Session 12:
Oozie and Flume
Assignment 1

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

To stream data to our database from twitter we should have the following pre-requisites.

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:

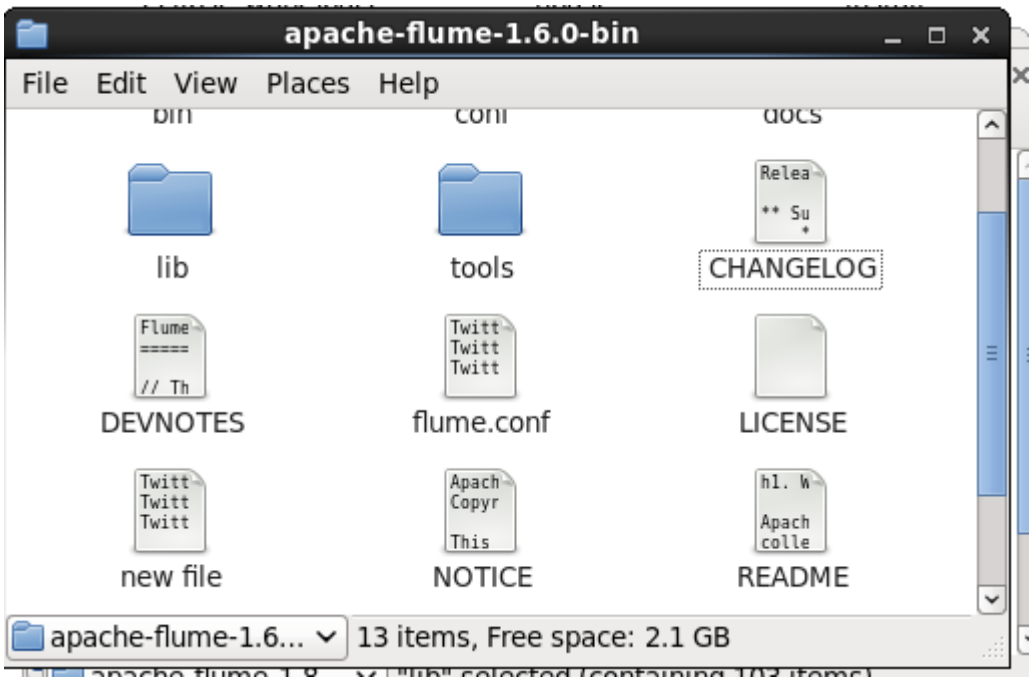
Copy the consumer key and the consumer secret code.

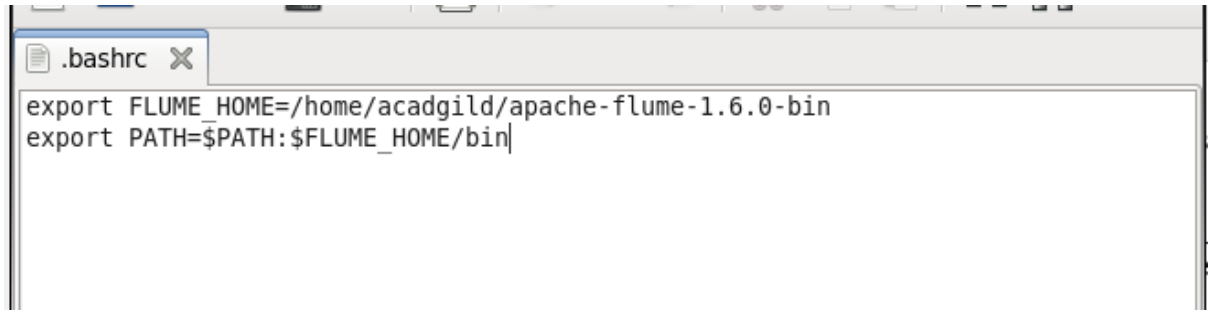
Step 7:

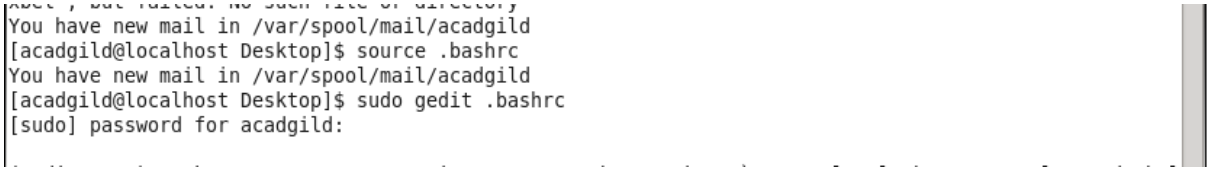
Scroll down further and select the 'create my access token' button.

Step 8:
Copy the Access Token and Access token Secret code.

Step 9: Go to flume directory and update the path of extracted flume directory in the .bashrc file as mentioned in the below image.

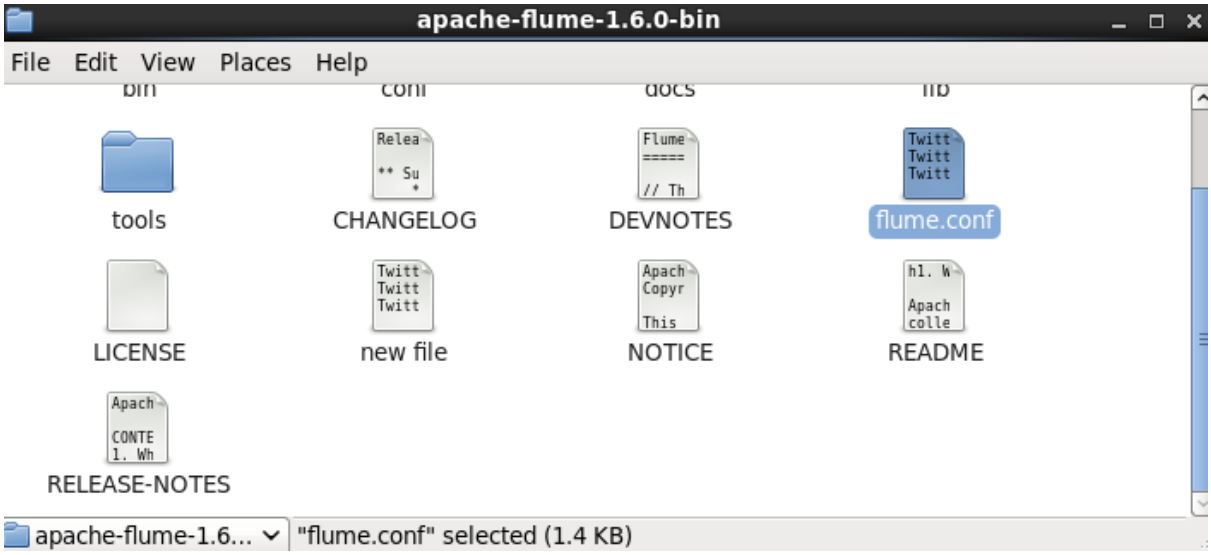






After setting the path of flume directory, save and close the .bashrc file. And then in the terminal type the below command to update the .bashrc file.

Step 10:
Create a new file inside the conf directory inside the Flume-extracted directory



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 11:
Copy the Flume configuration code from the below link and paste it in the newly created file.
https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk.

acadgild.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=uX0TWqkx0okYEjjqLzxIx6mD6
TwitterAgent.sources.Twitter.consumerSecret=rzHIs3TMJnADbZNvdGU7LQUo0kPxPISq3RGSLfqcBip39X5END
TwitterAgent.sources.Twitter.accessToken=559516596-yDA9xqOljo4CV32wSnqsx2BXh4RBIRKFxZGSZrPC
TwitterAgent.sources.Twitter.accessTokenSecret=zDxePILZitS5tIWBhre0GWqps0FIj9OadX8RZb6w8ZCwz
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

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 12:

Change the twitter api keys with the keys generated as shown in the step no 6 and step number 8.

Step 13:

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.

Step 14:

Open a new terminal and start all the Hadoop daemons, before running the flume command to fetch the twitter data.

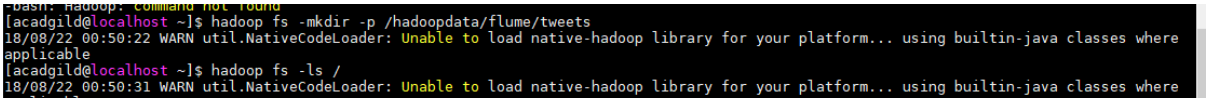
Use the ‘jps’ command to see the running Hadoop daemons.

```
[acadgild@localhost ~]$ jps
2992 NameNode
3456 ResourceManager
5619 Jps
3093 DataNode
3559 NodeManager
3279 SecondaryNameNode
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Step 15:

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

hHadoop fs –mkdir –p /hadoopdata/flume/tweets

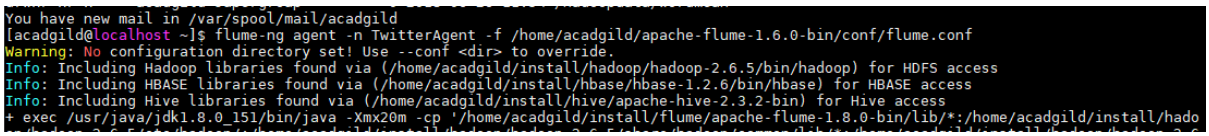


Step 16:

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/apache-flume-1.6.0-bin/conf/flume.conf



Once, the tweet data started streaming it into the given HDFS path we can use 'Ctrl+c' command to stop the streaming process.

Step 17:

To check the contents of the tweet data we can use the following command:

hadoop fs –ls / hadoopdata/flume/tweets

Step 18:

We can use the *'cat'* command to display the tweet data inside the / **hadoopdata/flume/tweets** /FlumeData.145* path.

hadoop fs –cat / hadoopdata/flume/tweets/<flumeData file name>