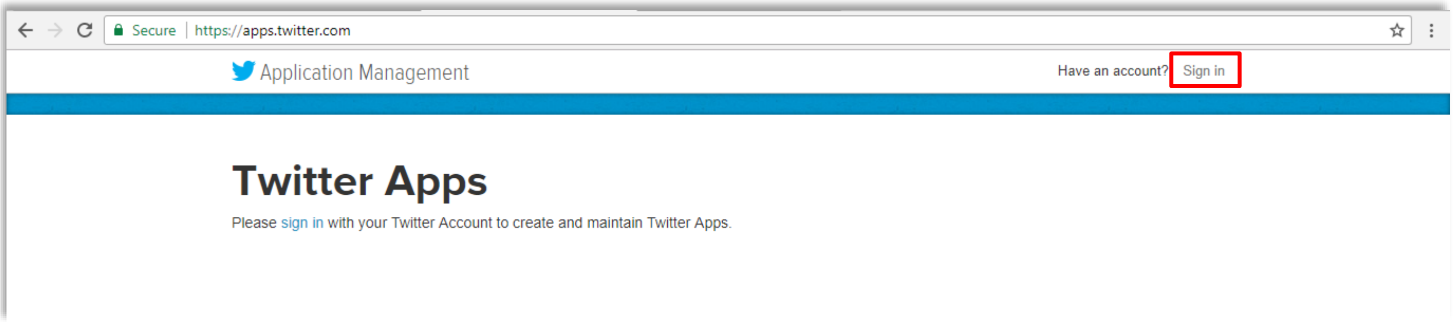
Module 10: Apache Oozie and Hadoop Project

Twitter Data Streaming using Apache Flume

edureka!

Step 1: Go to “apps.twitter.com” and click on “Sign in”



Secure | https://apps.twitter.com

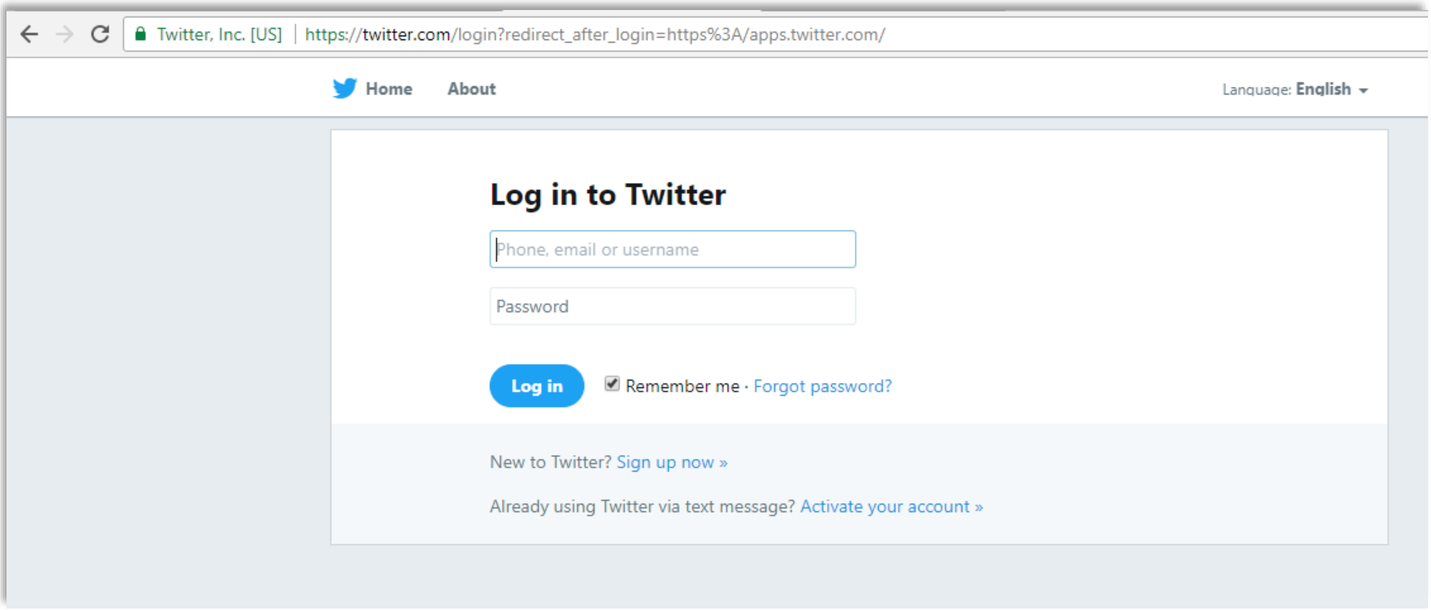
Application Management

Have an account? Sign in

Twitter Apps

Please sign in with your Twitter Account to create and maintain Twitter Apps.

Step 2: Log in to your twitter account



Twitter, Inc. [US] | https://twitter.com/login?redirect_after_login=https%3A/apps.twitter.com/

Log in to Twitter

Phone, email or username

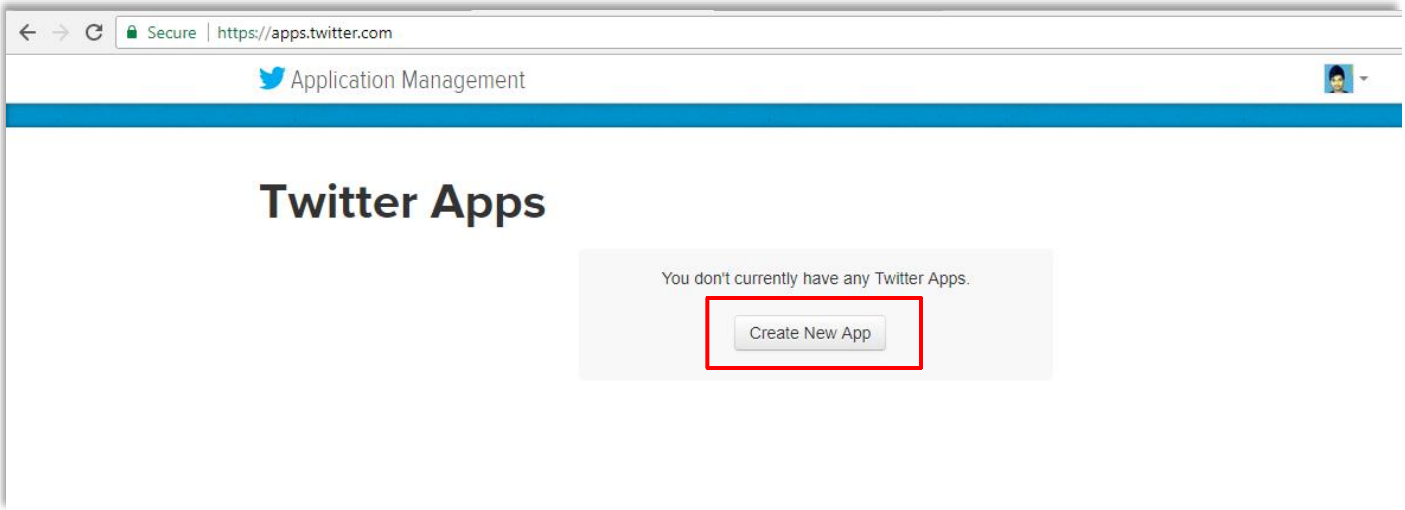
Password

Log in ☑ Remember me · Forgot password?

New to Twitter? Sign up now »

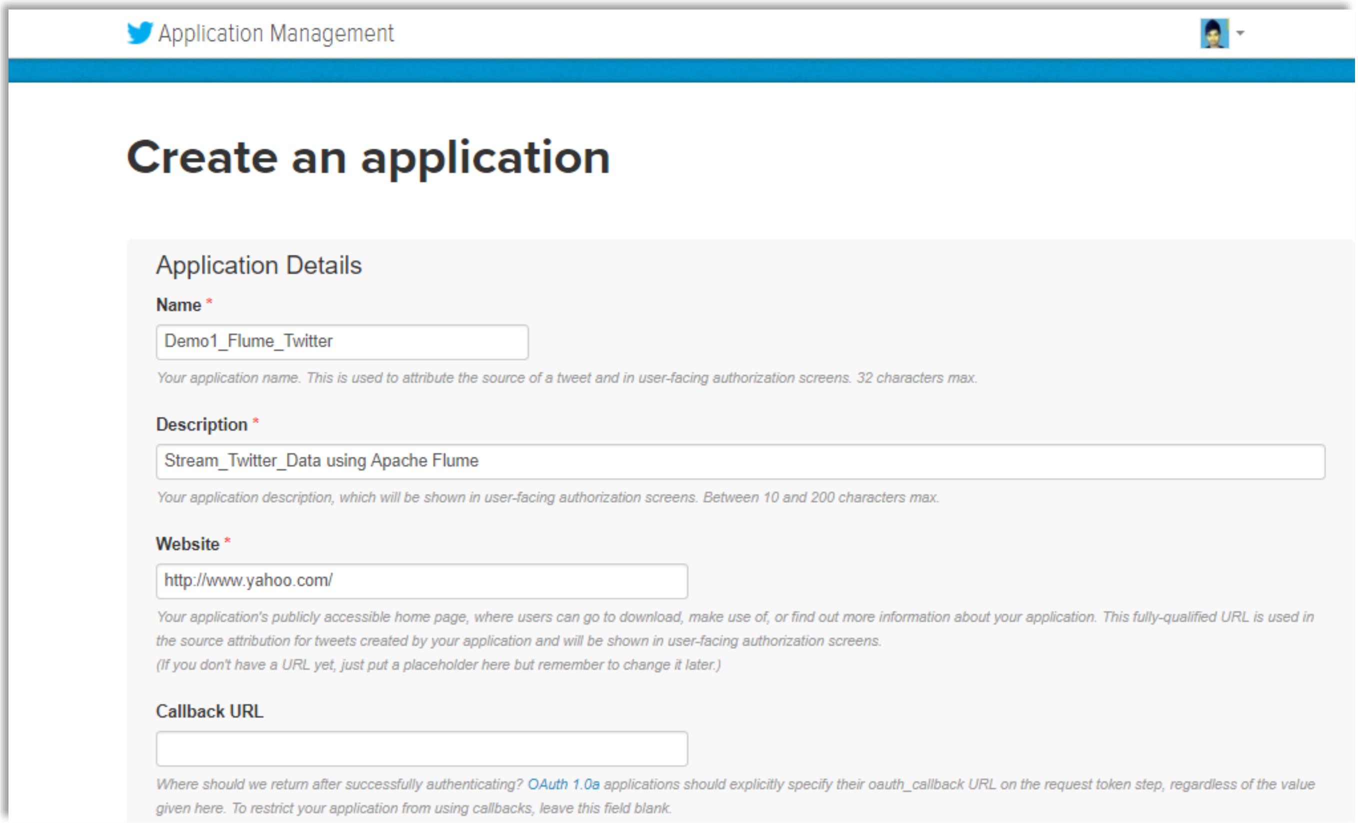
Already using Twitter via text message? Activate your account »

Step 3: Click on "Create New App"



Secure | https://apps.twitter.com
Application Management
Twitter Apps
You don't currently have any Twitter Apps.
Create New App

Step 4: Enter the application details



Create an application

Application Details

Name *

Demo1_Flume_Twitter

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Stream_Twitter_Data using Apache Flume

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

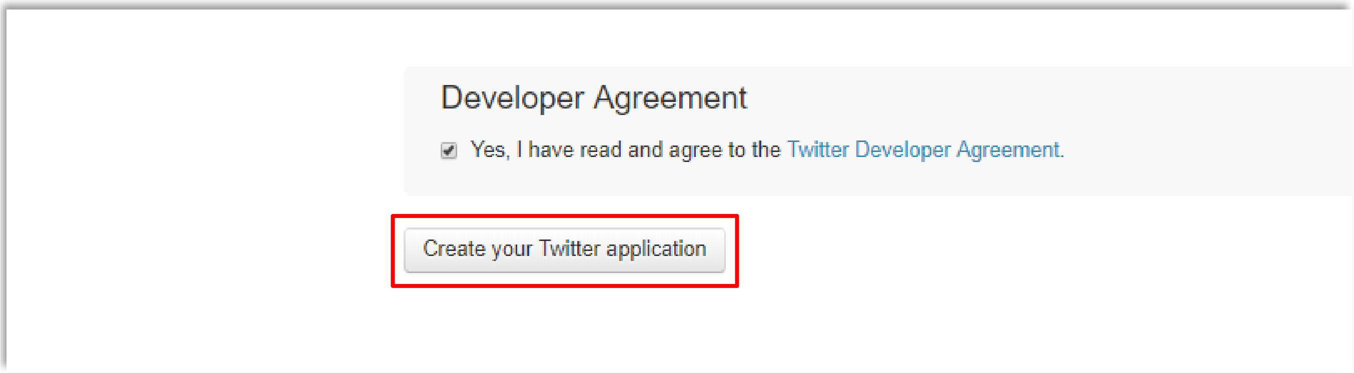
http://www.yahoo.com/

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

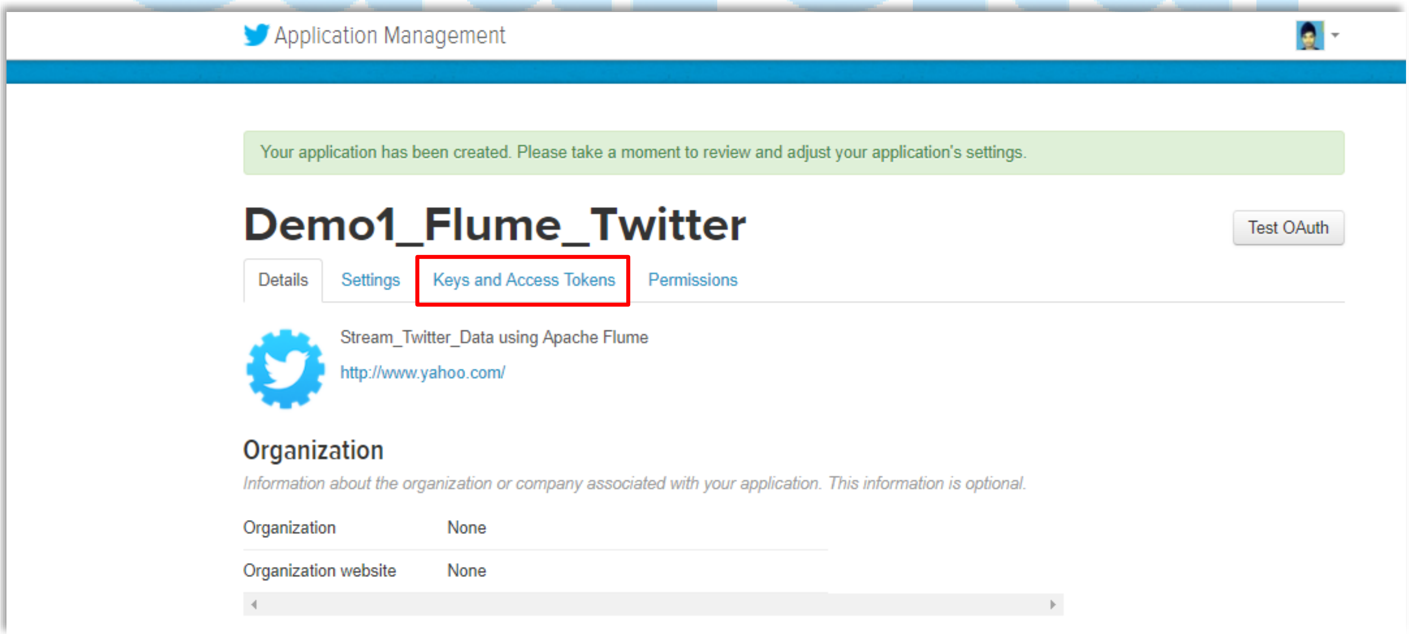
Step 5: Accept the Developer Agreement and click on “Create your Twitter application”



Developer Agreement

☑ Yes, I have read and agree to the Twitter Developer Agreement.

Step 6: Click on “Keys and Access Tokens”



Your application has been created. Please take a moment to review and adjust your application's settings.

Demo1_Flume_Twitter

Stream_Twitter_Data using Apache Flume

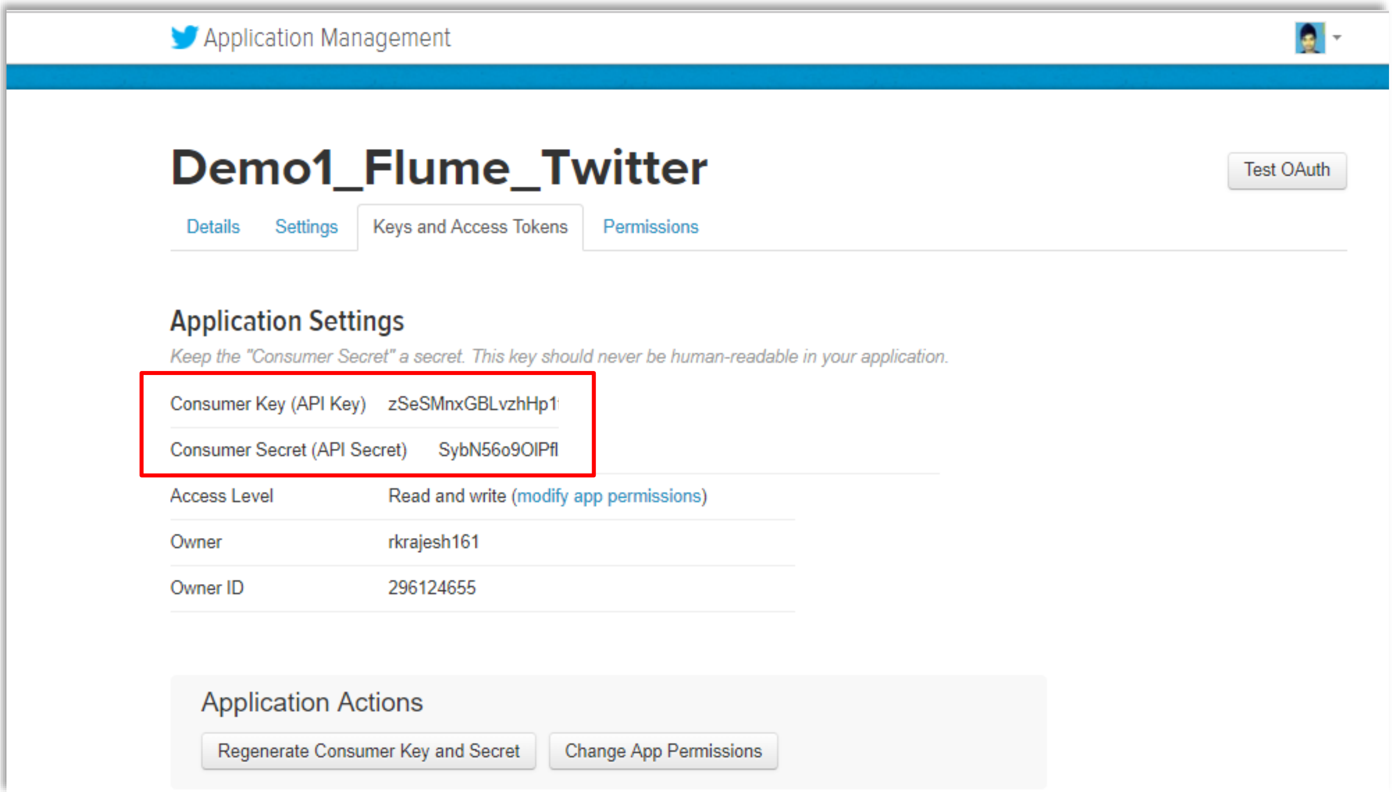
http://www.yahoo.com/

Organization

Information about the organization or company associated with your application. This information is optional.

Organization	None
Organization website	None

Step 7: Note the “Consumer Key”, “Consumer Secret” that is generated and click on “create my access token”



Application Management

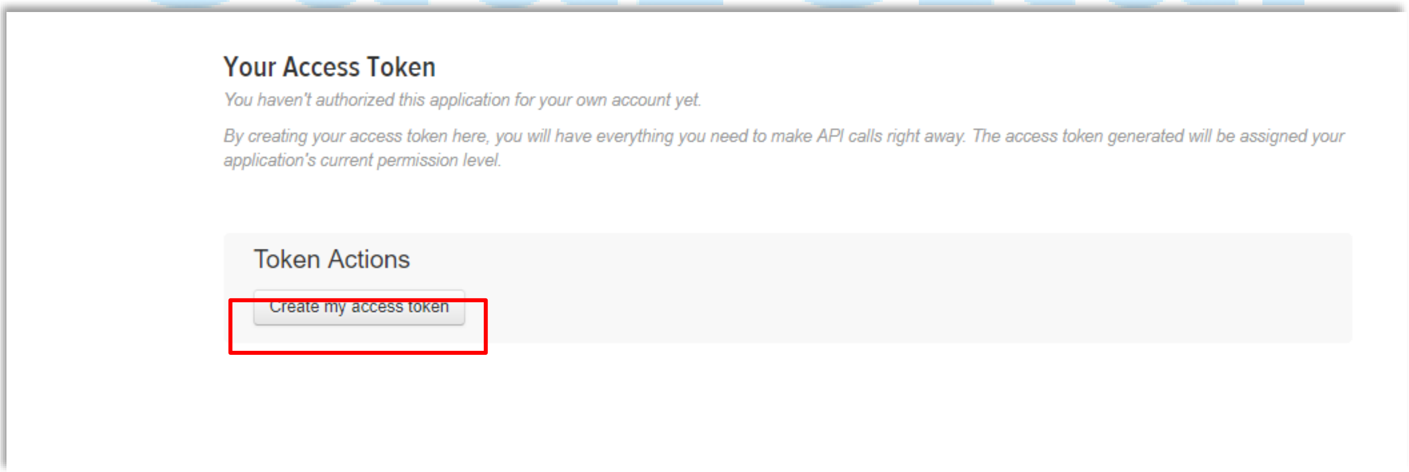
Demo1_Flume_Twitter

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	zSeSMnxGBLvzhHp1
Consumer Secret (API Secret)	SybN56o9OlPfl
Access Level	Read and write (modify app permissions)
Owner	rkrajesh161
Owner ID	296124655

Application Actions



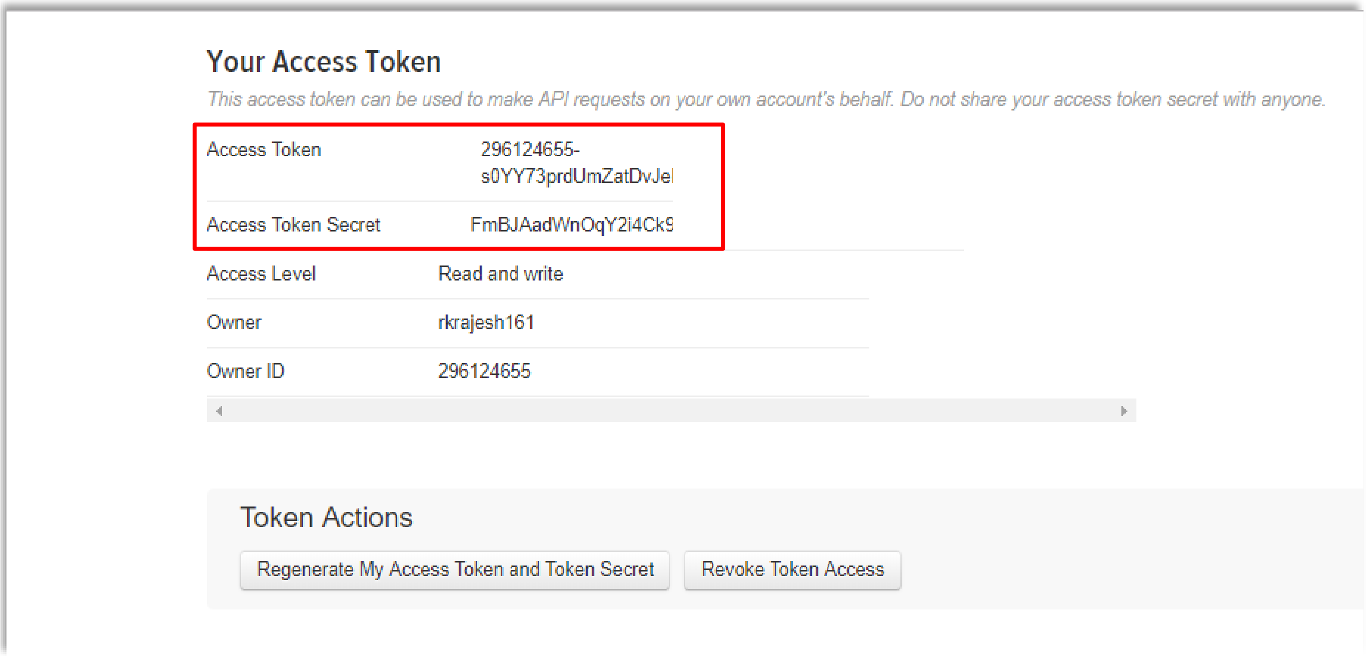
Your Access Token

You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assigned your application's current permission level.

Token Actions

Step 8: Here is the generated "Access Token" and "Access Token Secret" -



Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

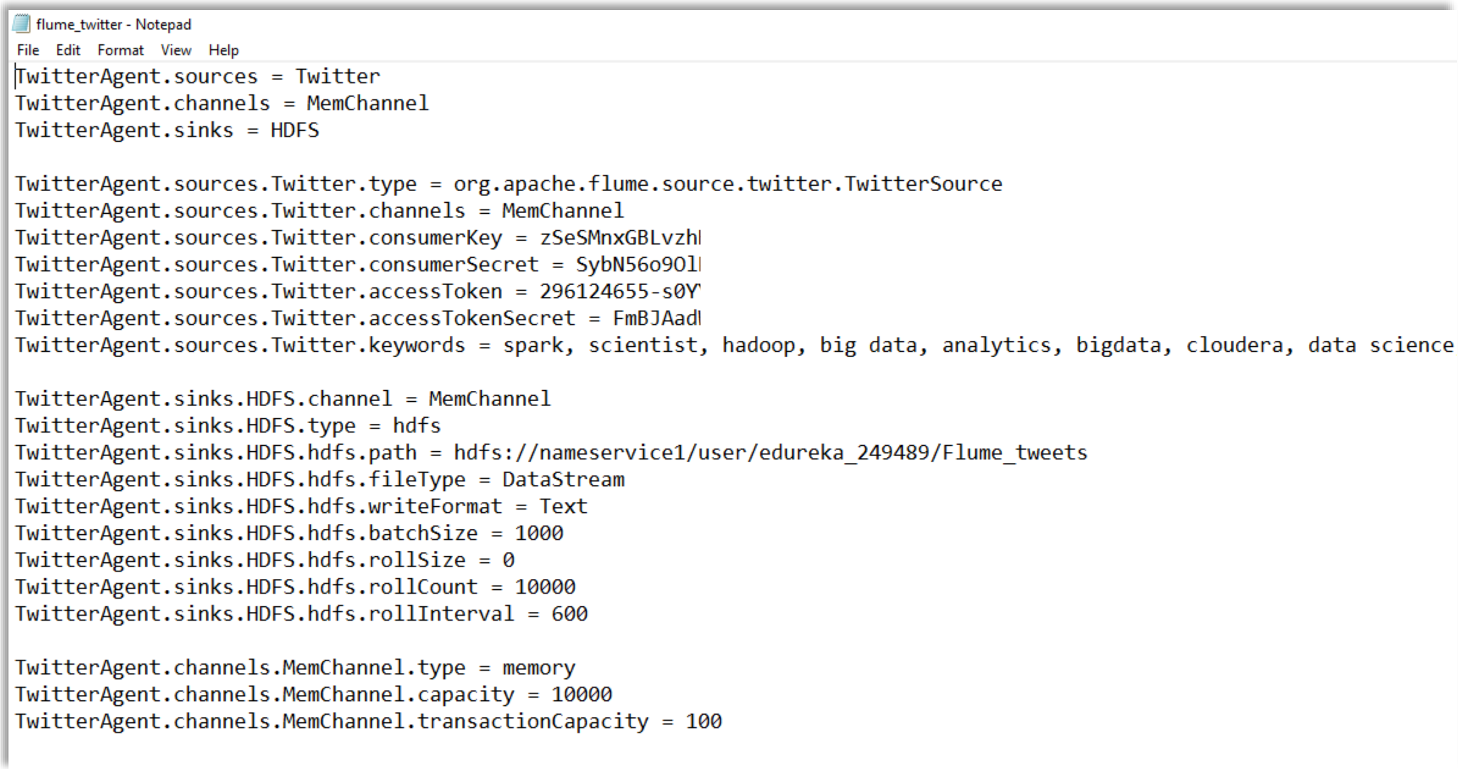
Access Token	296124655- s0YY73prdUmZatDvJel
Access Token Secret	FmBJAadWnOqY2i4Ck9
Access Level	Read and write
Owner	rkrajesh161
Owner ID	296124655

Step 9: Write the following configurations in a file “flume_twitter.conf”

Note: Here we will use the “Consumer Key”, “Consumer Secret”, “Access Token” and “Access Token Secret” generated in the above steps.

```
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 = zSeSMnxGBLvzhHp1
TwitterAgent.sources.Twitter.consumerSecret = SybN56o9OlPfl23OpJL6C8ahMSSdGuOmLOlUL8Lh
TwitterAgent.sources.Twitter.accessToken = 296124655-s0YY73prdUmZatDvJeMsTKv4ZJ4OfT
TwitterAgent.sources.Twitter.accessTokenSecret = FmBJAadWnOqY2i4Ck9ecssaQWYsCPLhmw4jjb

TwitterAgent.sources.Twitter.keywords = spark, scientist, hadoop, big data, analytics, bigdata, cloudera, data science, data scientist, 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://nameservice1/user/edureka_249489/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
```

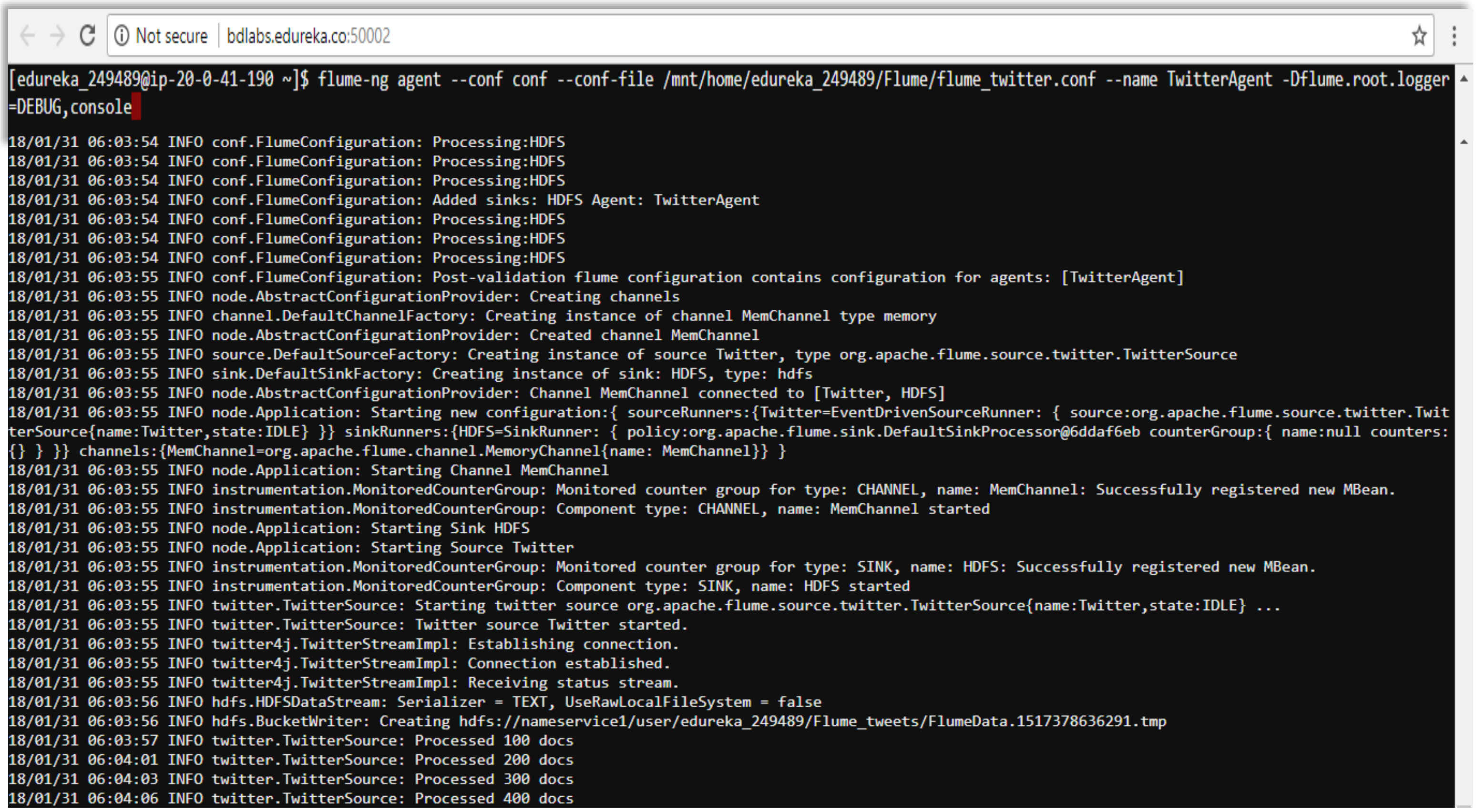
TwitterAgent.channels.MemChannel.transactionCapacity = 100



```
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 = zSeSMnxGBLvzh
TwitterAgent.sources.Twitter.consumerSecret = SybN56o9Ol
TwitterAgent.sources.Twitter.accessToken = 296124655-s0Y
TwitterAgent.sources.Twitter.accessTokenSecret = FmBJAad
TwitterAgent.sources.Twitter.keywords = spark, scientist, hadoop, big data, analytics, bigdata, cloudera, data science

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://nameservice1/user/edureka_249489/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 = 100
```

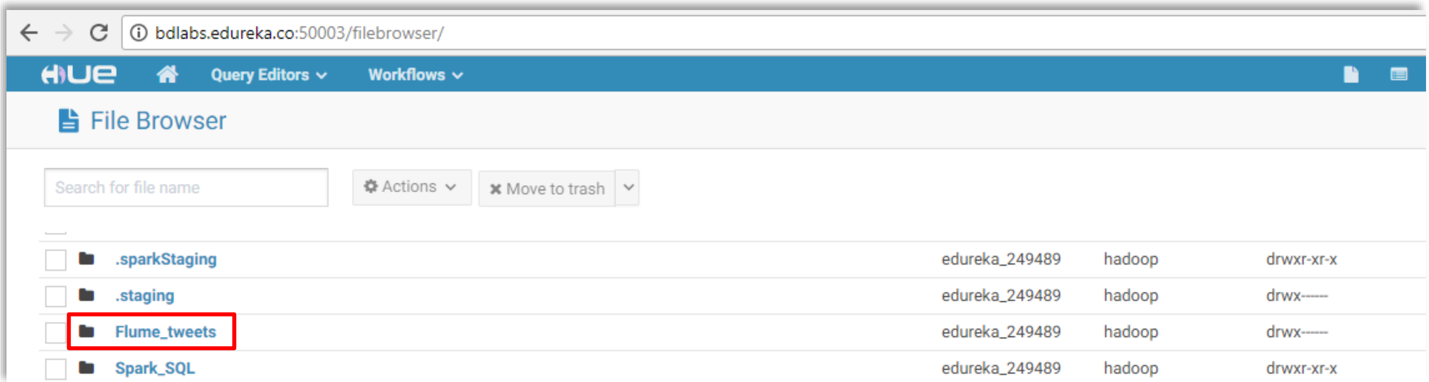
Step 10: Run the following command to capture twitter data

Command: `flume-ng agent --conf conf --conf-file /mnt/home/edureka_249489/Flume/flume_twitter.conf --name TwitterAgent -Dflume.root.logger=DEBUG,console`



```
[edureka_249489@ip-20-0-41-190 ~]$ flume-ng agent --conf conf --conf-file /mnt/home/edureka_249489/Flume/flume_twitter.conf --name TwitterAgent -Dflume.root.logger
=DEBUG,console

18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: TwitterAgent
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:54 INFO conf.FlumeConfiguration: Processing:HDFS
18/01/31 06:03:55 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAgent]
18/01/31 06:03:55 INFO node.AbstractConfigurationProvider: Creating channels
18/01/31 06:03:55 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
18/01/31 06:03:55 INFO node.AbstractConfigurationProvider: Created channel MemChannel
18/01/31 06:03:55 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.TwitterSource
18/01/31 06:03:55 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
18/01/31 06:03:55 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
18/01/31 06:03:55 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { source:org.apache.flume.source.twitter.Twit
terSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache.flume.sink.DefaultSinkProcessor@6ddaf6eb counterGroup:{ name:null counters:
{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name: MemChannel}} }
18/01/31 06:03:55 INFO node.Application: Starting Channel MemChannel
18/01/31 06:03:55 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.
18/01/31 06:03:55 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
18/01/31 06:03:55 INFO node.Application: Starting Sink HDFS
18/01/31 06:03:55 INFO node.Application: Starting Source Twitter
18/01/31 06:03:55 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered new MBean.
18/01/31 06:03:55 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/01/31 06:03:55 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} ...
18/01/31 06:03:55 INFO twitter.TwitterSource: Twitter source Twitter started.
18/01/31 06:03:55 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/01/31 06:03:55 INFO twitter4j.TwitterStreamImpl: Connection established.
18/01/31 06:03:55 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/01/31 06:03:56 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/01/31 06:03:56 INFO hdfs.BucketWriter: Creating hdfs://nameservice1/user/edureka_249489/Flume_tweets/FlumeData.1517378636291.tmp
18/01/31 06:03:57 INFO twitter.TwitterSource: Processed 100 docs
18/01/31 06:04:01 INFO twitter.TwitterSource: Processed 200 docs
18/01/31 06:04:03 INFO twitter.TwitterSource: Processed 300 docs
18/01/31 06:04:06 INFO twitter.TwitterSource: Processed 400 docs
```

Step 11: Go to Hue and browse to the HDFS directory and click on "Flume_tweets"

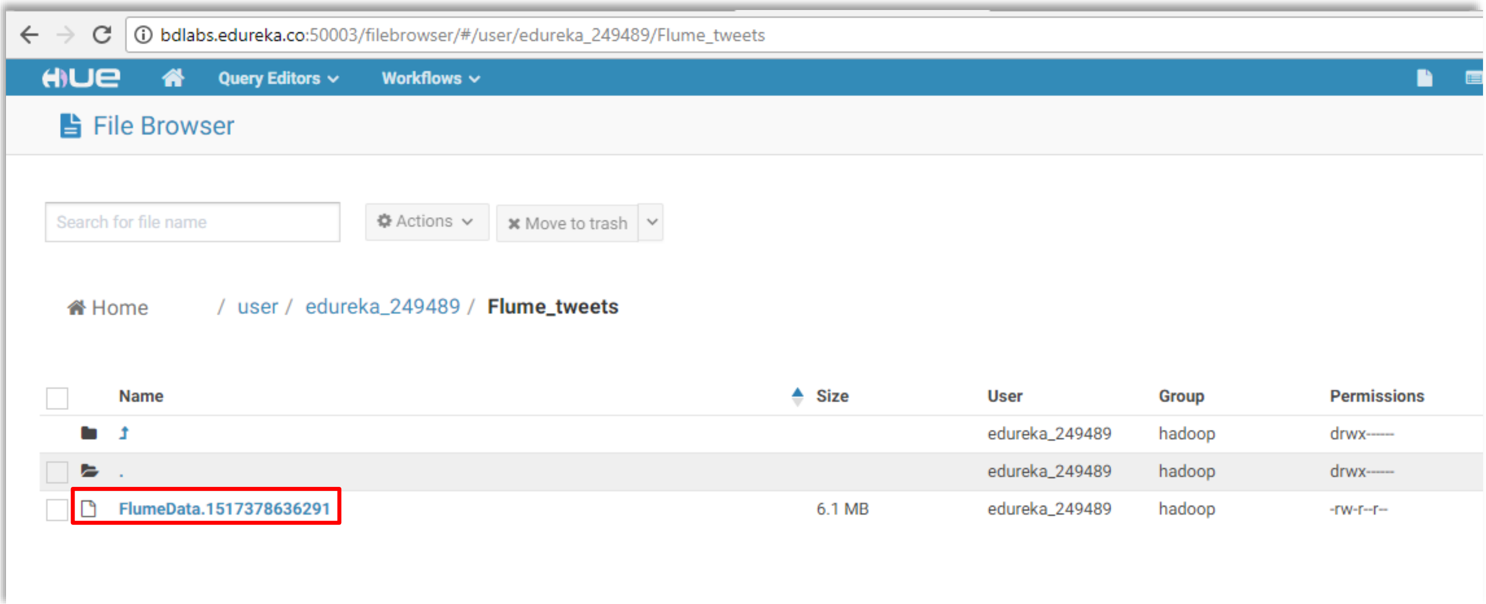


bdlabs.edureka.co:50003/filebrowser/

File Browser

.sparkStaging	edureka_249489	hadoop	drwxr-xr-x
.staging	edureka_249489	hadoop	drwx------
Flume_tweets	edureka_249489	hadoop	drwx------
Spark_SQL	edureka_249489	hadoop	drwxr-xr-x

Step 12: Click on "FlumeData.1517378636291" file



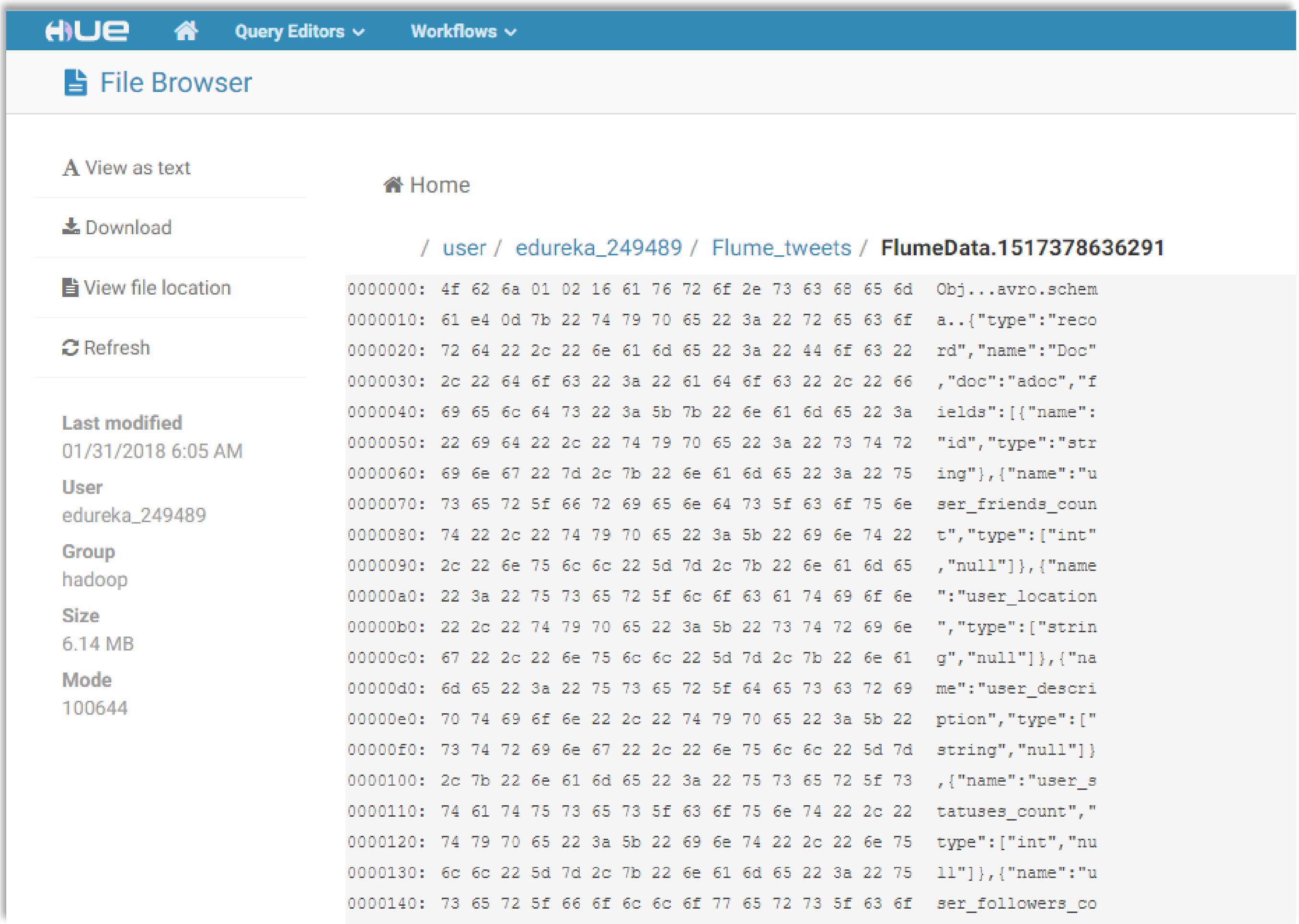
bdlabs.edureka.co:50003/filebrowser/#/user/edureka_249489/Flume_tweets

File Browser

Home / user / edureka_249489 / Flume_tweets

Name	Size	User	Group	Permissions
↥		edureka_249489	hadoop	drwx------
.		edureka_249489	hadoop	drwx------
FlumeData.1517378636291	6.1 MB	edureka_249489	hadoop	-rw-r--r--

Step13: Check the output. This is the data that has been downloaded from Twitter.



Last modified
01/31/2018 6:05 AM
User
edureka_249489
Group
hadoop
Size
6.14 MB
Mode
100644

Home

/ user / edureka_249489 / Flume_tweets / FlumeData.1517378636291

```
0000000: 4f 62 6a 01 02 16 61 76 72 6f 2e 73 63 68 65 6d  Obj...avro.schem
0000010: 61 e4 0d 7b 22 74 79 70 65 22 3a 22 72 65 63 6f  a..{"type":"reco
0000020: 72 64 22 2c 22 6e 61 6d 65 22 3a 22 44 6f 63 22  rd","name":"Doc"
0000030: 2c 22 64 6f 63 22 3a 22 61 64 6f 63 22 2c 22 66  ,"doc":"adoc","f
0000040: 69 65 6c 64 73 22 3a 5b 7b 22 6e 61 6d 65 22 3a  ields":[{"name":
0000050: 22 69 64 22 2c 22 74 79 70 65 22 3a 22 73 74 72  "id","type":"str
0000060: 69 6e 67 22 7d 2c 7b 22 6e 61 6d 65 22 3a 22 75  ing"},{"name":"u
0000070: 73 65 72 5f 66 72 69 65 6e 64 73 5f 63 6f 75 6e  ser_friends_coun
0000080: 74 22 2c 22 74 79 70 65 22 3a 5b 22 69 6e 74 22  t","type":["int"
0000090: 2c 22 6e 75 6c 6c 22 5d 7d 2c 7b 22 6e 61 6d 65  ,"null"]},{"name
00000a0: 22 3a 22 75 73 65 72 5f 6c 6f 63 61 74 69 6f 6e  ":"user_location
00000b0: 22 2c 22 74 79 70 65 22 3a 5b 22 73 74 72 69 6e  ","type":["strin
00000c0: 67 22 2c 22 6e 75 6c 6c 22 5d 7d 2c 7b 22 6e 61  g","null"]},{"na
00000d0: 6d 65 22 3a 22 75 73 65 72 5f 64 65 73 63 72 69  me":"user_descri
00000e0: 70 74 69 6f 6e 22 2c 22 74 79 70 65 22 3a 5b 22  ption","type":["
00000f0: 73 74 72 69 6e 67 22 2c 22 6e 75 6c 6c 22 5d 7d  string","null"]}
0000100: 2c 7b 22 6e 61 6d 65 22 3a 22 75 73 65 72 5f 73  ,{"name":"user_s
0000110: 74 61 74 75 73 65 73 5f 63 6f 75 6e 74 22 2c 22  tatuses_count","
0000120: 74 79 70 65 22 3a 5b 22 69 6e 74 22 2c 22 6e 75  type":["int","nu
0000130: 6c 6c 22 5d 7d 2c 7b 22 6e 61 6d 65 22 3a 22 75  ll"]},{"name":"u
0000140: 73 65 72 5f 66 6f 6c 6c 6f 77 65 72 73 5f 63 6f  ser_followers_co
```

Note: Press 'ctrl + c' on webconsole to stop the generation of tweets.