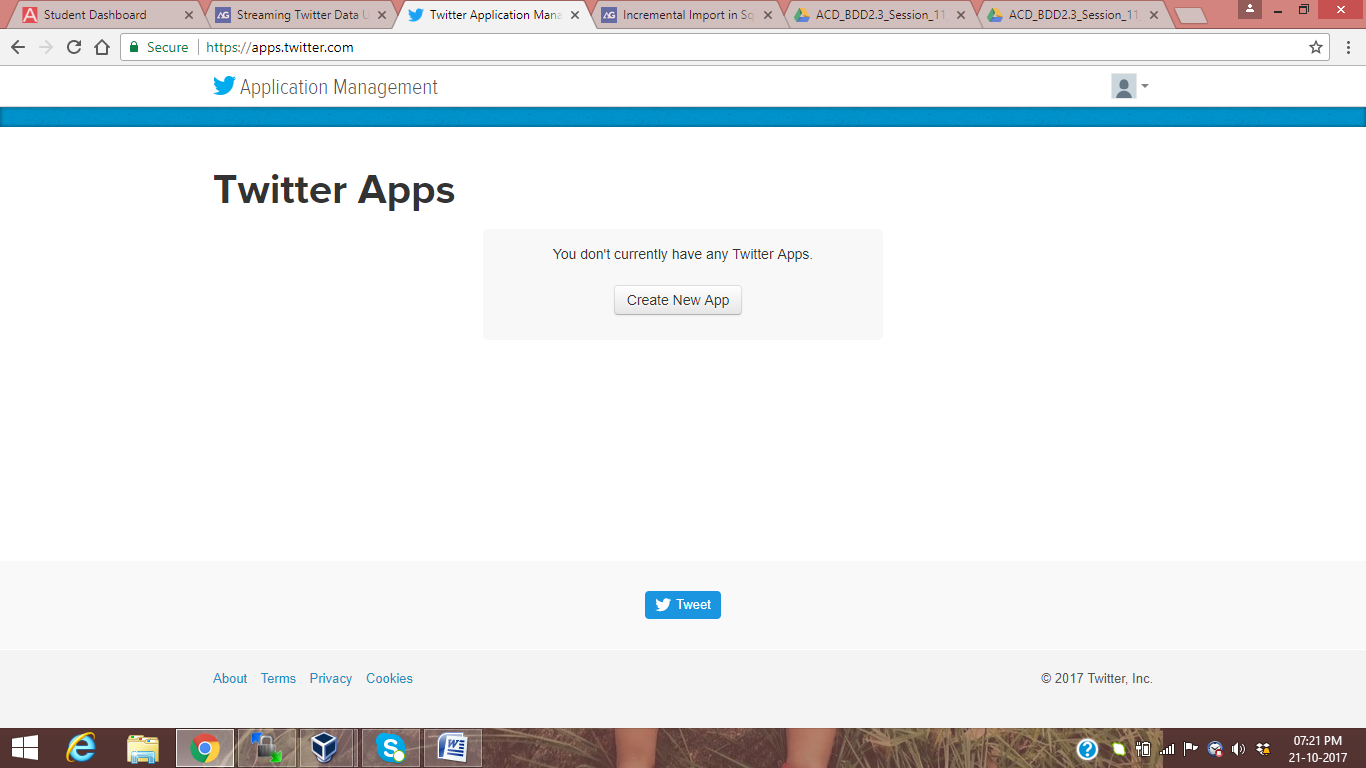
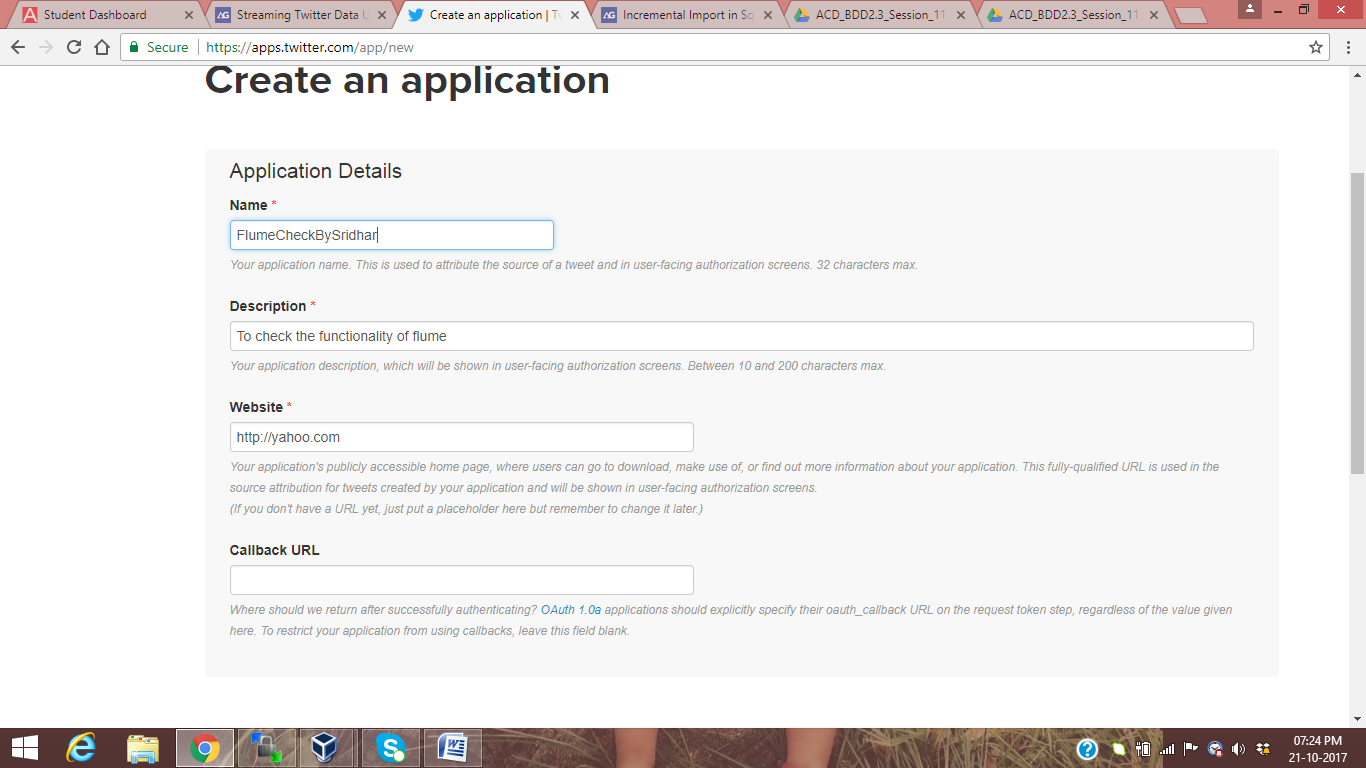
# Problem Statement

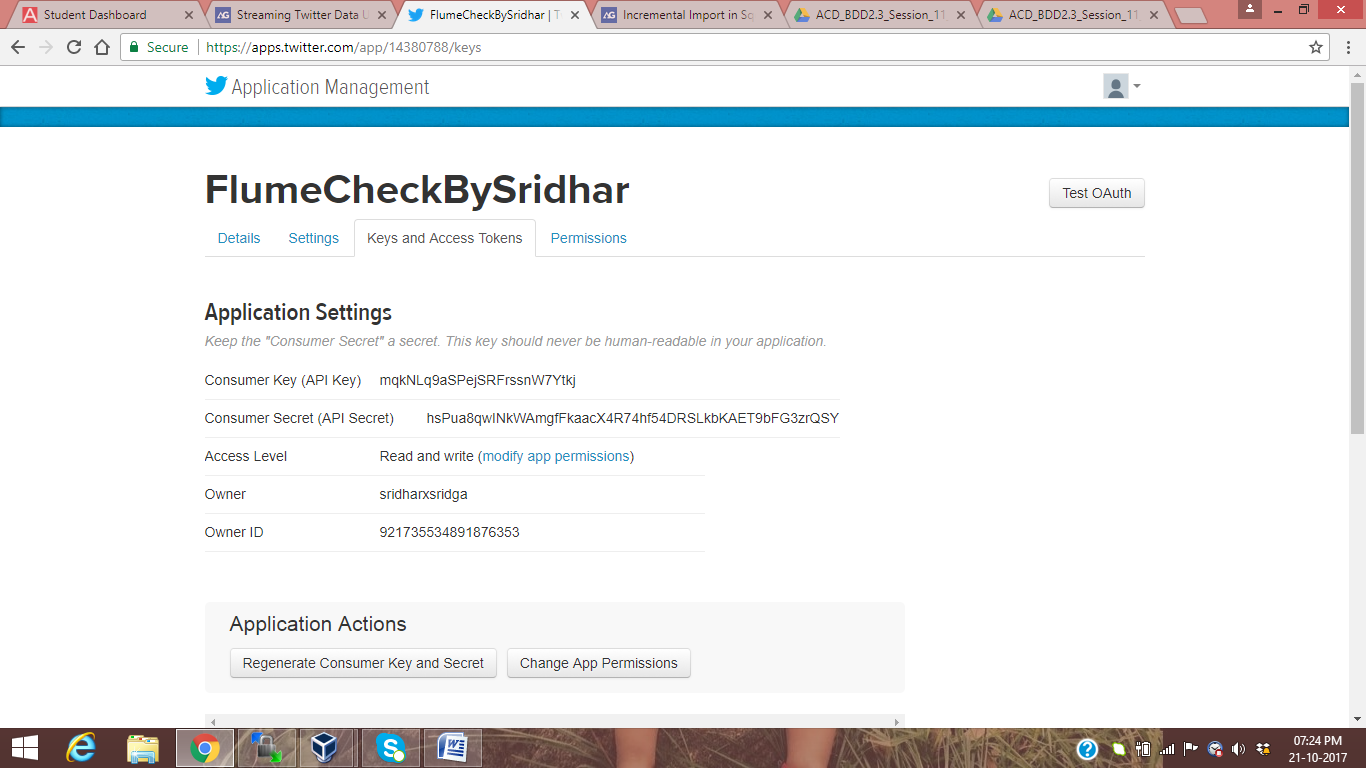
## Flume agent that streams data from Twitter and stores in the HDFS

* **First to retrieve the source data from Twitter we will create an app**

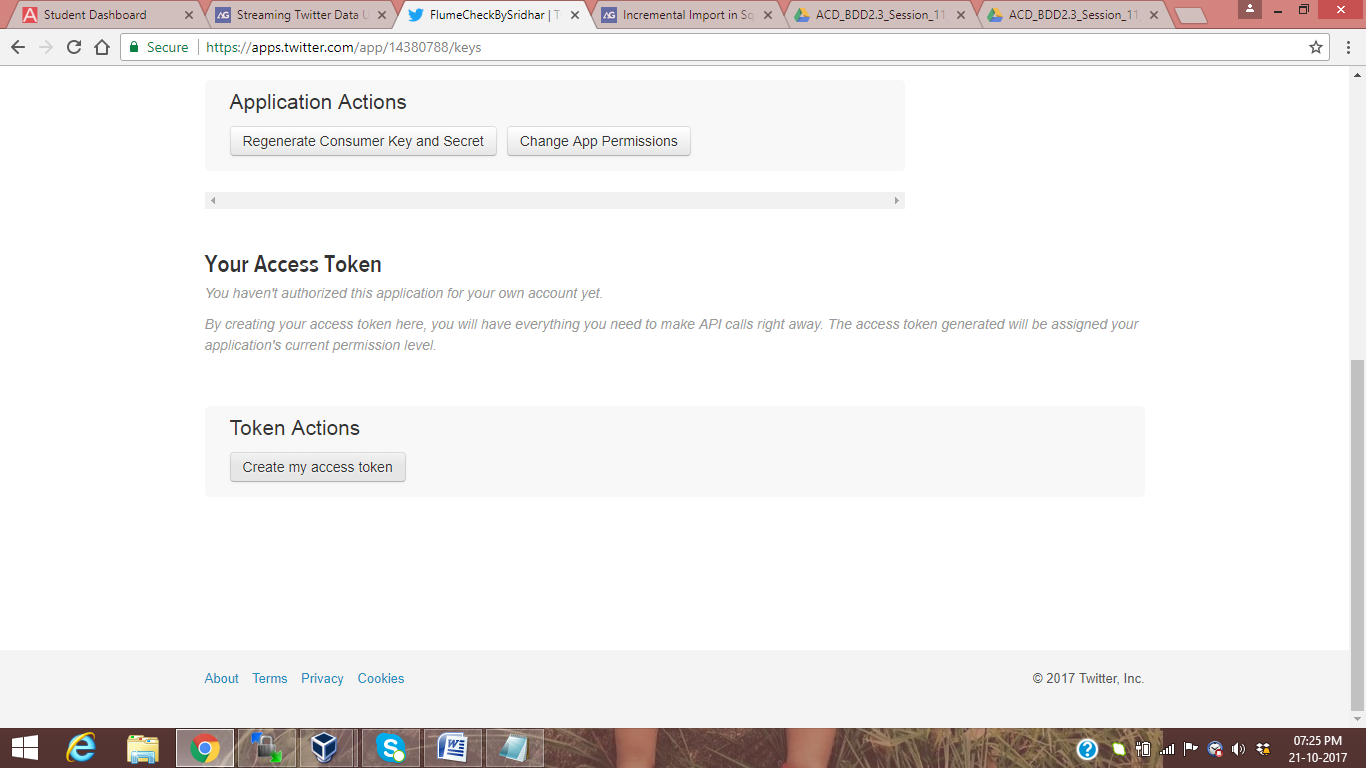
<https://apps.twitter.com/>

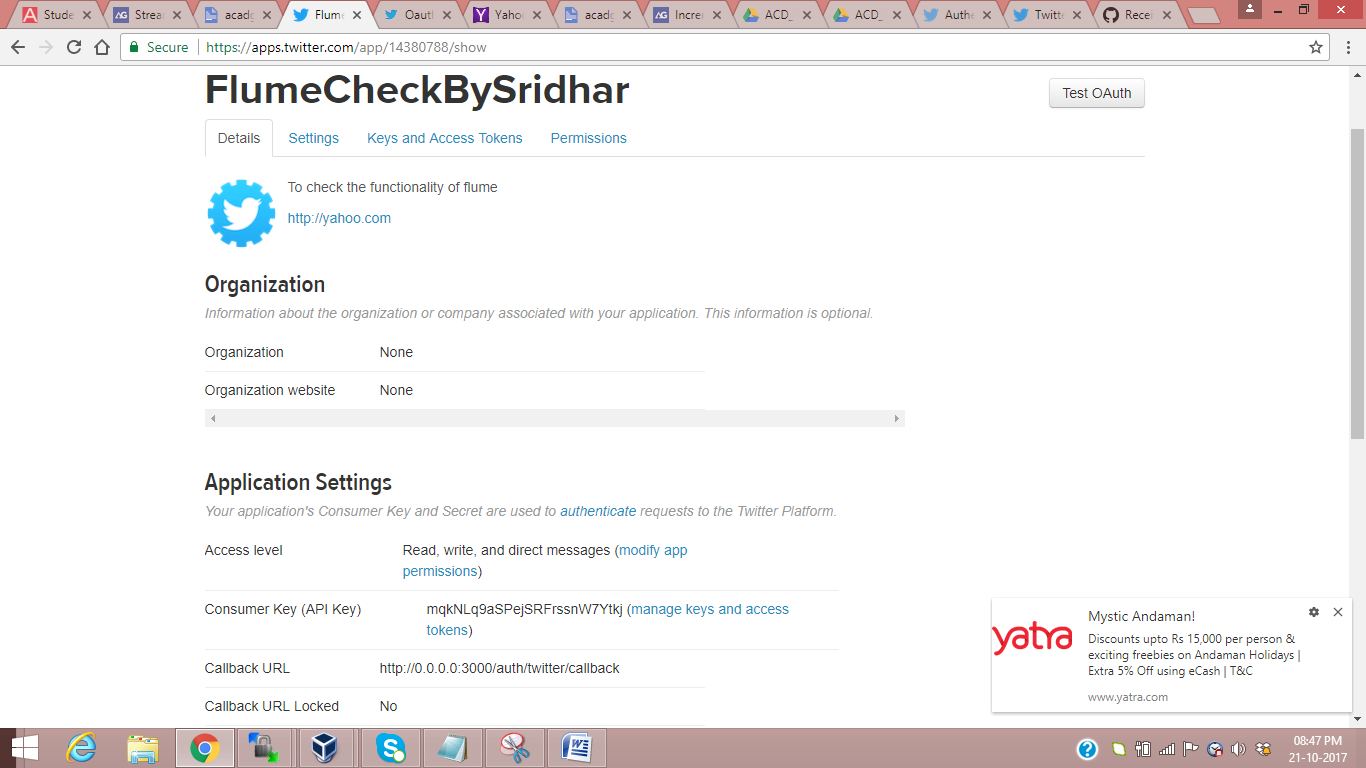


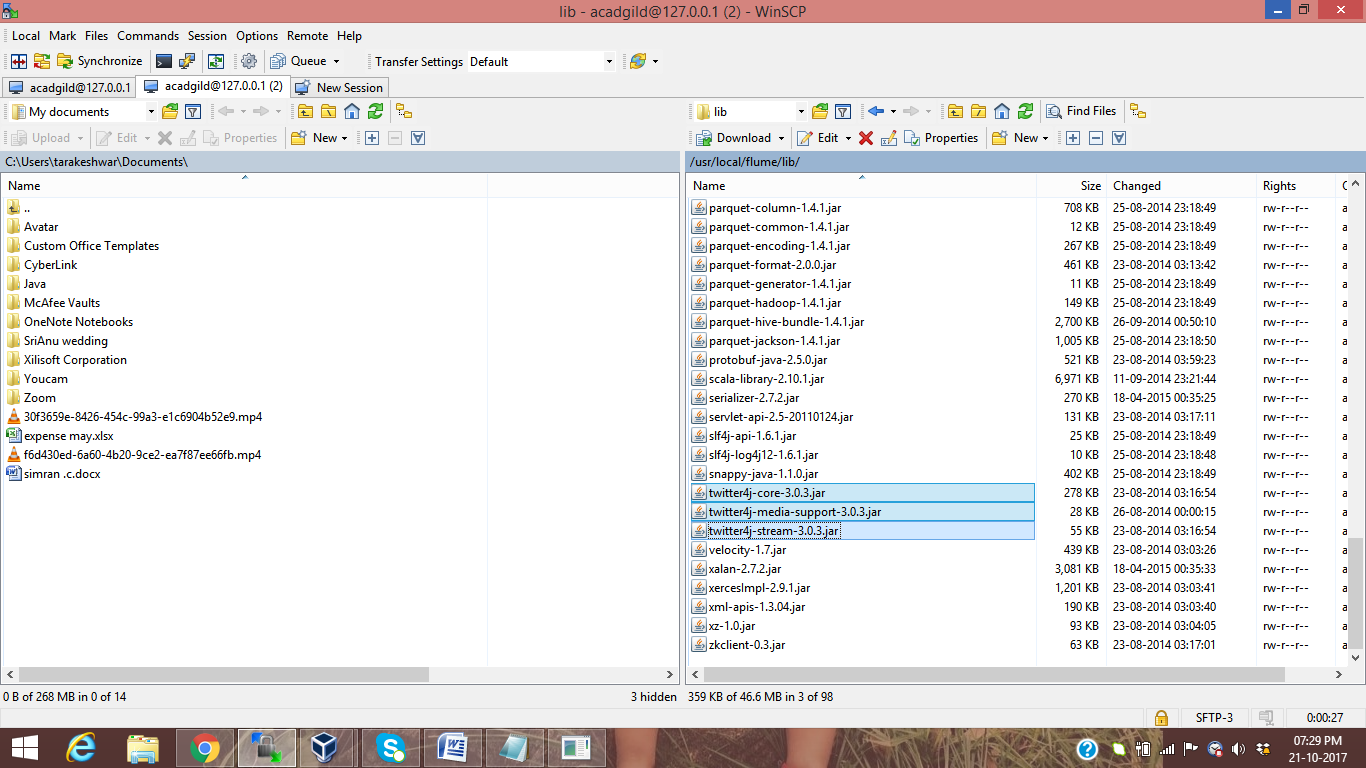




* **Copy Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret key which would be used for authentication to connect to Twitter App**





* **Make Sure that we have twitter related jars in the lib directory**
* **Place Twitter\_flume.conf configuration file in local file system** /home/acadgild/sridhar\_hadoop/flumeTest

## Twitter\_flume.conf

**#Specifying source name**

TwitterAgent.sources = Twitter

**#Specifying channel name**

TwitterAgent.channels = MemChannel

**#Specifying sink name**

TwitterAgent.sinks = HDFS

**# Describing/Configuring the source**

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

**# Change the twitter api keys with the keys generated**

TwitterAgent.sources.Twitter.consumerKey=mqkNLq9aSPejSRFrssnW7Ytk

TwitterAgent.sources.Twitter.consumerSecret=hsPua8qwINkWAmgfFkaacX4R74hf54DRSLkbKAET9bFG3zrQSY

TwitterAgent.sources.Twitter.accessToken=921735534891876353-CuMTlRN7r2llcpRXQAlAn4yM7cIYbQj

TwitterAgent.sources.Twitter.accessTokenSecret=lUU99XtzSQgvuS7aRbmUo9clsEDtIuBEE1OysABOGPj1R

**#tweet data to be collected from the twitter application**

TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql

**# Configuring the sink**

TwitterAgent.sinks.HDFS.channel=MemChannel

TwitterAgent.sinks.HDFS.type=hdfs

**# HDFS location where the data will be generated**

TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/flume\_output

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

**# Channel being used is memory , can be changed to file as well depending upon performance requirement**

TwitterAgent.channels.MemChannel.type=memory

TwitterAgent.channels.MemChannel.capacity=10000

TwitterAgent.channels.MemChannel.transactionCapacity=1000

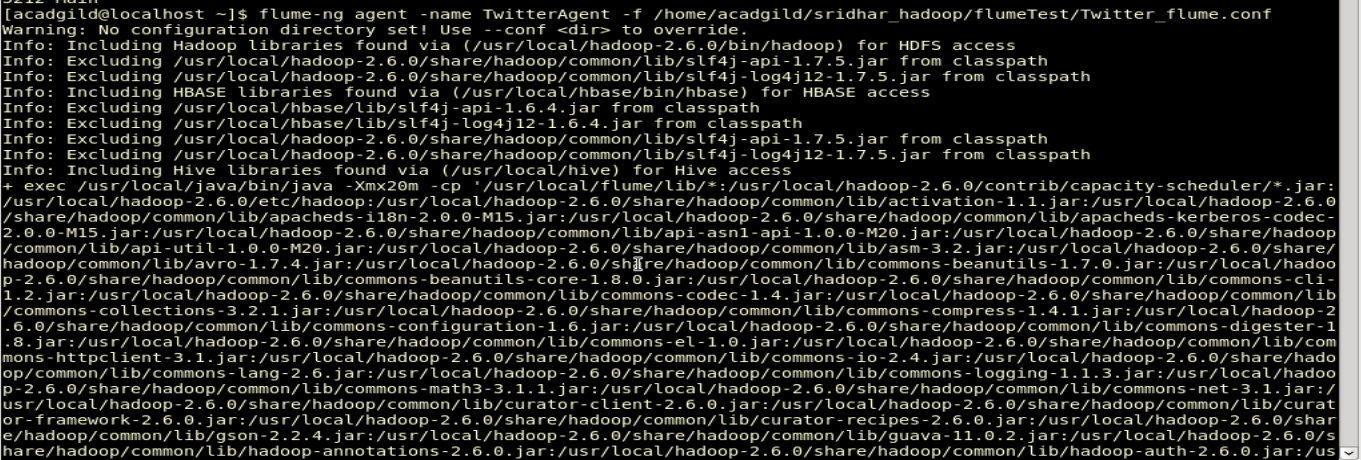
**# Connecting channel to source and sink**

TwitterAgent.sources.Twitter.channels = MemChannel

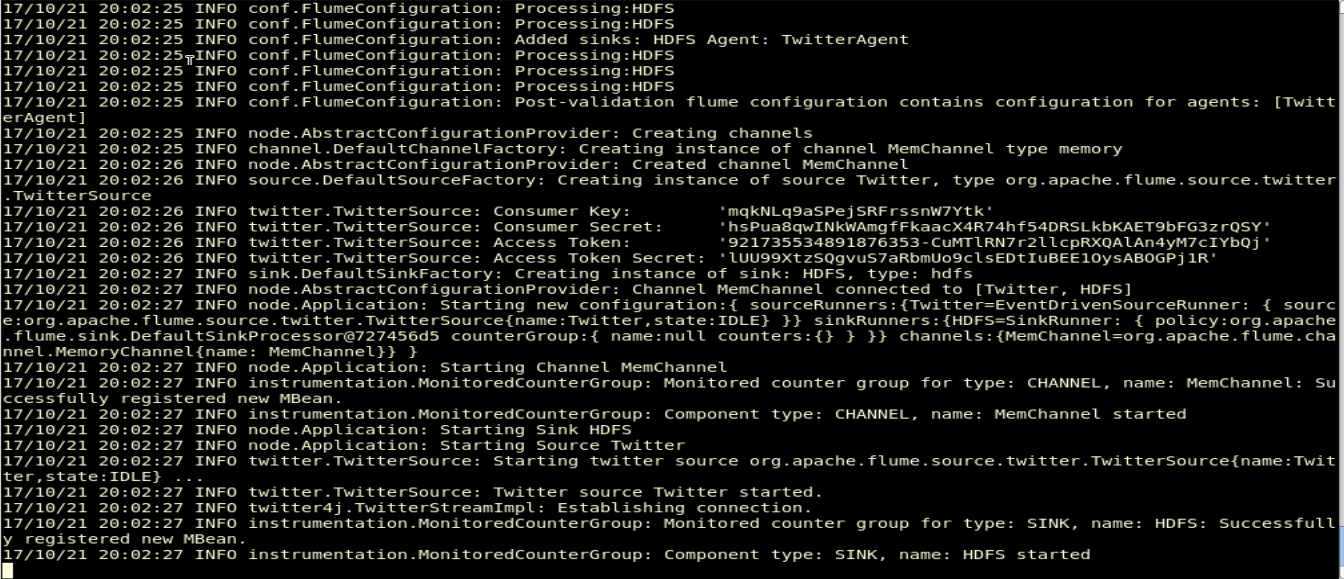
TwitterAgent.sinks.HDFS.channel = MemChannel

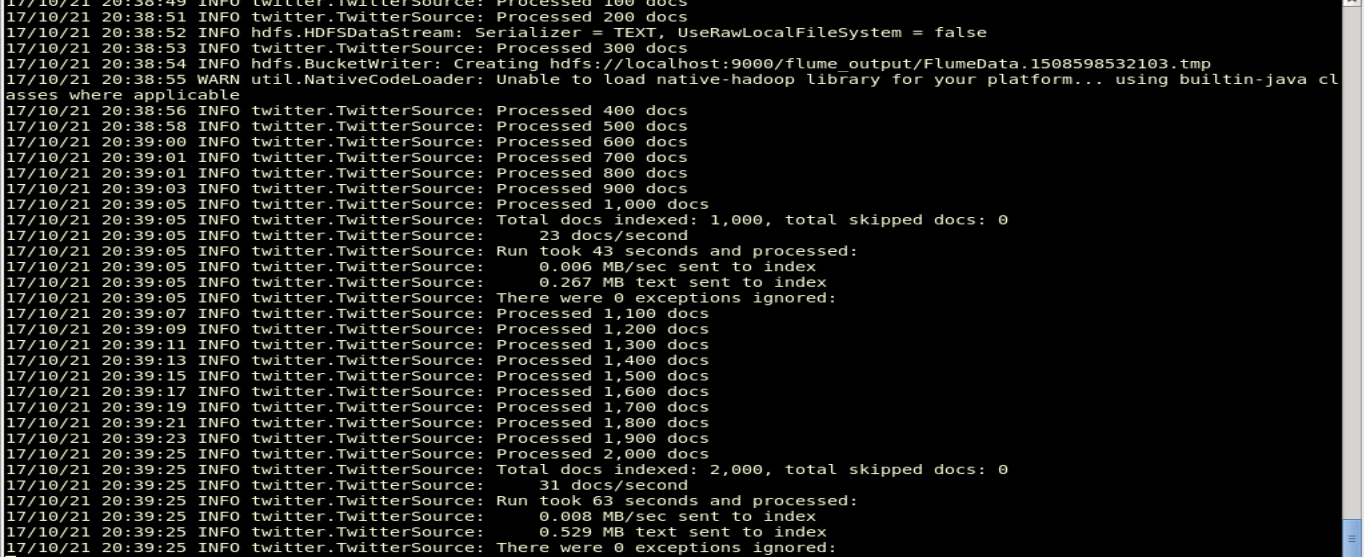
# Run the flume job using the command

## flume-ng agent -name TwitterAgent -f /home/acadgild/sridhar\_hadoop/flumeTest/ Twitter\_flume.conf



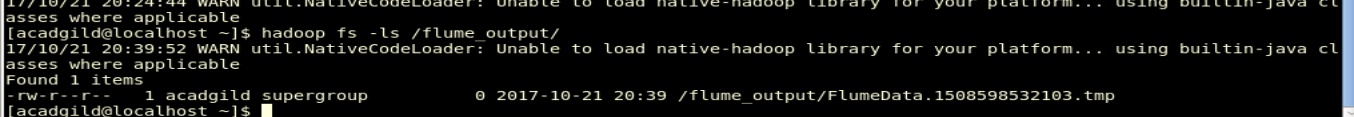
**Job has started:**





**Output from Twitter has been collected to source in hdfs location , use the below command to check**

hadoop fs -ls /flume\_output/



**Check the content using the below command**

hadoop fs -cat /flume\_output/FlumeData.1508598532103.tmp

