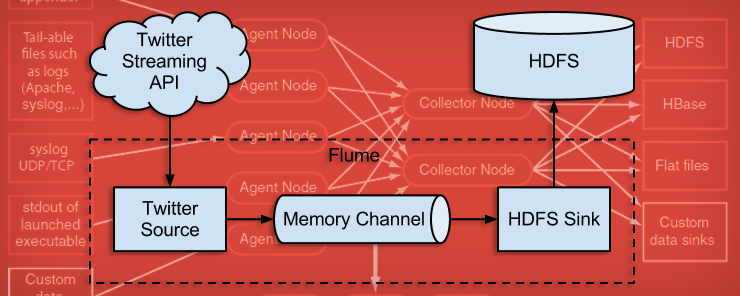
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/Flume-Installation-and-Streaming-Twitter-Data-Using-Flume.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/Flume-Installation-and-Streaming-Twitter-Data-Using-Flume.png)

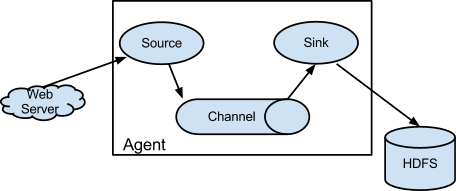
Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

Flume lets Hadoop users make the most of valuable log data. Specifically, Flume allows users to:

* Stream data from multiple sources into Hadoop for analysis
* Collect high-volume Web logs in real time
* Insulate themselves from transient spikes when the rate of incoming data exceeds the rate at which data can be written to the destination
* Guarantee data delivery
* Scale horizontally to handle additional data volume

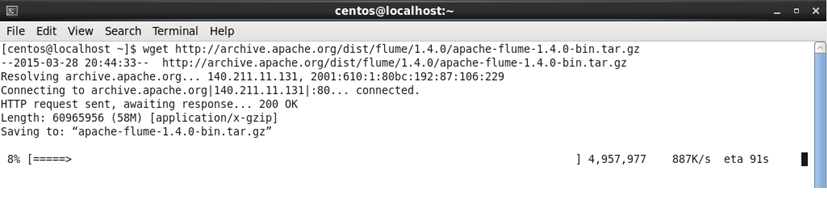
Flume’s high-level architecture is focused on delivering a streamlined codebase that is easy-to-use and easy-to-extend. The project team has designed Flume with the following components:

* **Event** – a singular unit of data that is transported by Flume (typically a single log entry
* **Source**– the entity through which data enters into Flume. Sources either actively poll for data or passively wait for data to be delivered to them. A variety of sources allow data to be collected, such as log4j logs and syslogs.
* **Sink**– the entity that delivers the data to the destination. A variety of sinks allow data to be streamed to a range of destinations. One example is the HDFS sink that writes events to HDFS.
* **Channel**– the conduit between the Source and the Sink. Sources ingest events into the channel and the sinks drain the channel.
* **Agent**– any physical Java virtual machine running Flume. It is a collection of sources, sinks and channels.
* **Client**– produces and transmits the Event to the Source operating within the Agent

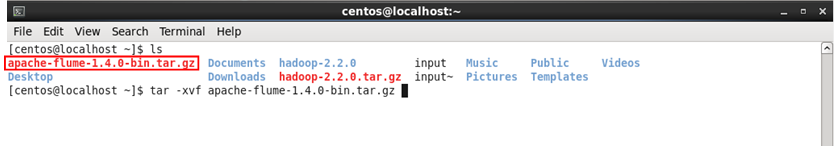
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/12.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/12.png)

A flow in Flume starts from the Client (Web Server). The Client transmits the event to a Source operating within the Agent. The Source receiving this event then delivers it to one or more Channels. These Channels are drained by one or more Sinks operating within the same Agent. Channels allow decoupling of ingestion rate from drain rate using the familiar producer-consumer model of data exchange. When spikes in client side activity cause data to be generated faster than what the provisioned capacity on the destination can handle, the channel size increases. This allows sources to continue normal operation for the duration of the spike. Flume agents can be chained together by connecting the sink of one agent to the source of another agent. This enables the creation of complex dataflow topologies.

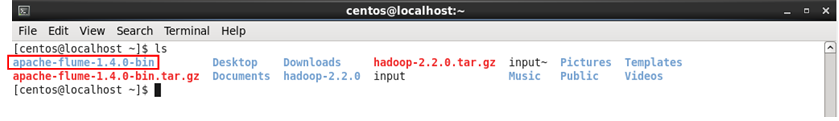
**Now we will install apache flume on our virtual machine.**  
  
**STEP 1:**  
  
Download flume:  
**Command:** wget http://archive.apache.org/dist/flume/1.4.0/apache-flume-1.4.0-bin.tar.gz

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/22.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/22.png)

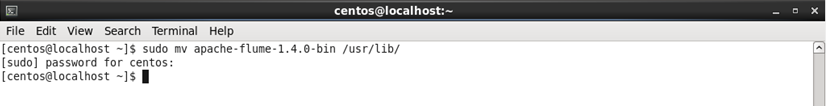
**Command:** ls

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/31.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/31.png)

**STEP 2:**  
  
Extract file from flume tar file.  
**Command:** tar -xvf apache-flume-1.4.0-bin.tar.gz  
**Command:** ls

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/41.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/41.png)

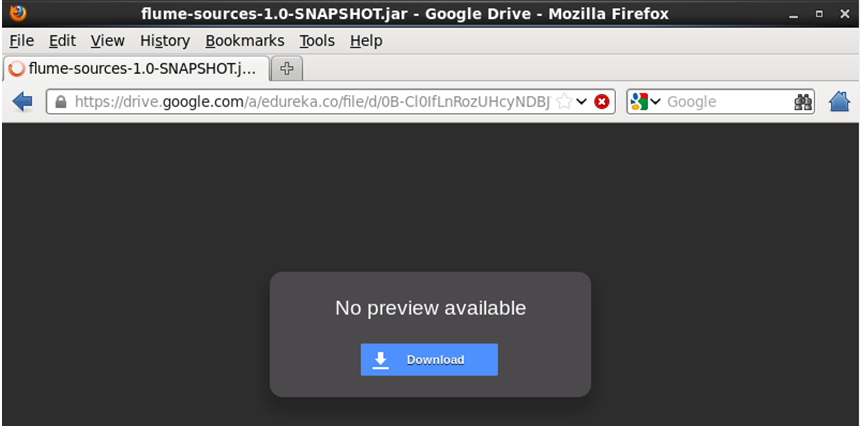
**STEP 3:**  
  
Put apache-flume-1.4.0-bin directory inside /usr/lib/ directory.  
  
**Command:** sudo mv apache-flume-1.4.0-bin /usr/lib/

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/51.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/51.png)

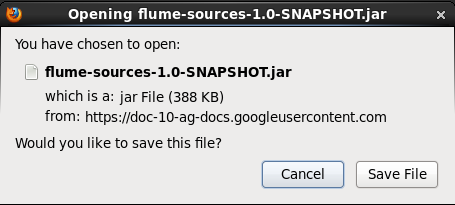
**STEP 4:**  
  
We need to remove protobuf-java-2.4.1.jar and guava-10.1.1.jar from lib directory of apache-flume-1.4.0-bin ( when using hadoop-2.x )  
  
**Command:** sudo rm /usr/lib/apache-flume-1.4.0-bin/lib/protobuf-java-2.4.1.jar /usr/lib/apache-flume-1.4.0-bin/lib/guava-10.0.1.jar

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/61.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/61.png)

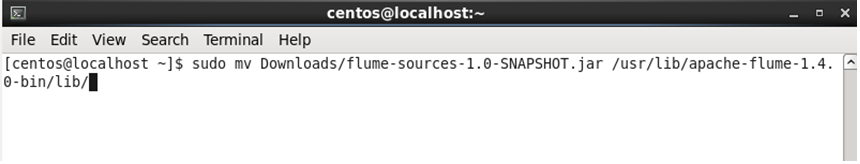
**STEP 5:**  
  
Use below link and download flume-sources-1.0-SNAPSHOTS.jar  
https://drive.google.com/file/d/0B-Cl0IfLnRozUHcyNDBJWnNxdHc/view?usp=sharing

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/71.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/71.png)

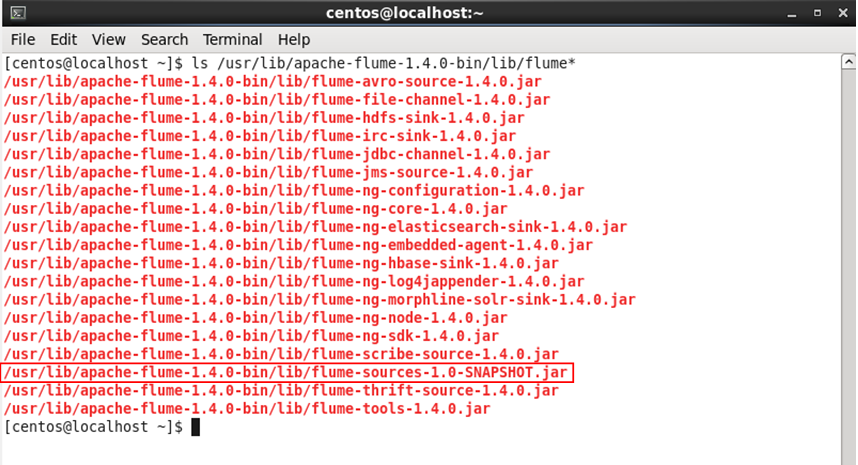
Save the file.

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/81.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/81.png)

**STEP 6:**  
  
Move the flume-sources-1.0-SNAPSHOT.jar file from Downloads directory to lib directory of apache flume:  
  
**Command:** sudo mv Downloads/flume-sources-1.0-SNAPSHOT.jar /usr/lib/apache-flume-1.4.0-bin/lib/

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/9.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/9.png)

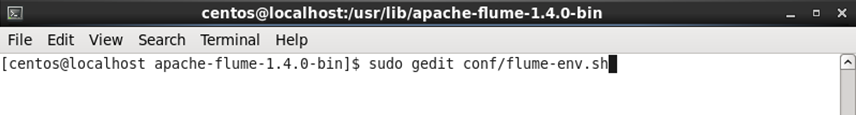
**STEP 7:**  
  
Check whether flume SNAPSHOT has moved to the lib folder of apache flume:  
  
**Command:** ls /usr/lib/apache-flume-1.4.0-bin/lib/flume\*

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/10.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/10.png)

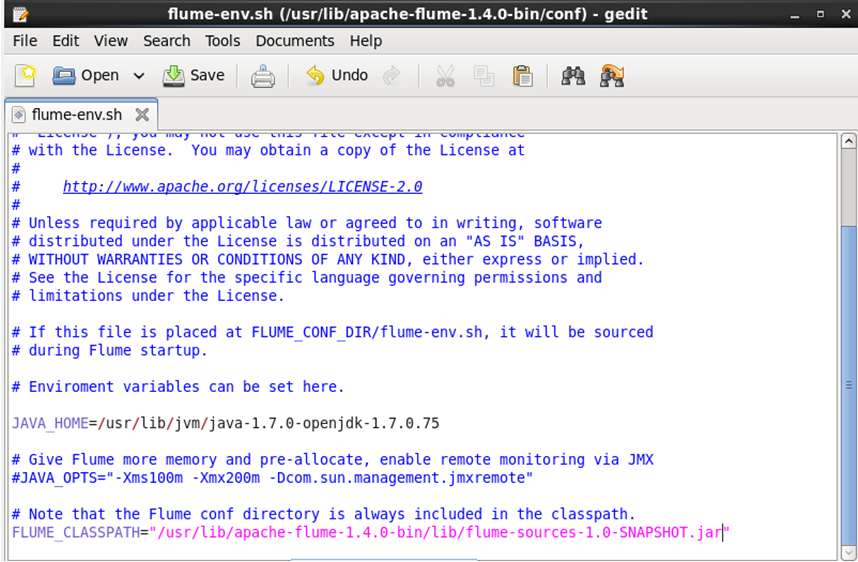
**STEP 8:**  
  
Copy flume-env.sh.template content to flume-env.sh  
  
**Command:** cd /usr/lib/apache-flume-1.4.0-bin/  
  
**Command:** sudo cp conf/flume-env.sh.template conf/flume-env.sh

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/111.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/111.png)

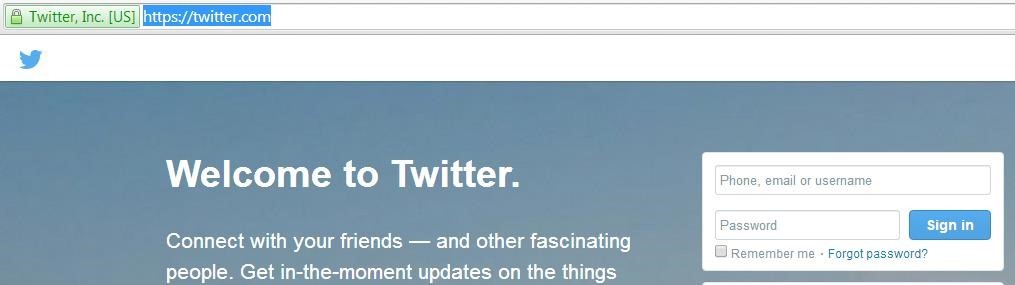
**STEP 9:**  
  
Edit flume-env.sh as mentioned in below snapshot.  
  
**command:** sudo gedit conf/flume-env.sh

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/121.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/121.png)

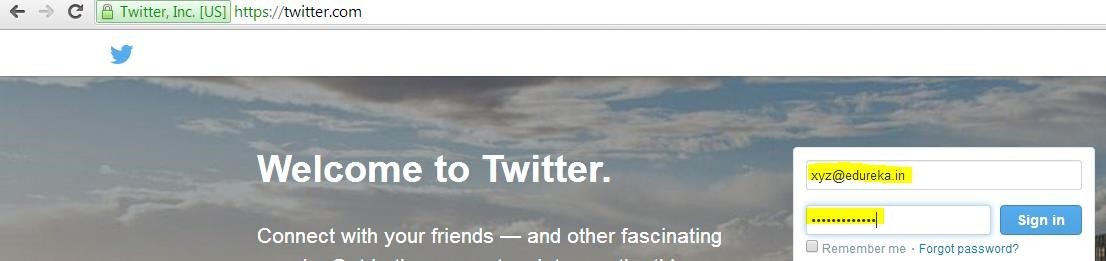
Set JAVA\_HOME and FLUME\_CLASSPATH as shown in below snapshot.

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/13.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/13.png)

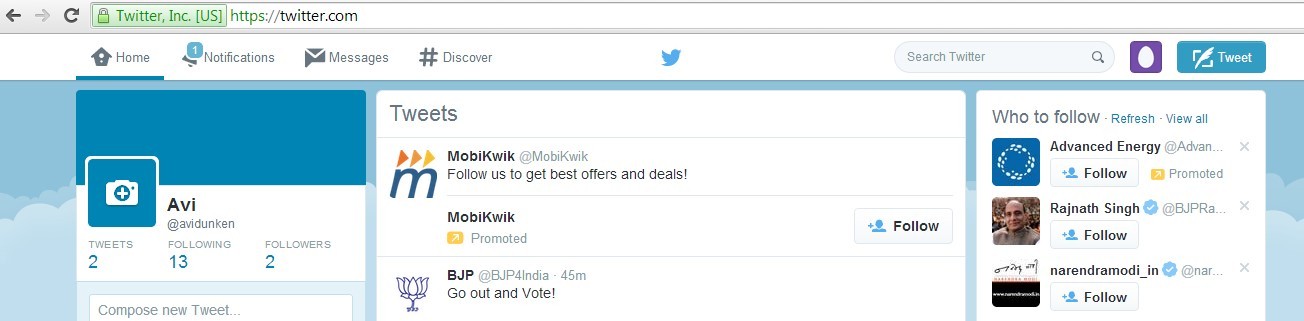
Now we have installed flume on our machine. Lets run flume to stream twitter data on to HDFS.  
We need to create an application in twitter and use its credentials to fetch data.  
  
**STEP 10:**  
  
Open a Browser and go to the below URL:  
  
URL:https://twitter.com/

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/14.jpg)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/14.jpg)

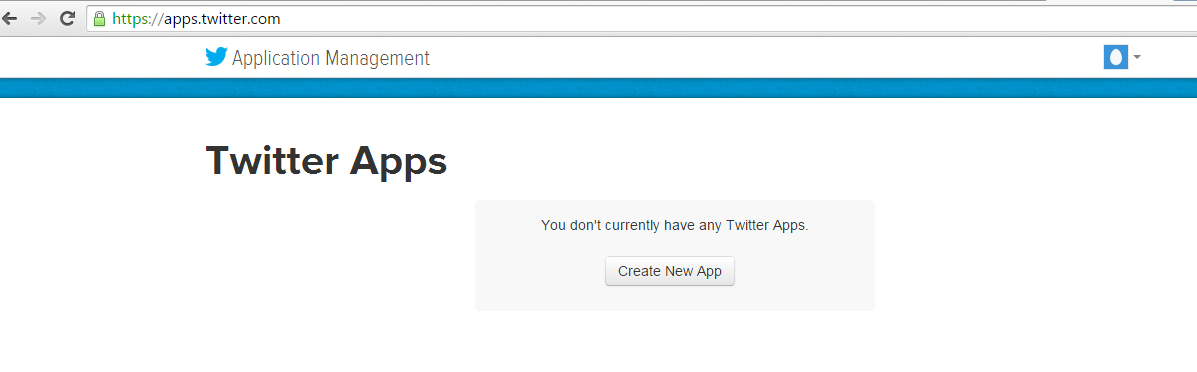
**STEP 11:**  
  
Enter your Twitter account credentials and sign in:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/15.jpg)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/15.jpg)

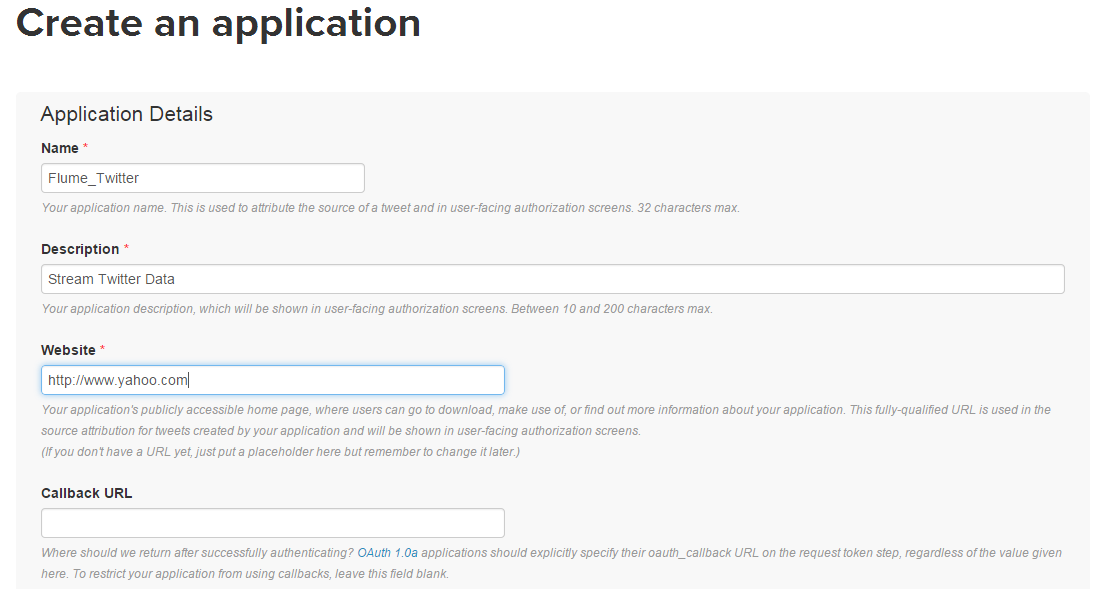
**STEP 12:**  
  
Your twitter home page will open:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/16.jpg)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/16.jpg)

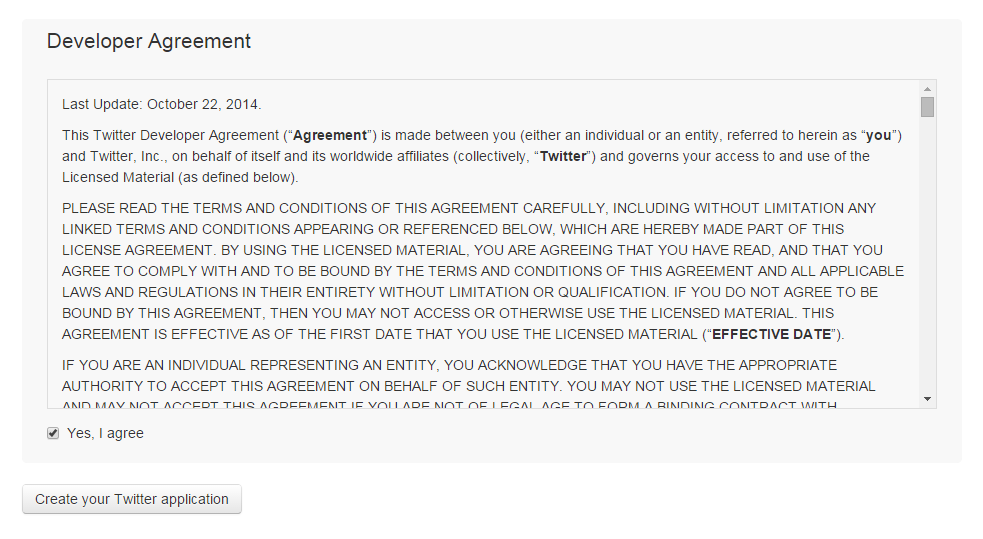
**STEP 13:**  
  
Change the URL to https://apps.twitter.com

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/17.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/17.png)

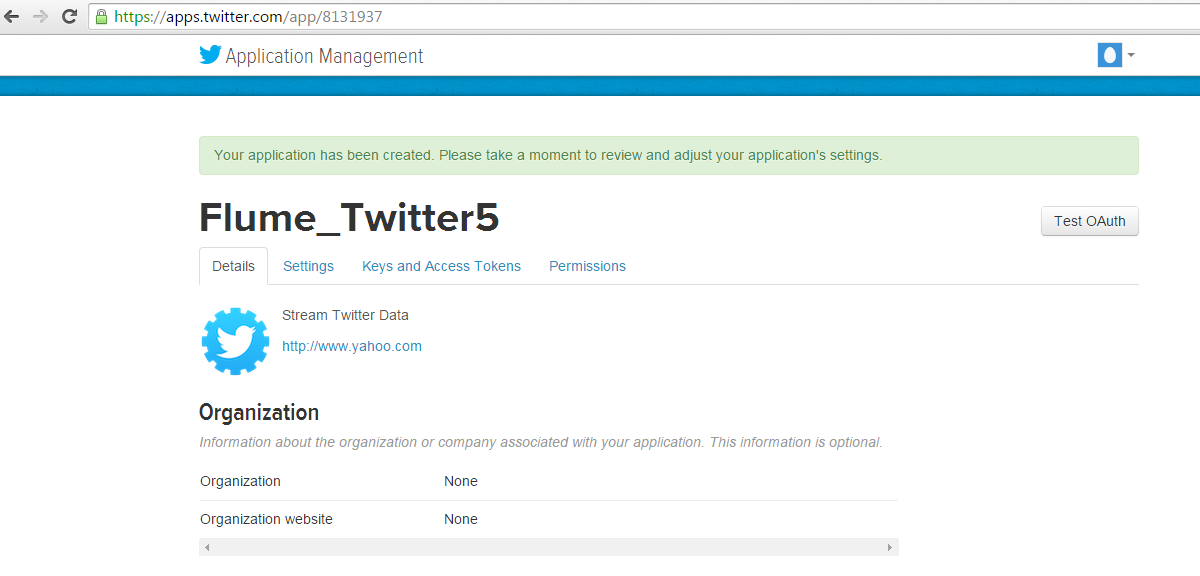
**STEP 14:**  
  
Click on Create New App to create a new application and enter all the details in the application:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/18.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/18.png)

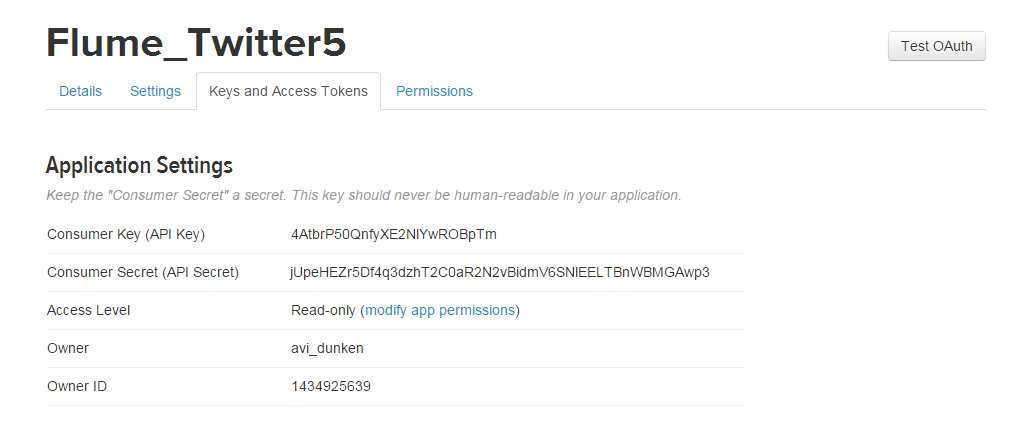
**STEP 15:**  
  
Check Yes, I agree and click on Create your Twitter application:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/19.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/19.png)

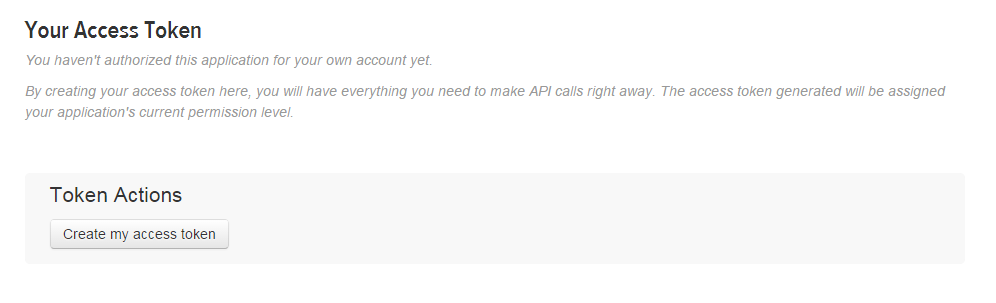
**STEP 16:**  
  
Your Application will be created:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/20.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/20.png)

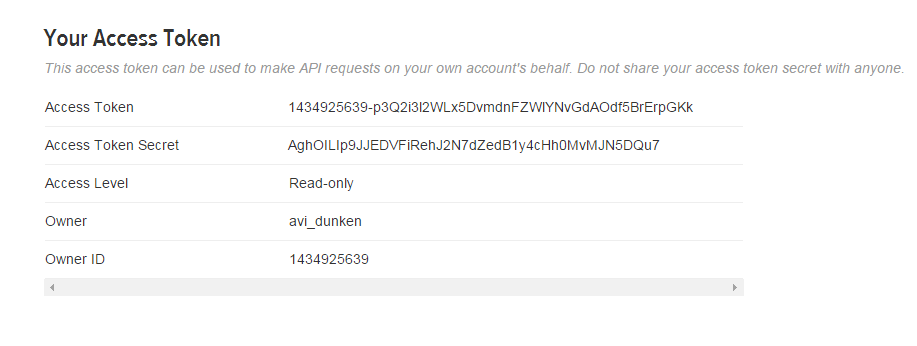
**STEP 17:**  
  
Click on Keys and Access Tokens, you will get Consumer Key and Consumer Secret.

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/211.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/211.png)

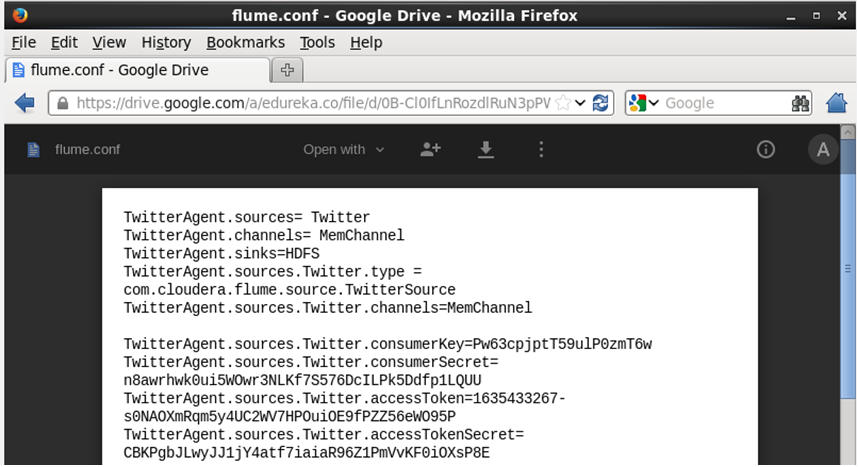
**STEP 18:**  
  
Scroll down and Click on Create my access token:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/221.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/221.png)

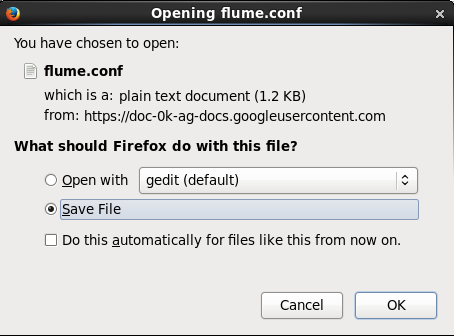
Your Access token got created:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/23.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/23.png)

**Consumer Key (API Key)** 4AtbrP50QnfyXE2NlYwROBpTm  
**Consumer Secret (API Secret)** jUpeHEZr5Df4q3dzhT2C0aR2N2vBidmV6SNlEELTBnWBMGAwp3  
**Access Token** 1434925639-p3Q2i3l2WLx5DvmdnFZWlYNvGdAOdf5BrErpGKk  
**Access Token Secret** AghOILIp9JJEDVFiRehJ2N7dZedB1y4cHh0MvMJN5DQu7  
  
**STEP 19:**  
  
Use below link to download flume.conf file  
https://drive.google.com/file/d/0B-Cl0IfLnRozdlRuN3pPWEJ1RHc/view?usp=sharing

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/24.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/24.png)

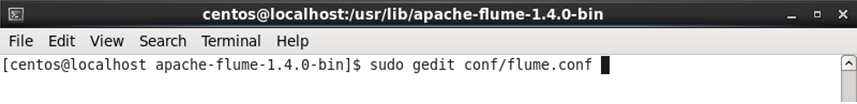
Save the file.

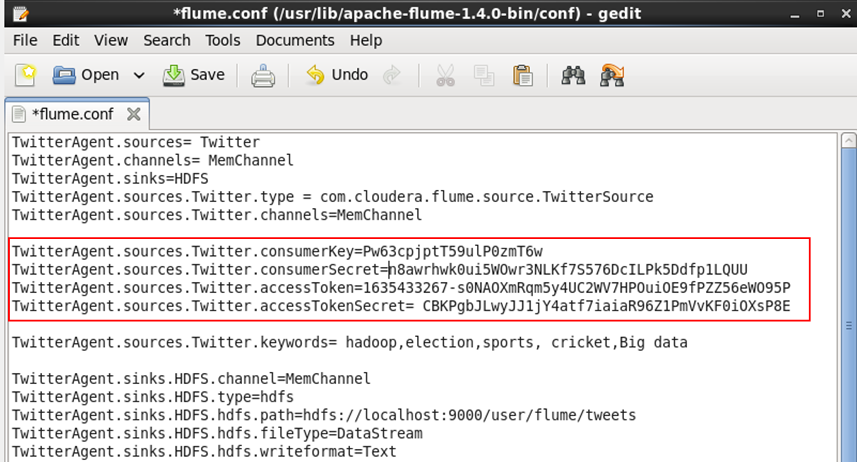
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/25.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/25.png)

**STEP 20:**  
  
Put the flume.conf in the conf directory of apache-flume-1.4.0-bin  
**Command:** sudo cp /home/centos/Downloads/flume.conf /usr/lib/apache-flume-1.4.0-bin/conf/

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/26.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/26.png)

**STEP 21:**  
  
Edit flume.conf  
  
**Command:** sudo gedit conf/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.

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/27.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/27.png)

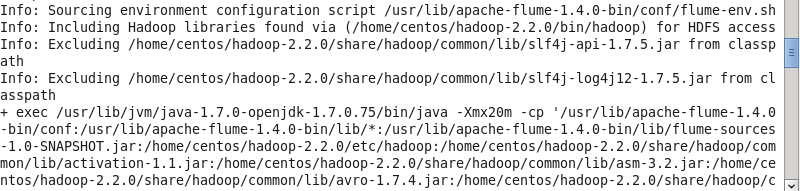
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/28.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/28.png)

**STEP 22:**  
  
Change permissions for flume directory.  
  
Command: sudo chmod -R 755 /usr/lib/apache-flume-1.4.0-bin/

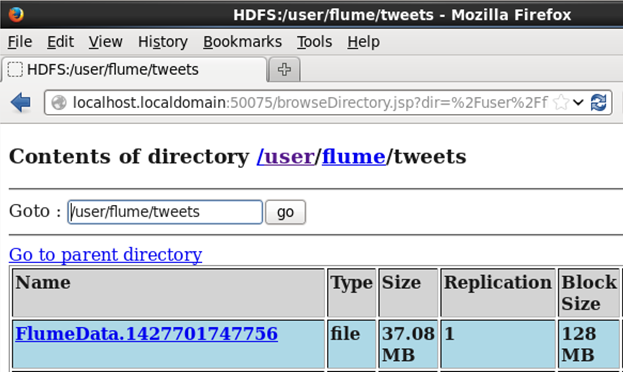
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/29.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/29.png)

**STEP 23:**  
  
Start fetching the data from twitter:  
  
Command: ./bin/flume-ng agent -n TwitterAgent -c conf -f /usr/lib/apache-flume-1.4.0-bin/conf/flume.conf

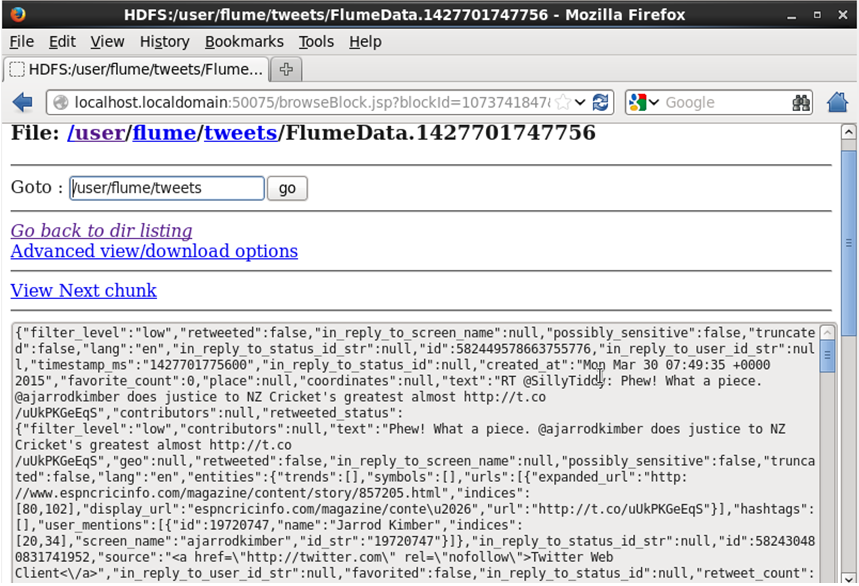
[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/30.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/30.png)

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/311.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/311.png)

Now wait for 20-30 seconds and let flume stream the data on HDFS, after that press ctrl + c to break the command and stop the streaming. (Since you are stopping the process, you may get few exceptions, ignore it)  
  
**STEP 24:**  
  
Open the Mozilla browser in your VM, and go to /user/flume/tweets in HDFS  
  
Click on FlumeData file which got created:

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/32.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/32.png)

If you can see data similar as shown in below snapshot, then the unstructured data has been streamed from twitter on to HDFS successfully. Now you can do analytics on this twitter data using Hive.

[[](http://www.eduonix.com/blog/wp-content/uploads/2015/06/33.png)](http://www.eduonix.com/blog/wp-content/uploads/2015/06/33.png)