

## Assignment- 12 Flume

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

**Solution:** I will explain the step-by-step process to stream data from Twitter and stores in the HDFS.

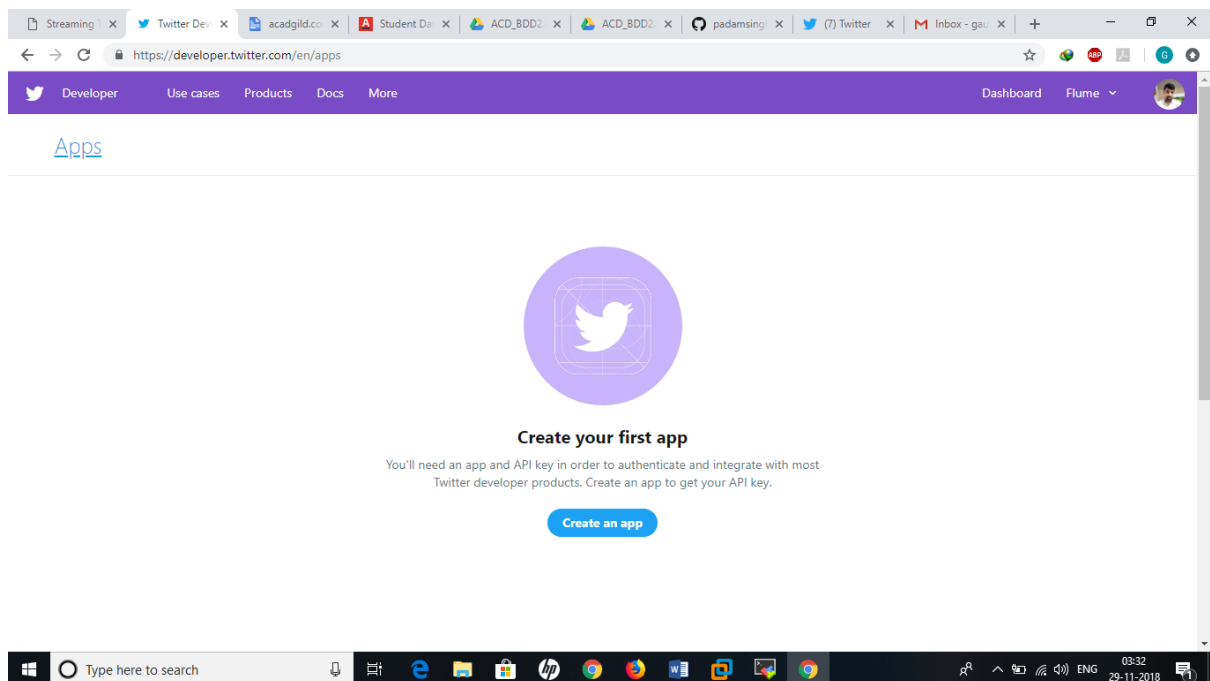
### **Step 1:**

Login to the twitter account.

### **Step 2:**

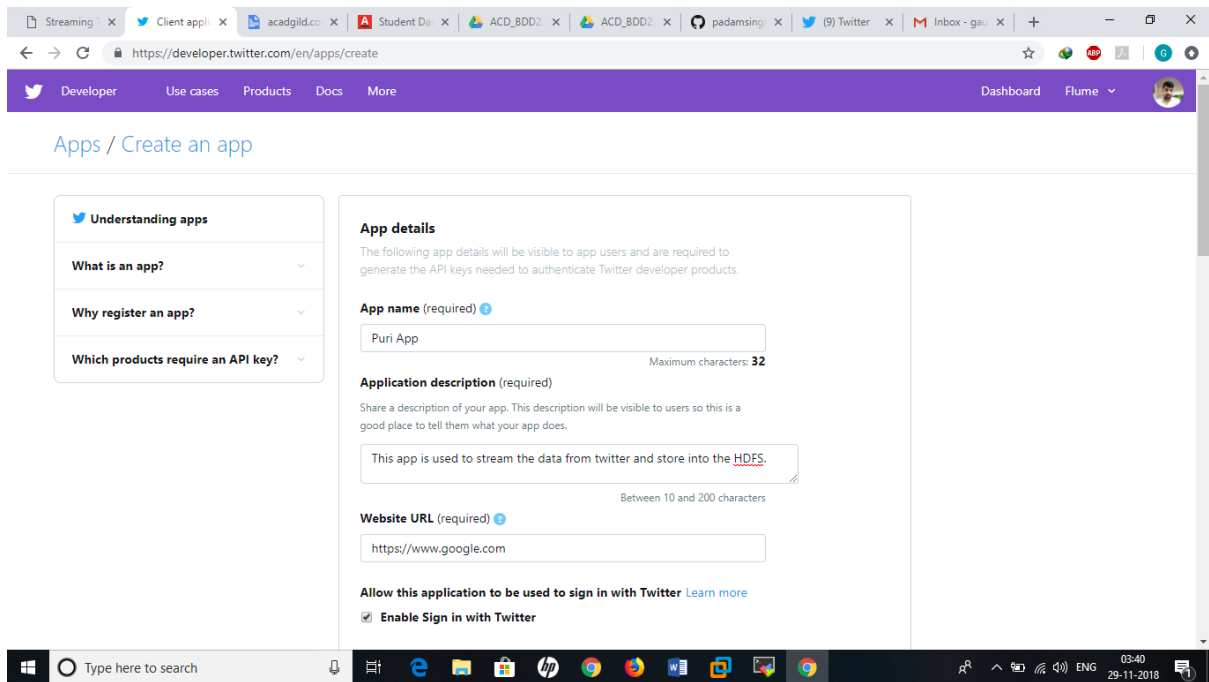
Go to the following link and click the 'create new app' button.

<https://apps.twitter.com/app>



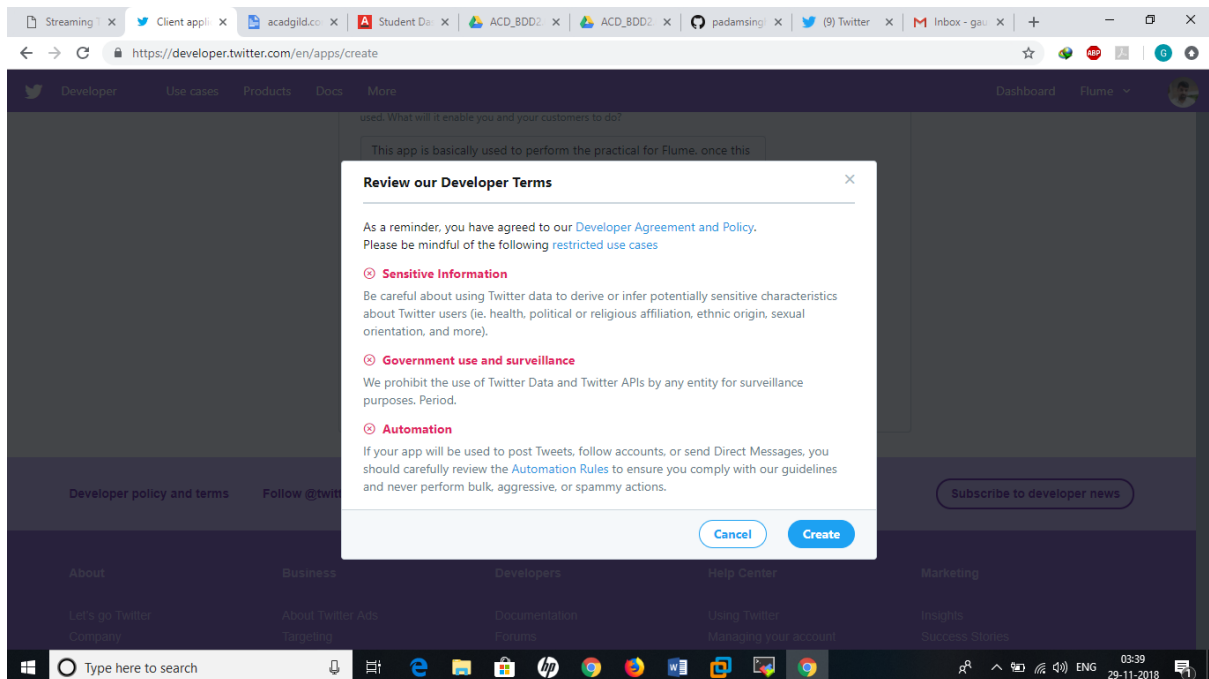
### **Step 3:**

Enter the necessary details.



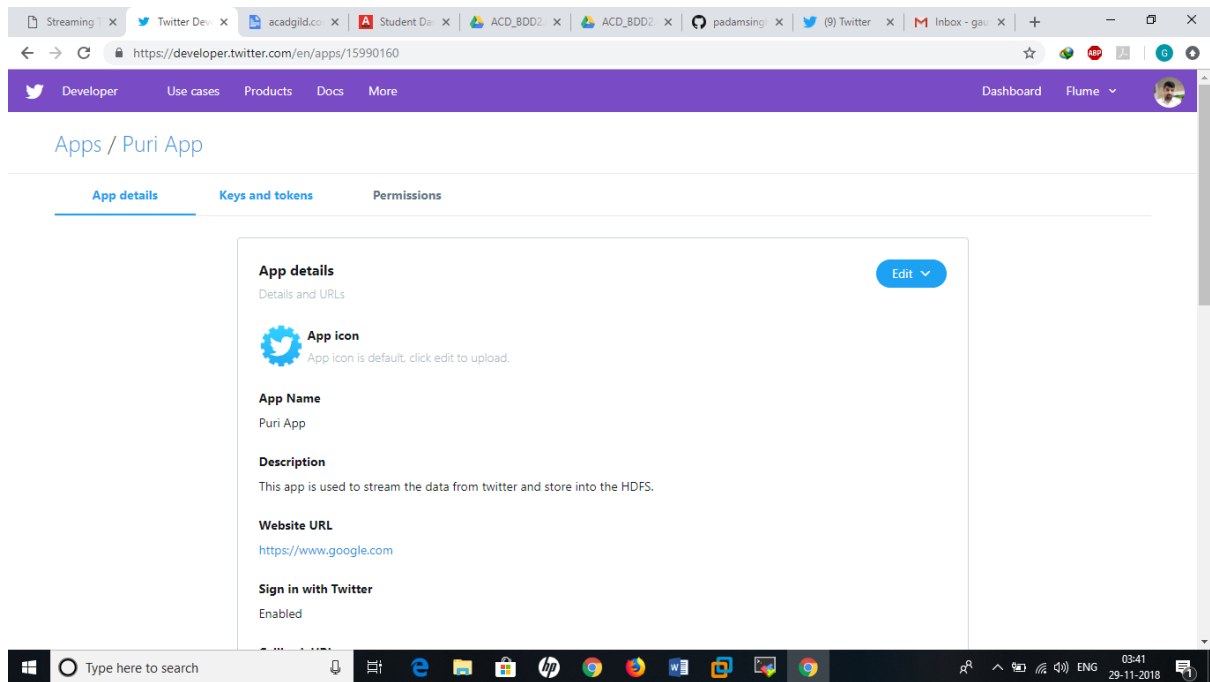
#### Step 4:

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.

#### Step 9:

Create a new file inside the conf directory inside the Flume-extracted directory and named as flume.conf with the configuration.

```
agent.channels.memoryChannel.capacity = 100
[acagild@localhost conf]$ nano flume.conf
You have new mail in /var/spool/mail/acagild
[acagild@localhost conf]$ cat 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=9WlovtYbu9NLSeWlsEEHCohMa
TwitterAgent.sources.Twitter.consumerSecret=fv7dR44JCv7J89lnxEWz0Z0FPuEgE59SYJXyl6Jk80agEKa3I
TwitterAgent.sources.Twitter.accessToken=989497225314107392-Hok9TP22U04t2HLsagltimH2z0QXUIC
TwitterAgent.sources.Twitter.accessTokenSecret=3FRK2N8HQL9aJLR61vuxEzSzjin0Ew9f2hAMRL3P1PFo
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink
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
[acagild@localhost conf]$ ls
flume.conf flume-conf.properties.template flume-env.ps1.template flume-env.sh.template log4j.properties
[acagild@localhost conf]$
```

### Step 10:

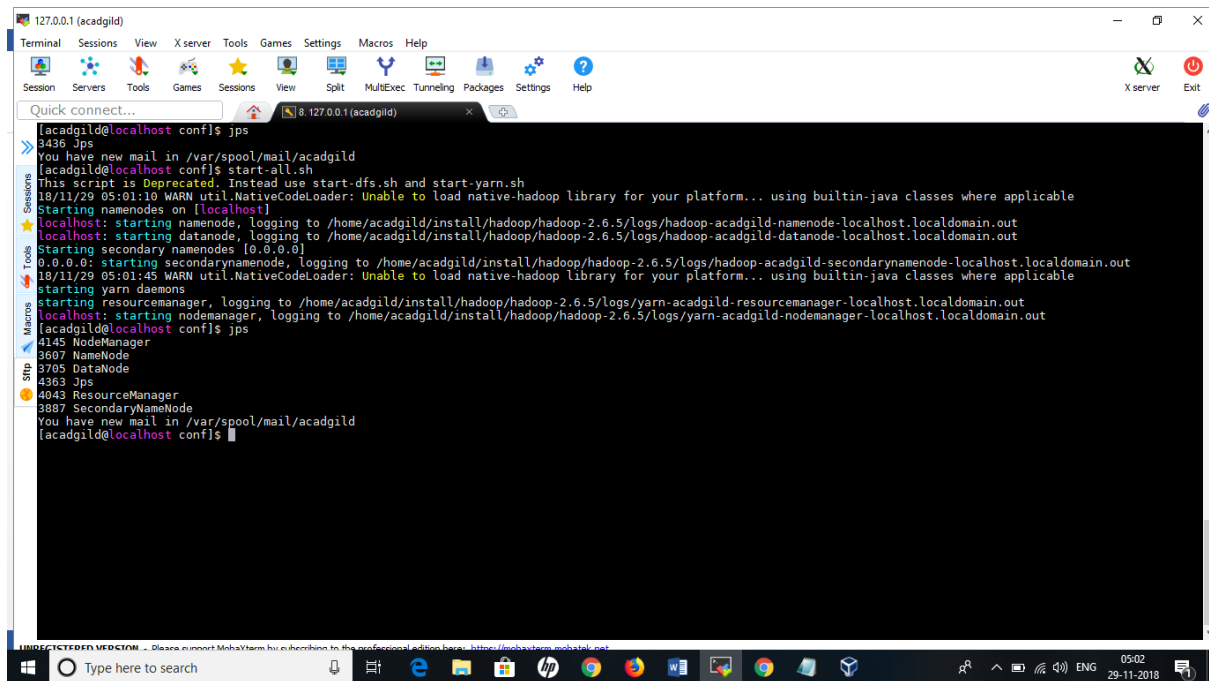
We have to decide which keywords tweet data to be collected from the twitter application. So, you can change the keywords in the `TwitterAgent.sources.Twitter.keywords` command.

In our example, we are fetching tweet data related to KartarpurCorridor, Modi

### Step 11:

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.

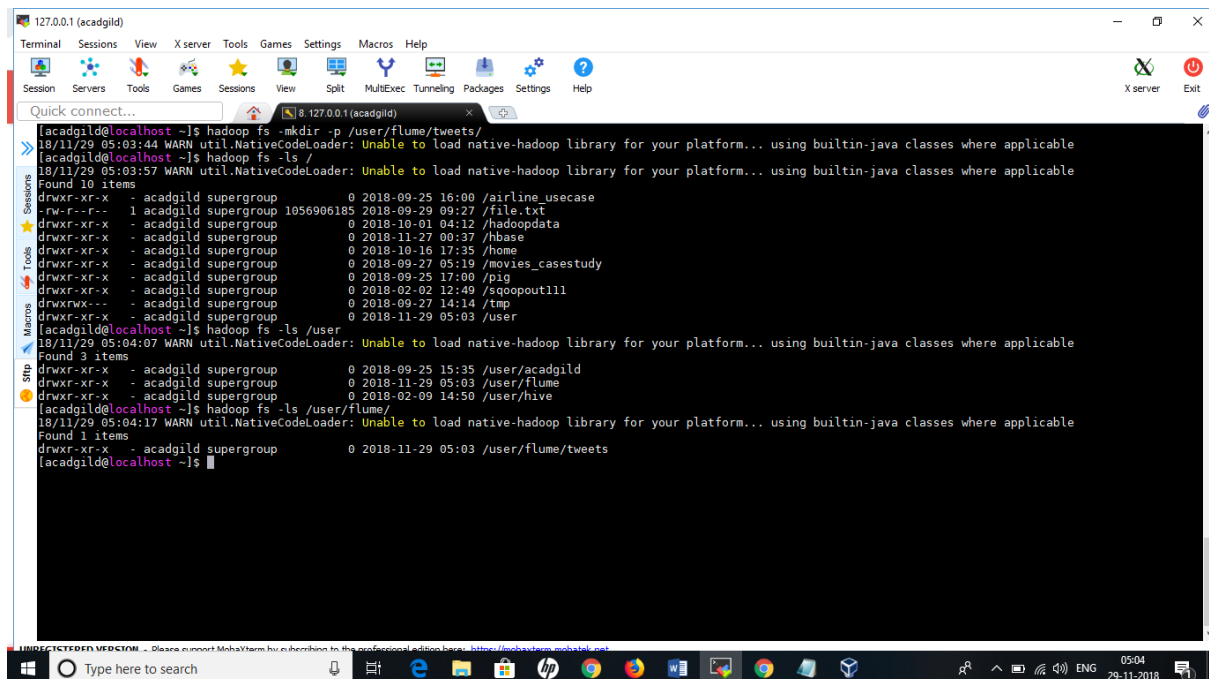


```
[acagild@localhost conf]$ jps
3436 Jps
[acagild@localhost conf]$ start-all.sh
This script is deprecated. Instead use start-dfs.sh and start-yarn.sh
18/11/29 05:01:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acagild/install/hadoop/hadoop-2.6.5/logs/hadoop-acagild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /home/acagild/install/hadoop/hadoop-2.6.5/logs/hadoop-acagild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acagild/install/hadoop/hadoop-2.6.5/logs/hadoop-acagild-secondarynamenode-localhost.localdomain.out
18/11/29 05:01:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/acagild/install/hadoop/hadoop-2.6.5/logs/yarn-acagild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /home/acagild/install/hadoop/hadoop-2.6.5/logs/yarn-acagild-nodemanager-localhost.localdomain.out
[acagild@localhost conf]$ jps
4145 NodeManager
3607 NameNode
3705 DataNode
4363 Jps
4043 ResourceManager
3887 SecondaryNameNode
You have new mail in /var/spool/mail/acagild
[acagild@localhost conf]$
```

## Step 12:

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

*Hadoop fs -mkdir /user/flume/tweets*



```
[acagild@localhost ~]$ hadoop fs -mkdir -p /user/flume/tweets/
18/11/29 05:03:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[acagild@localhost ~]$ hadoop fs -ls /
18/11/29 05:03:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 10 items
drwxr-xr-x - acagild supergroup 0 2018-09-25 16:00 /airline usecase
-rw-r--r-- 1 acagild supergroup 1056906185 2018-09-29 09:27 /file.txt
drwxr-xr-x - acagild supergroup 0 2018-10-01 04:12 /hadoopdata
drwxr-xr-x - acagild supergroup 0 2018-11-27 00:37 /hbase
drwxr-xr-x - acagild supergroup 0 2018-10-16 17:35 /home
drwxr-xr-x - acagild supergroup 0 2018-09-27 05:19 /movies_casestudy
drwxr-xr-x - acagild supergroup 0 2018-09-25 17:00 /pig
drwxr-xr-x - acagild supergroup 0 2018-02-02 12:49 /sqoopout111
drwxrwx-- - acagild supergroup 0 2018-09-27 14:14 /tmp
drwxr-xr-x - acagild supergroup 0 2018-11-29 05:03 /user
[acagild@localhost ~]$ hadoop fs -ls /user
18/11/29 05:04:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x - acagild supergroup 0 2018-09-25 15:35 /user/acagild
drwxr-xr-x - acagild supergroup 0 2018-11-29 05:03 /user/flume
drwxr-xr-x - acagild supergroup 0 2018-02-09 14:50 /user/hive
[acagild@localhost ~]$ hadoop fs -ls /user/flume/
18/11/29 05:04:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - acagild supergroup 0 2018-11-29 05:03 /user/flume/tweets
[acagild@localhost ~]$
```

## Step 13:

For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

[illegible]

The screenshot displays a Windows 10 desktop environment. At the top, a taskbar shows the Start button, a search bar, and several pinned application icons including File Explorer, Edge, and various utility tools. The main area of the screen is occupied by a terminal window titled '127.0.0.1'. The terminal shows the output of a Java application running on a local machine. The application is a Twitter data processing tool that connects to a Twitter stream and writes data to a local HDFS sink. The output shows the progress of processing documents, including warnings about native Hadoop libraries and a timeout error. The desktop background is a Windows 10 taskbar with various icons and a search bar.

```

127.0.0.1
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultExec Tunneling Packages Settings Help

Quick connect... 4 127.0.0.1 5 127.0.0.1

18/12/13 17:08:55 INFO twitter4j.TwitterStreamImpl: Connection established.
18/12/13 17:08:55 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/12/13 17:08:58 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/12/13 17:09:01 INFO twitter.TwitterSource: Processed 100 docs
18/12/13 17:09:01 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/flume/tweets/FlumeData.1544701137969.tmp
18/12/13 17:09:01 INFO twitter.TwitterSource: Processed 200 docs
18/12/13 17:09:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/12/13 17:09:03 INFO twitter.TwitterSource: Processed 300 docs
18/12/13 17:09:06 INFO twitter.TwitterSource: Processed 400 docs
18/12/13 17:09:09 INFO twitter.TwitterSource: Processed 500 docs
18/12/13 17:09:11 WARN hdfs.HDFSEventsSink: HDFS IO error
java.io.IOException: Callabte timed out after 10000 ms on file: hdfs://localhost:8020/user/flume/tweets/FlumeData.1544701137969.tmp
    at org.apache.flume.sink.hdfs.BucketWriter.callWithTimeout(BucketWriter.java:715)
    at org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:252)
    at org.apache.flume.sink.hdfs.BucketWriter.append(BucketWriter.java:541)
    at org.apache.flume.sink.hdfs.HDFSEventsSink.process(HDFSEventsSink.java:401)
    at org.apache.flume.sink.DefaultSinkProcessor.process(DefaultSinkProcessor.java:67)
    at org.apache.flume.SinkRunnersPollingRunner.run(SinkRunner.java:145)
    at java.lang.Thread.run(Thread.java:748)
Caused by: java.util.concurrent.TimeoutException
    at java.util.concurrent.FutureTask.get(FutureTask.java:205)
    at org.apache.flume.sink.hdfs.BucketWriter.callWithTimeout(BucketWriter.java:708)
    ... 6 more
18/12/13 17:09:11 INFO twitter.TwitterSource: Processed 600 docs
18/12/13 17:09:14 INFO twitter.TwitterSource: Processed 700 docs
18/12/13 17:09:14 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/flume/tweets/FlumeData.1544701137970.tmp
18/12/13 17:09:16 INFO twitter.TwitterSource: Processed 800 docs
18/12/13 17:09:19 INFO twitter.TwitterSource: Processed 900 docs
18/12/13 17:09:21 INFO twitter.TwitterSource: Processed 1,000 docs
18/12/13 17:09:21 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
18/12/13 17:09:21 INFO twitter.TwitterSource: 31 docs/second
18/12/13 17:09:21 INFO twitter.TwitterSource: Run took 32 seconds and processed:
18/12/13 17:09:21 INFO twitter.TwitterSource: 0.008 MB/sec sent to index
18/12/13 17:09:21 INFO twitter.TwitterSource: 0.269 MB text sent to index
18/12/13 17:09:21 INFO twitter.TwitterSource: There were 0 exceptions ignored:
18/12/13 17:09:23 INFO twitter.TwitterSource: Processed 1,100 docs
18/12/13 17:09:26 INFO twitter.TwitterSource: Processed 1,200 docs
18/12/13 17:09:29 INFO twitter.TwitterSource: Processed 1,300 docs

UNREGISTERED VERSION - Please support MobatTerm by subscribing to the professional edition here: https://mobatxterm.mobatek.net

```

```
hadoop fs -ls /user/flume/tweets
```

