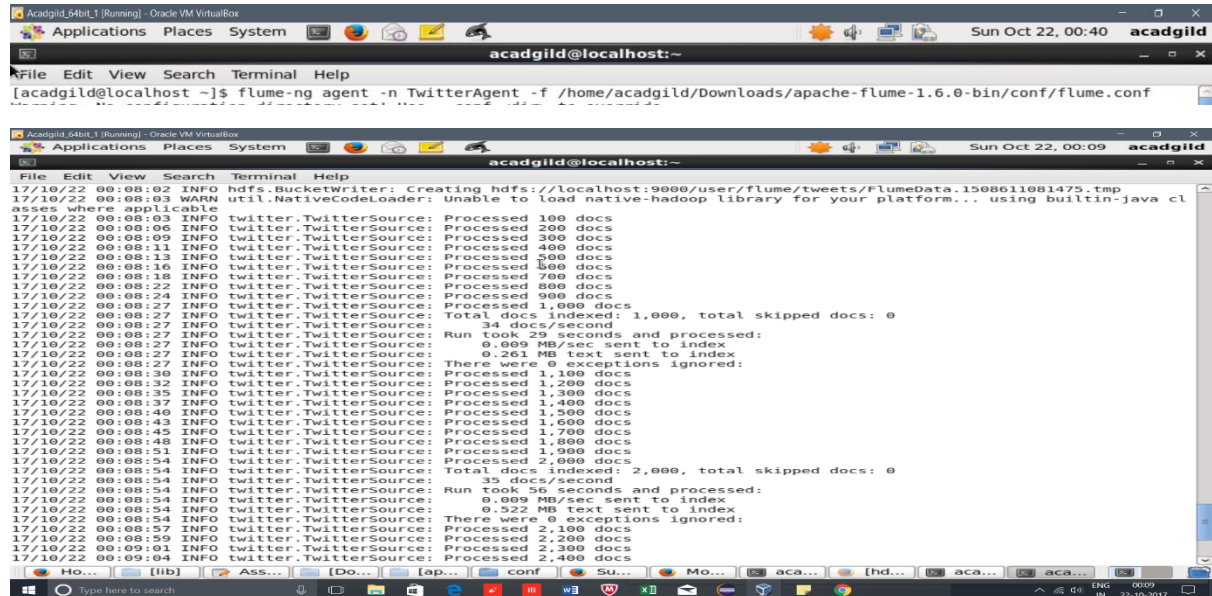
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

STEP 25:

Start fetching the data from twitter:

Command:

```
[acadgild@quickstart ~]$ ./bin/flume-ng agent -n TwitterAgent -c conf -f /home/acadgild/Downloads/apache-flume-1.6.0-bin/conf/flume.conf
```

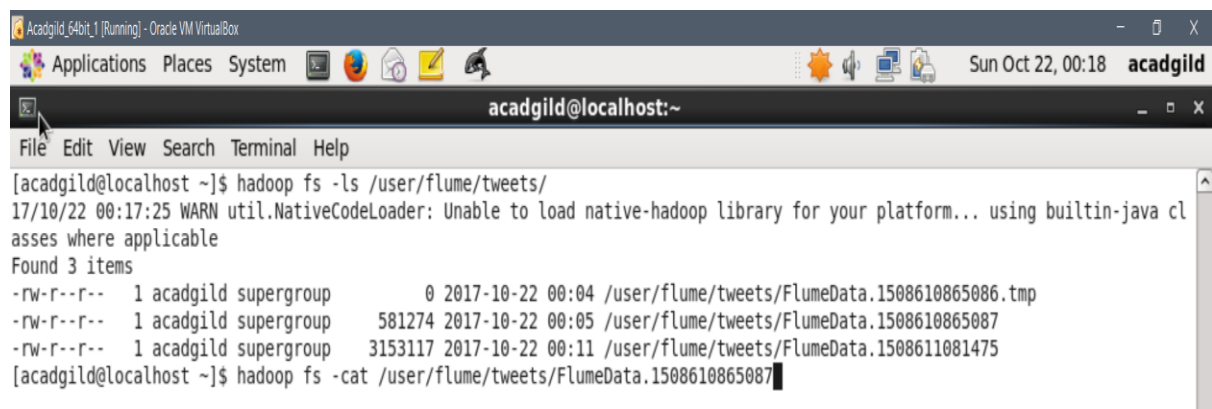


```
Acadgild_64bit_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/Downloads/apache-flume-1.6.0-bin/conf/flume.conf

Acadgild_64bit_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
17/10/22 00:08:02 INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/flume/tweets/FlumeData.1508611081475.tmp
17/10/22 00:08:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/10/22 00:08:05 INFO twitter.TwitterSource: Processed 100 docs
17/10/22 00:08:06 INFO twitter.TwitterSource: Processed 200 docs
17/10/22 00:08:09 INFO twitter.TwitterSource: Processed 300 docs
17/10/22 00:08:11 INFO twitter.TwitterSource: Processed 400 docs
17/10/22 00:08:13 INFO twitter.TwitterSource: Processed 500 docs
17/10/22 00:08:16 INFO twitter.TwitterSource: Processed 600 docs
17/10/22 00:08:18 INFO twitter.TwitterSource: Processed 700 docs
17/10/22 00:08:22 INFO twitter.TwitterSource: Processed 800 docs
17/10/22 00:08:24 INFO twitter.TwitterSource: Processed 900 docs
17/10/22 00:08:27 INFO twitter.TwitterSource: Processed 1,000 docs
17/10/22 00:08:27 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
17/10/22 00:08:27 INFO twitter.TwitterSource: 34 docs/second
17/10/22 00:08:27 INFO twitter.TwitterSource: Run took 29 seconds and processed:
0.009 MB/sec sent to index
17/10/22 00:08:27 INFO twitter.TwitterSource: 0.261 MB text sent to index
17/10/22 00:08:30 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/10/22 00:08:30 INFO twitter.TwitterSource: Processed 1,100 docs
17/10/22 00:08:32 INFO twitter.TwitterSource: Processed 1,200 docs
17/10/22 00:08:35 INFO twitter.TwitterSource: Processed 1,300 docs
17/10/22 00:08:37 INFO twitter.TwitterSource: Processed 1,400 docs
17/10/22 00:08:40 INFO twitter.TwitterSource: Processed 1,500 docs
17/10/22 00:08:43 INFO twitter.TwitterSource: Processed 1,600 docs
17/10/22 00:08:45 INFO twitter.TwitterSource: Processed 1,700 docs
17/10/22 00:08:48 INFO twitter.TwitterSource: Processed 1,800 docs
17/10/22 00:08:51 INFO twitter.TwitterSource: Processed 1,900 docs
17/10/22 00:08:54 INFO twitter.TwitterSource: Processed 2,000 docs
17/10/22 00:08:54 INFO twitter.TwitterSource: Total docs indexed: 2,000, total skipped docs: 0
17/10/22 00:08:54 INFO twitter.TwitterSource: 35 docs/second
17/10/22 00:08:54 INFO twitter.TwitterSource: Run took 56 seconds and processed:
0.009 MB/sec sent to index
17/10/22 00:08:54 INFO twitter.TwitterSource: 0.522 MB text sent to index
17/10/22 00:08:57 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/10/22 00:08:59 INFO twitter.TwitterSource: Processed 2,100 docs
17/10/22 00:08:59 INFO twitter.TwitterSource: Processed 2,200 docs
17/10/22 00:09:01 INFO twitter.TwitterSource: Processed 2,300 docs
17/10/22 00:09:04 INFO twitter.TwitterSource: Processed 2,400 docs
```

STEP 26:

```
[acadgild@quickstart ~]$ hadoop fs -cat /user/flume/tweets/
```



```
Acadgild_64bit_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets/
17/10/22 00:17:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 3 items
-rw-r--r-- 1 acadgild supergroup 0 2017-10-22 00:04 /user/flume/tweets/FlumeData.1508610865086.tmp
-rw-r--r-- 1 acadgild supergroup 581274 2017-10-22 00:05 /user/flume/tweets/FlumeData.1508610865087
-rw-r--r-- 1 acadgild supergroup 3153117 2017-10-22 00:11 /user/flume/tweets/FlumeData.1508611081475
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.1508610865087
```

Final Output: Streamed data from twitter
