**Name** : GOKULAKANNAN A

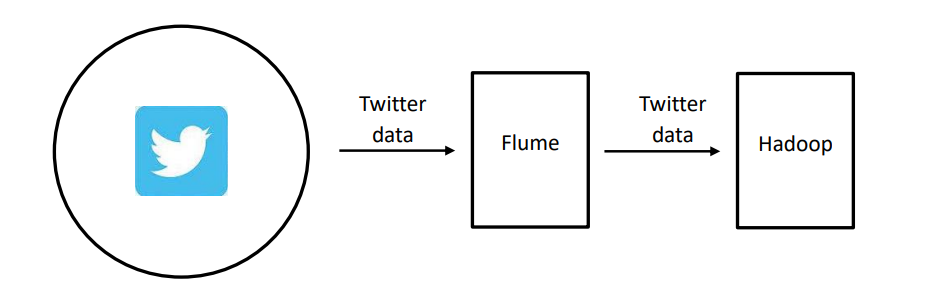
**RegNo :** 913020106002

**Department :** ELECTRONICS AND COMMUNICATION ENGINEERING

**Project** : BIG DATA ANALYSIS

**Problem Statement:**

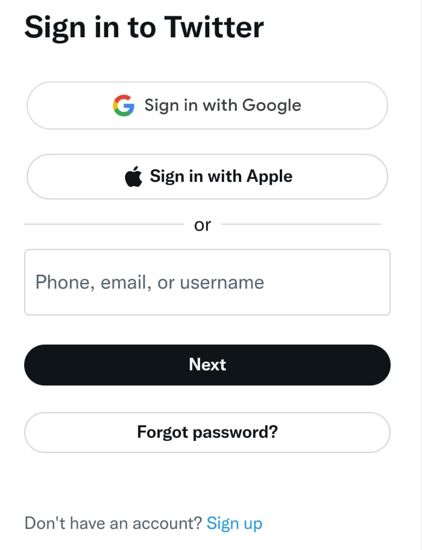
Real-time Data Collection Imagine you are a Big Data Engineer, and you need to fetch Twitter data into your Hadoop Cluster for doing some analyses to generate some business insights. The following figure illustrates a scenario where we need to ingest Twitter Data into the Hadoop clusters and then use the ingested data as required.



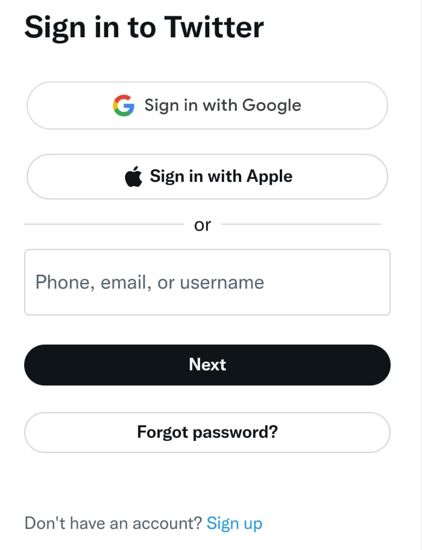
As a Big Data Engineer, your task is to ingest the Twitter Data into HDFS using Flume agent.

NOTE: Follow the following steps to get started with the project.

**Step 1:** Go to <https://developer.twitter.com/apps>



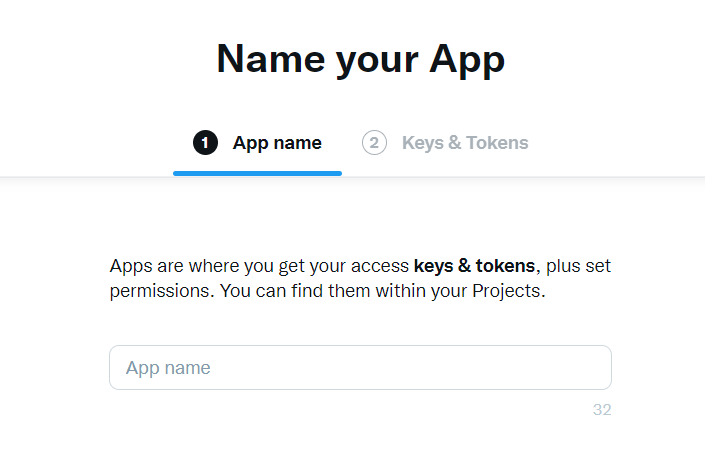
**Step 2:** Click Sign up if you don’t have account or click sign in if you have account



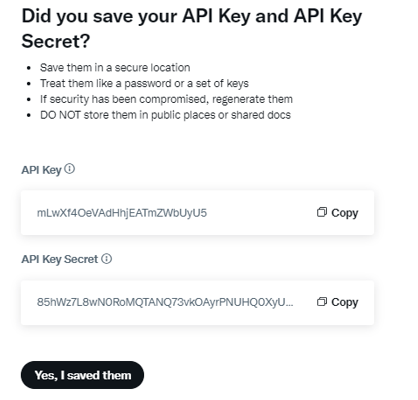
**Step 3:** After signing up it will go to twitter website. Again go to <https://developer.twitter.com/apps> and there you can find **CREATE APP** option



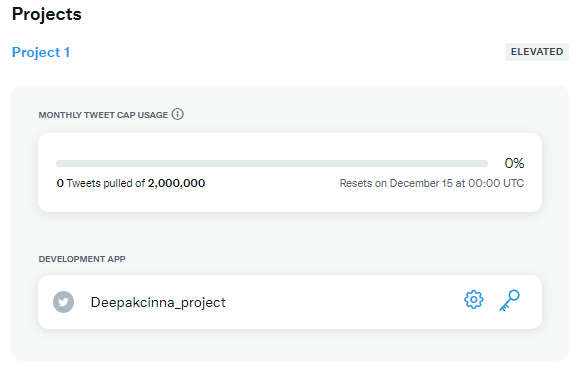
**Step 4:** You should name your App and go to keys and token section



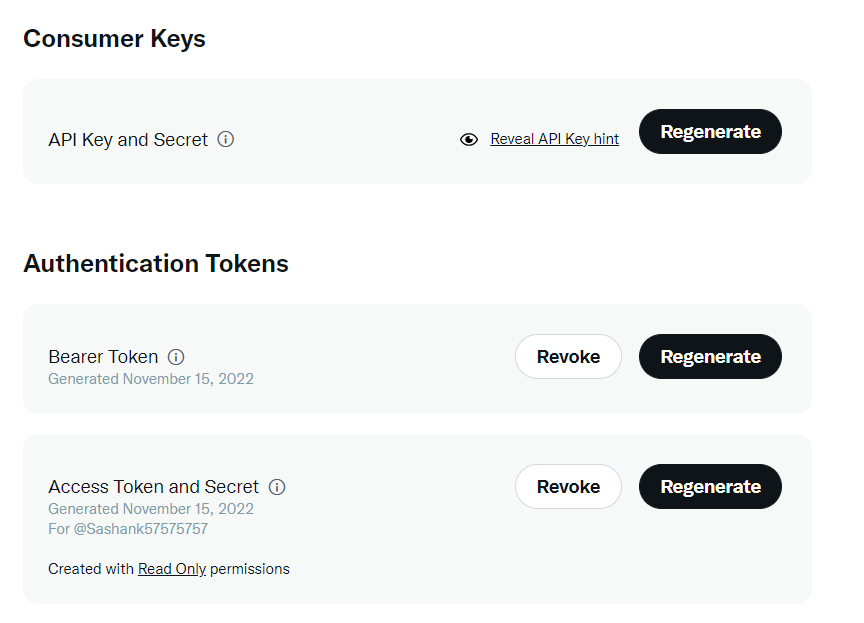
**Step 5:** Copy both API key and API key Secret somewhere safe.



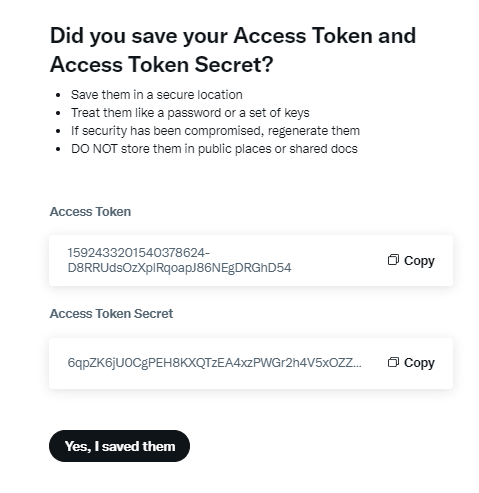
**Step 6:** After that click dashboard option that is present and it will take you to dashboard



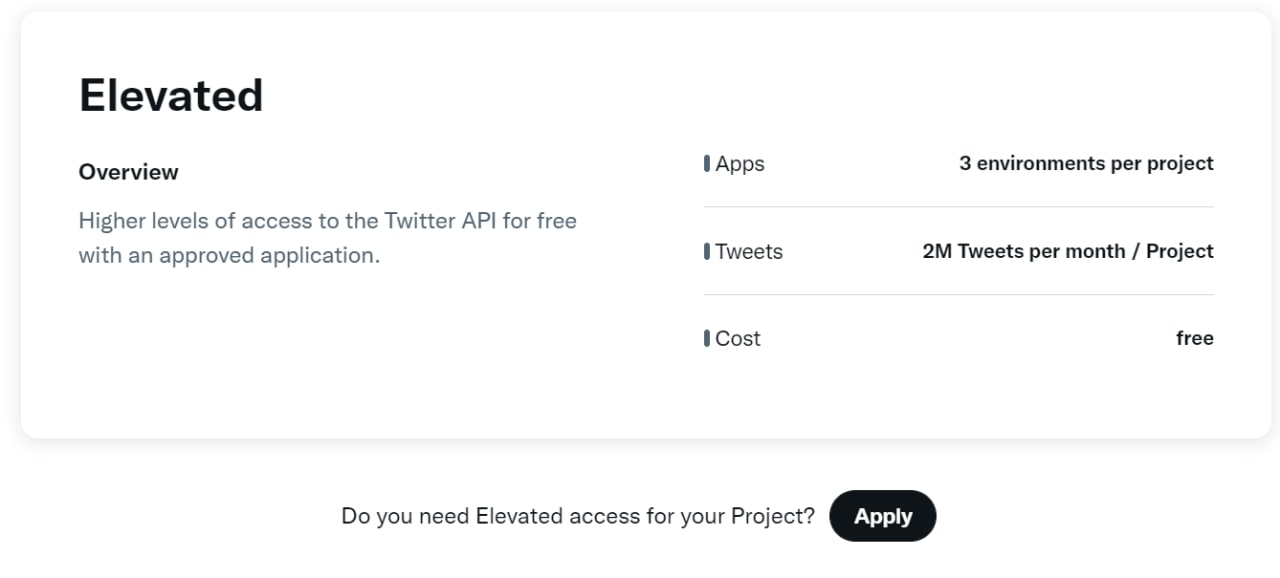
**Step 7:** Click key icon present on the APP



**Step 8:** Click Regenerate option in Access Token and Secret. Save Access Token and Access Token Secret Keys safe.



**Step 9:** We need to get elevated access for our project. So go to <https://developer.twitter.com/en/portal/products/elevated> and complete the process.



Click Apply and you can fill the required details to get elevated access.

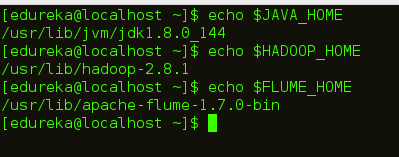
**Step 10:** Open Virtual box and open Edureka VM and click the terminal. In the terminal check whether java,Hadoop,flume installed or not.

Use:

Java: **echo $JAVA\_HOME**

Hadoop: **echo $HADOOP\_HOME**

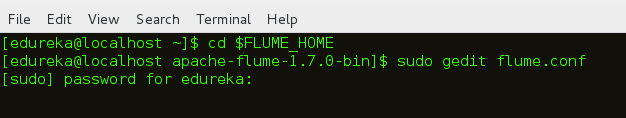
Flume: **echo $FLUME\_HOME**



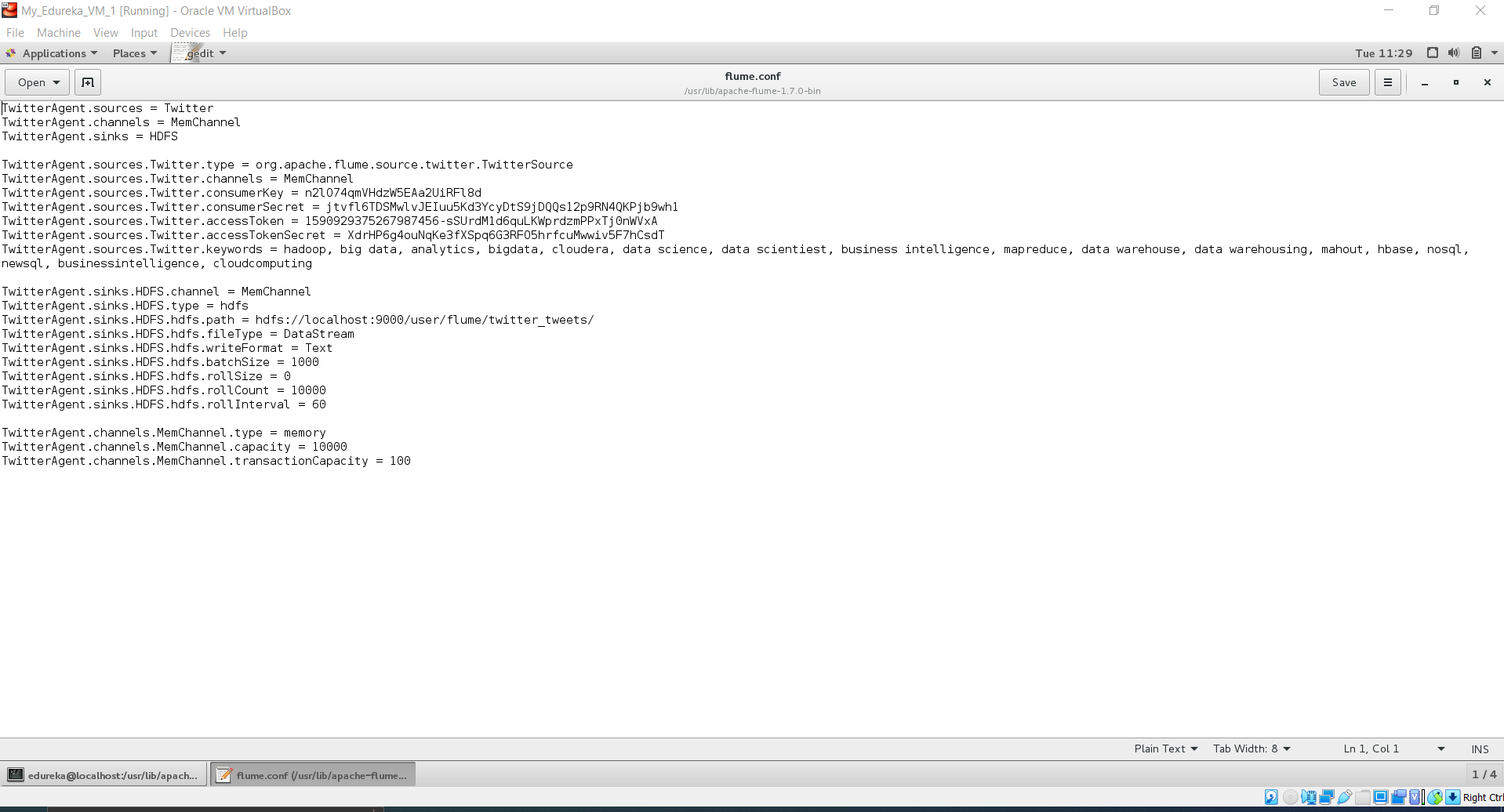
**Step 11:** Open flume folder by using **cd $FLUME\_HOME** and the type the

command: **sudo gedit flume.conf**

enter the password as **edureka**.



**Step 12:** It will open a text editor with name flume.conf. Type the below code in the text editor



TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

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

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = mLwXf4OeVAdHhjEATmZWbUyU5

TwitterAgent.sources.Twitter.consumerSecret = 85hWz7L8wN0RoMQTANQ73vkOAyrPNUHQ0XyUo56WWUTb9yU2AG

TwitterAgent.sources.Twitter.accessToken = 1592433201540378624-D8RRUdsOzXplRqoapJ86NEgDRGhD54

TwitterAgent.sources.Twitter.accessTokenSecret = 6qpZK6jU0CgPEH8KXQTzEA4xzPWGr2h4V5xOZZjrlefpk

TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientiest, business intelligence, mapreduce, data warehouse, data warehousing, mahout, hbase, nosql, newsql, businessintelligence, cloudcomputing

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/user/flume/twitter/

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 = 60

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

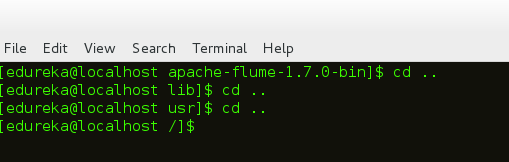
# After typing the above code save it by pressing Ctrl+S.

**Step 13:** Now go to Hadoop directory with command: **cd $HADOOP\_HOME**

Then go to sbin directory with command: **cd sbin** and run the command: **./start-all.sh**

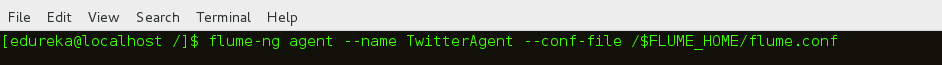


**Step 14:** In the terminal type command: **cd ..** until you get back to root folder(/).

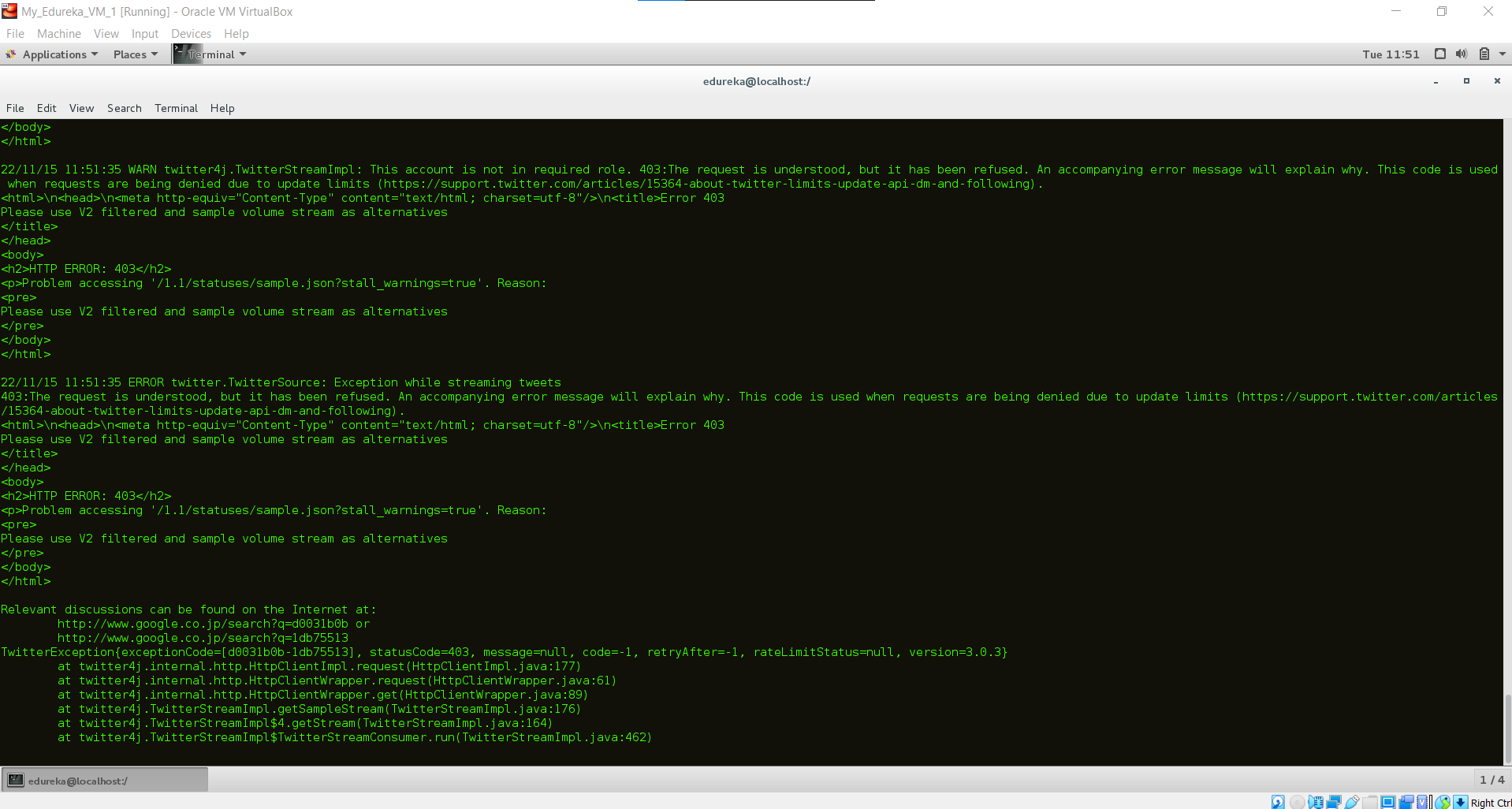


**Step 15:** After completion of step 14, use flume-ng agent to retrieve the data, to do that use the following command:

**flume-ng agent --name TwitterAgent --conf-file /$FLUME\_HOME/flume.conf**

****

**Output:**

****