Name : GOKULAKANNAN A

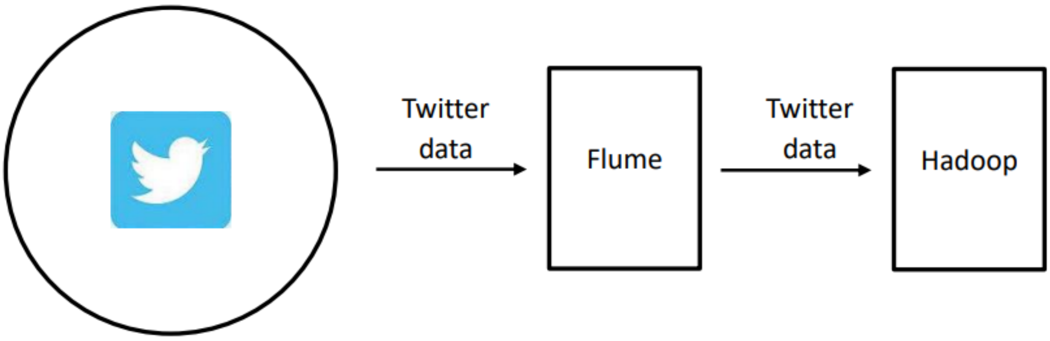
RegNo : 913020106002

Department : ELECTRONICS AND COMMUNICATION ENGINEERING

Project : BIG DATA ANALYSIS

Problem Statement:

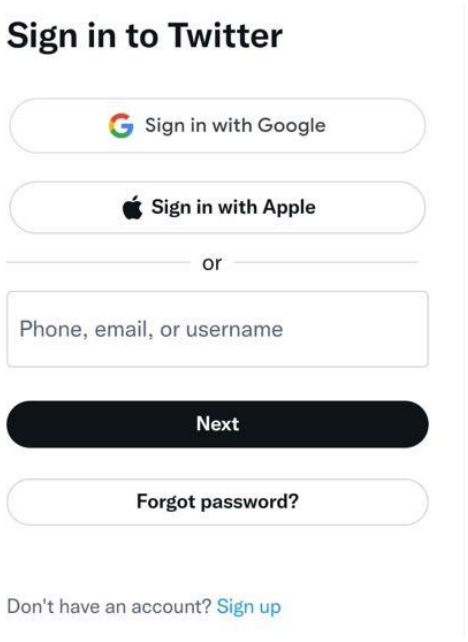
Real-time Data Collection Imagine you are a Big Data Engineer, and you need to fetch Twitter data into your Hadoop Cluster for doing some analyses to generate some business insights. The following figure illustrates a scenario where we need to ingest Twitter Data into the Hadoop clusters and then use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

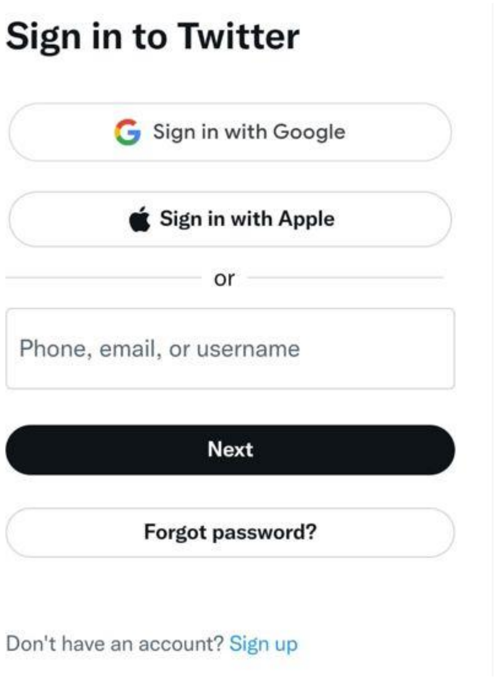
Step 1: Go to https://developer.twitter.com/apps



Sign in to Twitter

Don't have an account? Sign up

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

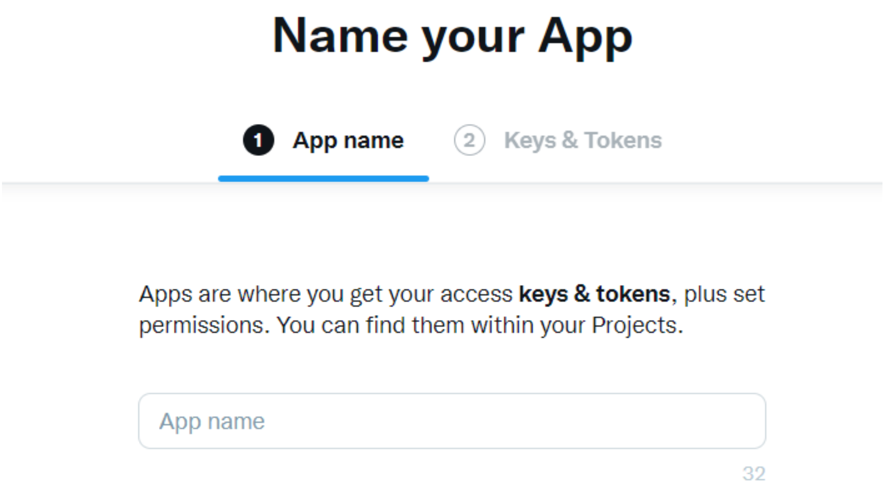


Sign in to Twitter

Don't have an account? Sign up

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

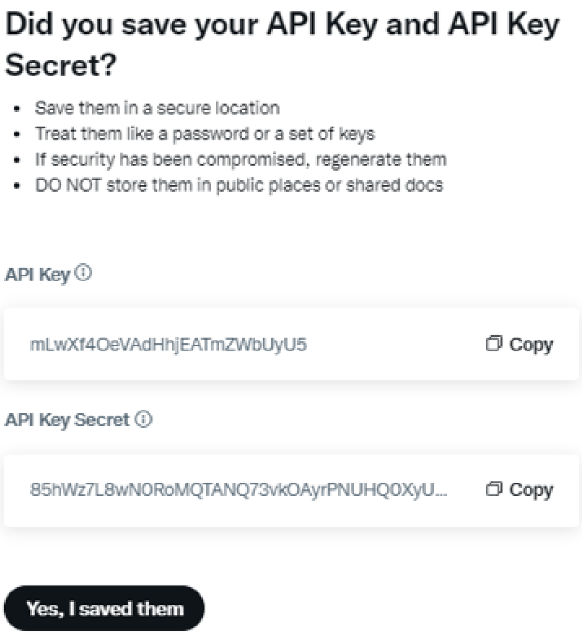


Name your App

Apps are where you get your access **keys & tokens**, plus set permissions. You can find them within your Projects.

App name

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



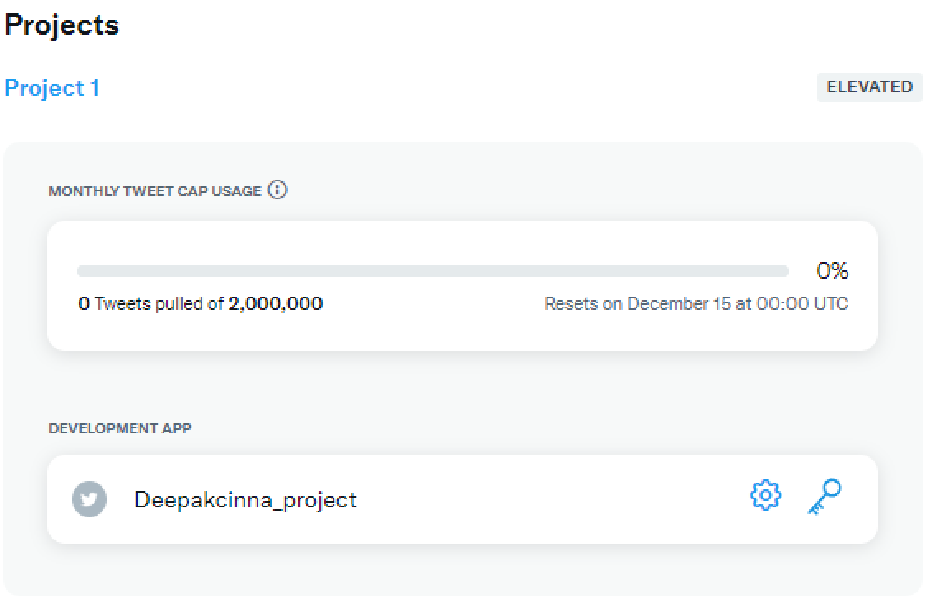
Did you save your API Key and API Key Secret?

- Save them in a secure location
- Treat them like a password or a set of keys
- If security has been compromised, regenerate them
- DO NOT store them in public places or shared docs

API Key: mLwXf4OeVAdHhjEATmZWbUyU5

API Key Secret: 85hWz7L8wN0RoMQTANQ73vkOAyrPNUHQ0XyU...

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



Projects

Project 1

ELEVATED

MONTHLY TWEET CAP USAGE

0%

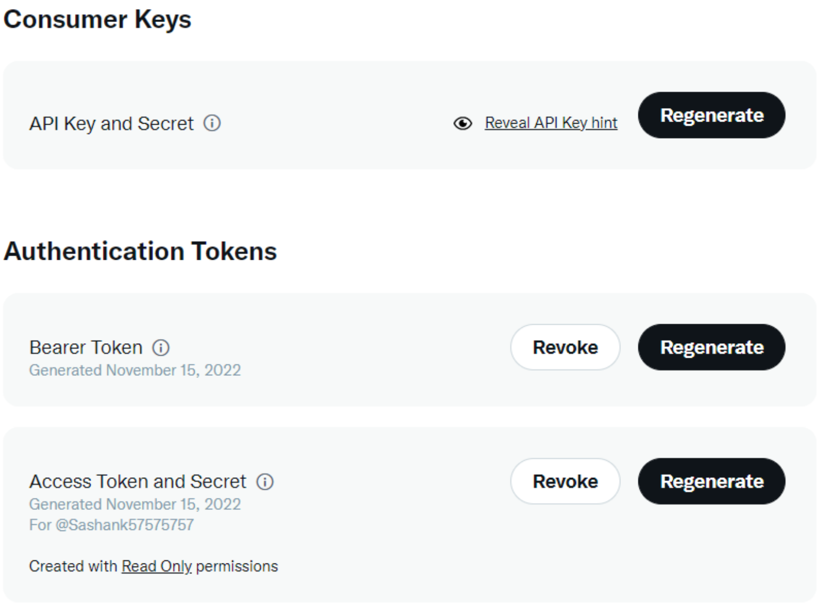
0 Tweets pulled of 2,000,000

Resets on December 15 at 00:00 UTC

DEVELOPMENT APP

Deepakcinna_project

Step 7: Click key icon present on the APP



Consumer Keys

API Key and Secret

Authentication Tokens

Bearer Token

Generated November 15, 2022

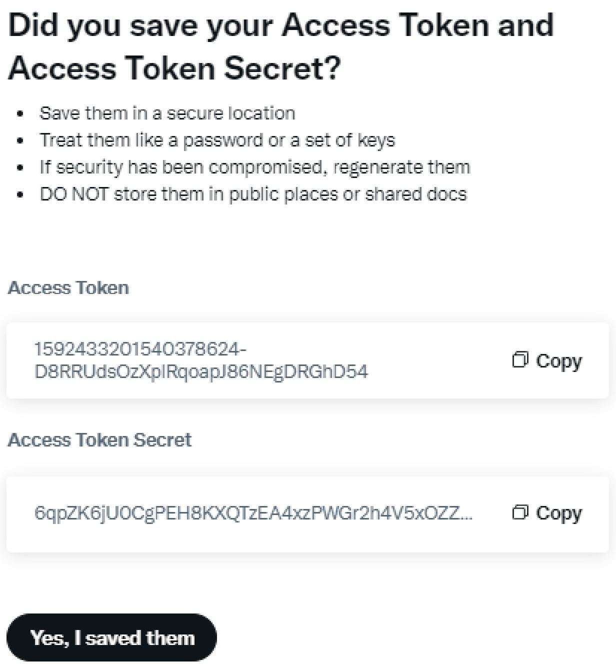
Access Token and Secret

Generated November 15, 2022

For @Sashank57575757

Created with Read Only permissions

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



Did you save your Access Token and Access Token Secret?

- Save them in a secure location
- Treat them like a password or a set of keys
- If security has been compromised, regenerate them
- DO NOT store them in public places or shared docs

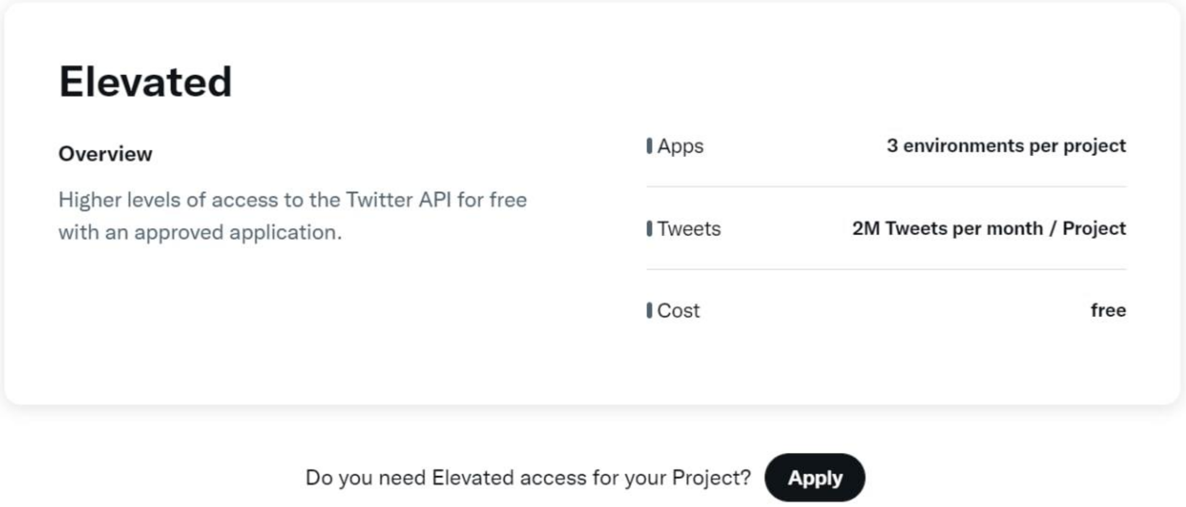
Access Token

1592433201540378624-D8RRUdsOzXplRqoapJ86NEgDRGhD54

Access Token Secret

6qpZK6jU0CgPEH8KXQTzEA4xzPWGr2h4V5xOZZ...

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.



Elevated

Overview

Higher levels of access to the Twitter API for free with an approved application.

Apps	3 environments per project
Tweets	2M Tweets per month / Project
Cost	free

Do you need Elevated access for your Project?

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

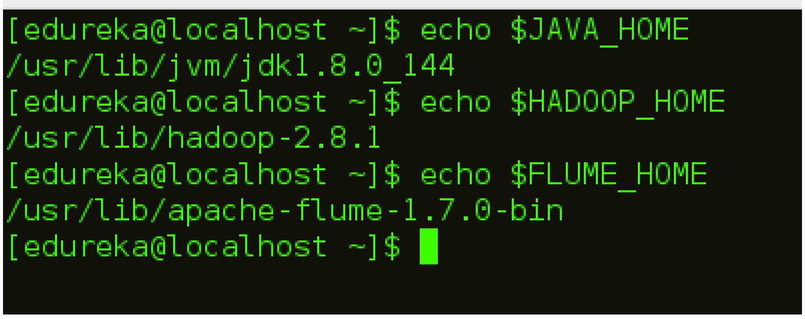
Step 10: Open Virtual box and open Edureka VM and click the terminal. In the terminal check whether java,Hadoop,flume installed or not.

Use:

Java: **echo $JAVA_HOME**

Hadoop: **echo $HADOOP_HOME**

Flume: **echo $FLUME_HOME**

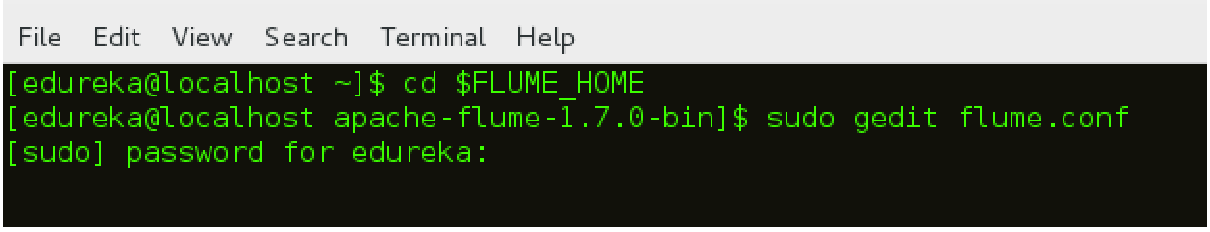


```
[edureka@localhost ~]$ echo $JAVA_HOME
/usr/lib/jvm/jdk1.8.0_144
[edureka@localhost ~]$ echo $HADOOP_HOME
/usr/lib/hadoop-2.8.1
[edureka@localhost ~]$ echo $FLUME_HOME
/usr/lib/apache-flume-1.7.0-bin
[edureka@localhost ~]$ 
```

Step 11: Open flume folder by using **cd $FLUME_HOME** and the type the

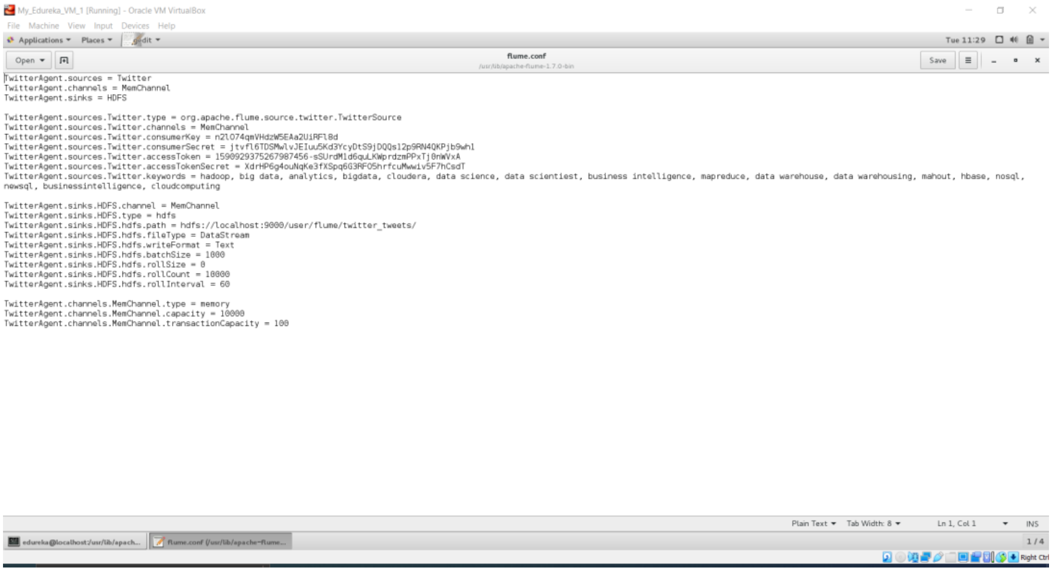
command: **sudo gedit flume.conf**

enter the password as **edureka**.



```
[edureka@localhost ~]$ cd $FLUME_HOME
[edureka@localhost apache-flume-1.7.0-bin]$ sudo gedit flume.conf
[sudo] password for edureka:
```

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


My_Edureka_VM_1 [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Applications
Places
gedit
Tue 11:29
Open
flume.conf
/usr/lib/apache-flume-1.7.0-bin
Save
Plain Text
Tab Width: 8
Ln 1, Col 1
INS
edureka@localhost:/usr/lib/apach...
flume.conf (/usr/lib/apache-flume...
1 / 4
Right Ctrl

```
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 = n2lO74qmVHdzW5EAa2UiRFl8d
TwitterAgent.sources.Twitter.consumerSecret = jtvfl6TDSMwlvJEIuu5Kd3YcyDtS9jDQQsl2p9RN4QKPjb9whl
TwitterAgent.sources.Twitter.accessToken = 1590929375267987456-sSUrdM1d6quLKWprdzmPPxTj0nWVxA
TwitterAgent.sources.Twitter.accessTokenSecret = XdrHP6g4ouNqKe3fXSpq6G3RFO5hrfcuMwwiv5F7hCsdT
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_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 = 60

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

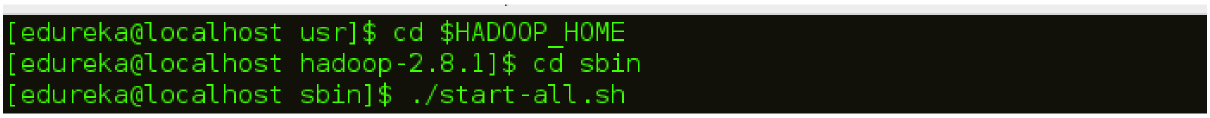
TwitterAgent.sources = Twitter

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.
```

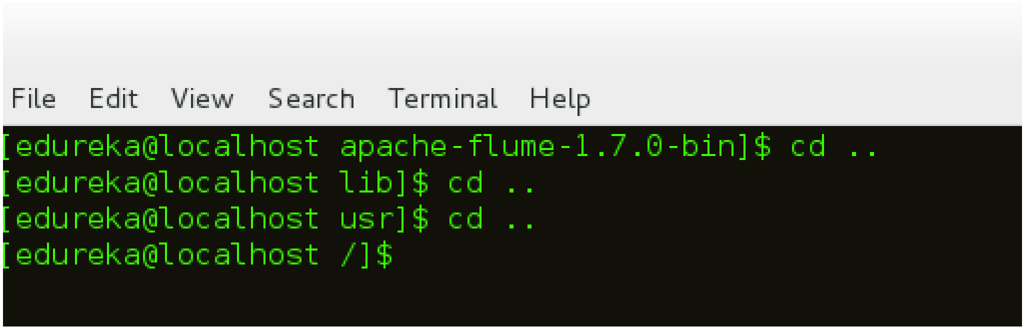
Step 13: Now go to Hadoop directory with command: **cd $HADOOP_HOME**

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



```
[edureka@localhost usr]$ cd $HADOOP_HOME
[edureka@localhost hadoop-2.8.1]$ cd sbin
[edureka@localhost sbin]$ ./start-all.sh
```

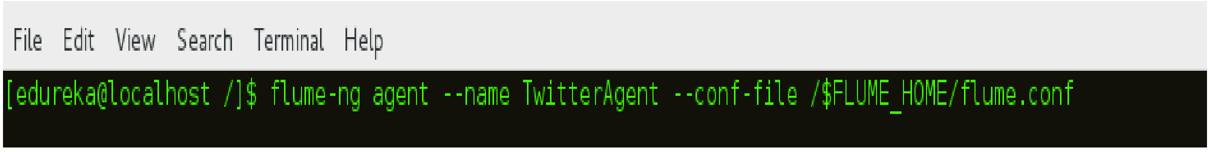
Step 14: In the terminal type command: **cd ..** until you get back to root folder(/).



```
[edureka@localhost apache-flume-1.7.0-bin]$ cd ..
[edureka@localhost lib]$ cd ..
[edureka@localhost usr]$ cd ..
[edureka@localhost /]$
```

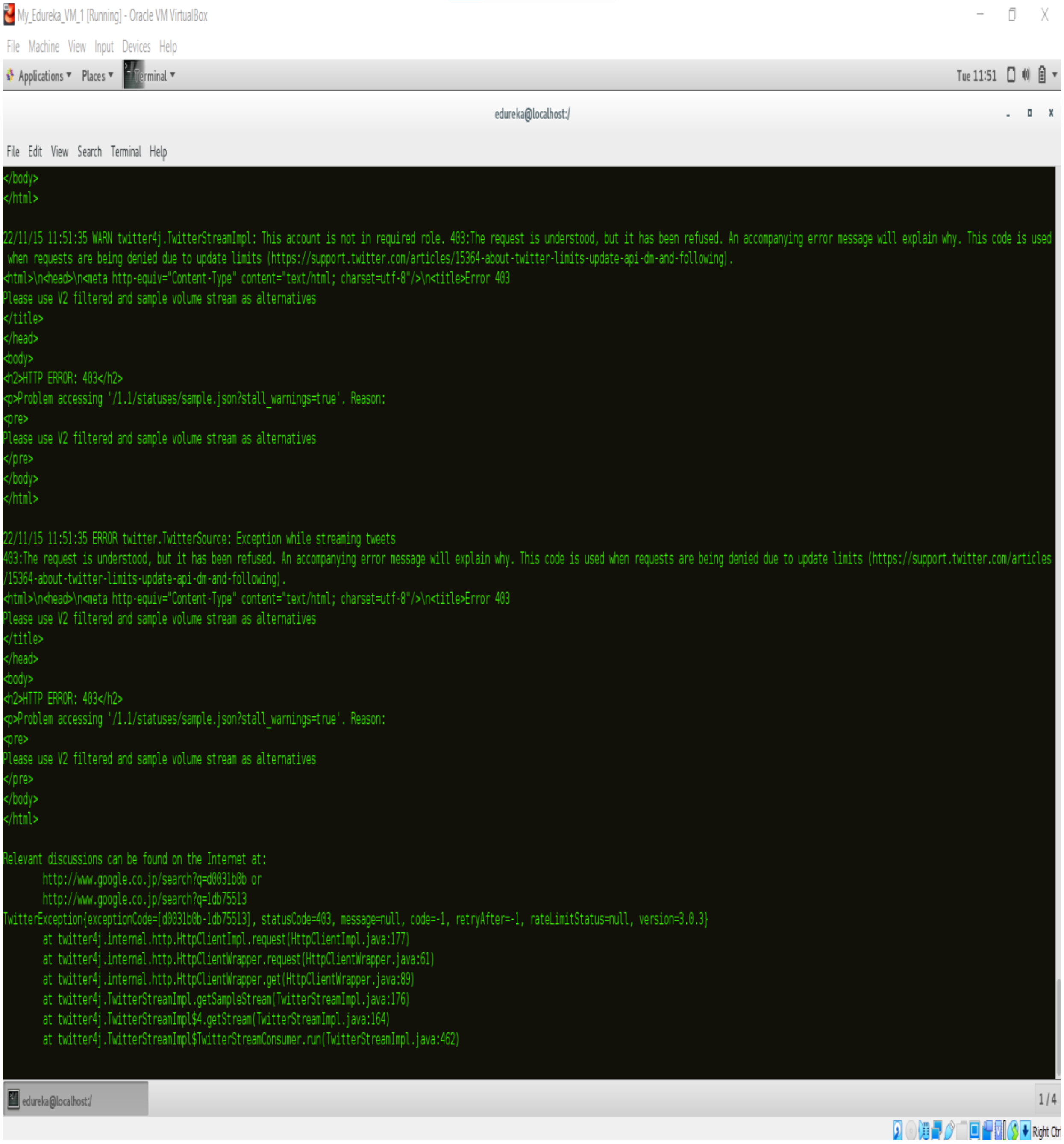
Step 15: After completion of step 14, use flume-ng agent to retrieve the data, to do that use the following command:

flume-ng agent --name TwitterAgent --conf-file /$FLUME_HOME/flume.conf



```
[edureka@localhost /]$ flume-ng agent --name TwitterAgent --conf-file /$FLUME_HOME/flume.conf
```

Output:



My_Edureka_VM_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places Terminal
Tue 11:51
edureka@localhost:/
File Edit View Search Terminal Help
edureka@localhost:/
1 / 4
Right Ctrl

```
</body>
</html>

22/11/15 11:51:35 WARN twitter4j.TwitterStreamImpl: This account is not in required role. 403:The request is understood, but it has been refused. An accompanying error message will explain why. This code is used
 when requests are being denied due to update limits (https://support.twitter.com/articles/15364-about-twitter-limits-update-api-dm-and-following).
<html>\n<head>\n<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>\n<title>Error 403
Please use V2 filtered and sample volume stream as alternatives
</title>
</head>
<body>
<h2>HTTP ERROR: 403</h2>
<p>Problem accessing '/1.1/statuses/sample.json?stall_warnings=true'. Reason:
<pre>
Please use V2 filtered and sample volume stream as alternatives
</pre>
</body>
</html>

22/11/15 11:51:35 ERROR twitter.TwitterSource: Exception while streaming tweets
403:The request is understood, but it has been refused. An accompanying error message will explain why. This code is used when requests are being denied due to update limits (https://support.twitter.com/articles
/15364-about-twitter-limits-update-api-dm-and-following).
<html>\n<head>\n<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>\n<title>Error 403
Please use V2 filtered and sample volume stream as alternatives
</title>
</head>
<body>
<h2>HTTP ERROR: 403</h2>
<p>Problem accessing '/1.1/statuses/sample.json?stall_warnings=true'. Reason:
<pre>
Please use V2 filtered and sample volume stream as alternatives
</pre>
</body>
</html>

Relevant discussions can be found on the Internet at:
        http://www.google.co.jp/search?q=d0031b0b or
        http://www.google.co.jp/search?q=1db75513
TwitterException{exceptionCode=[d0031b0b-1db75513], statusCode=403, message=null, code=-1, retryAfter=-1, rateLimitStatus=null, version=3.0.3}
        at twitter4j.internal.http.HttpClientImpl.request(HttpClientImpl.java:177)
        at twitter4j.internal.http.HttpClientWrapper.request(HttpClientWrapper.java:61)
        at twitter4j.internal.http.HttpClientWrapper.get(HttpClientWrapper.java:89)
        at twitter4j.TwitterStreamImpl.getSampleStream(TwitterStreamImpl.java:176)
        at twitter4j.TwitterStreamImpl$4.getStream(TwitterStreamImpl.java:164)
        at twitter4j.TwitterStreamImpl$TwitterStreamConsumer.run(TwitterStreamImpl.java:462)
```