Fetching twitter data using Flume
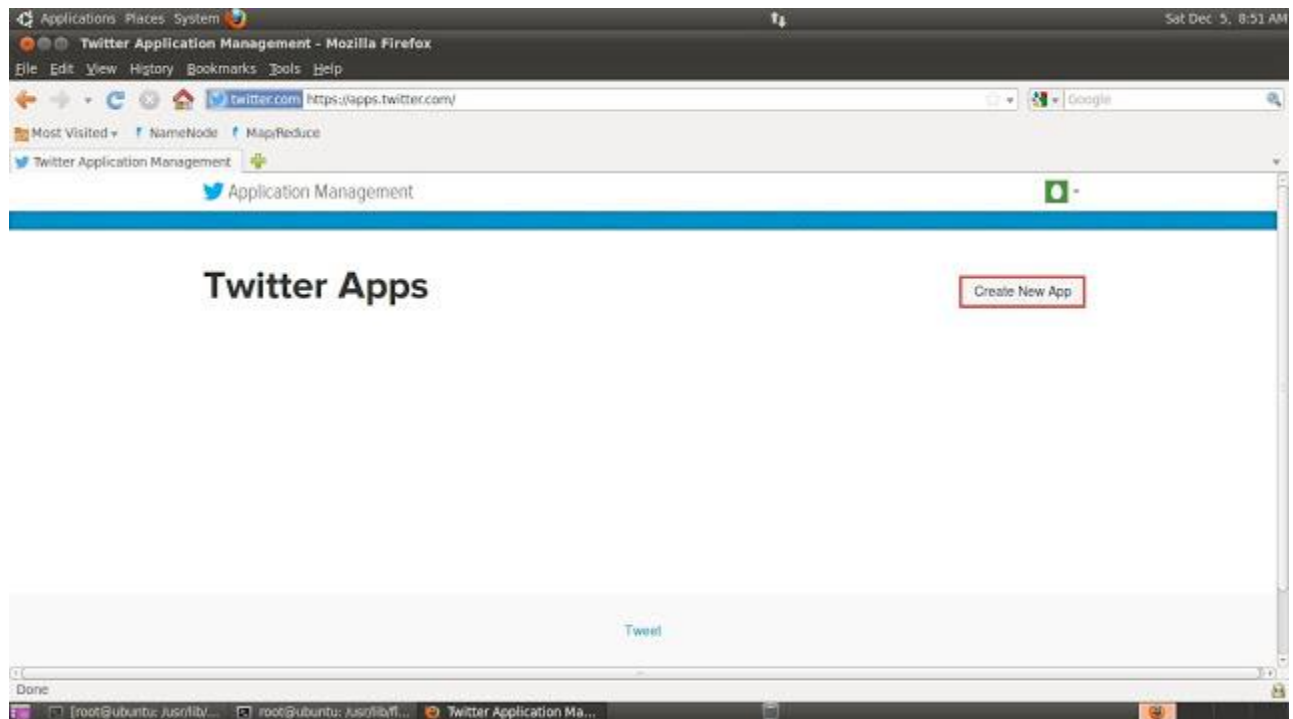
Setup for Twitter

Open https://apps.twitter.com
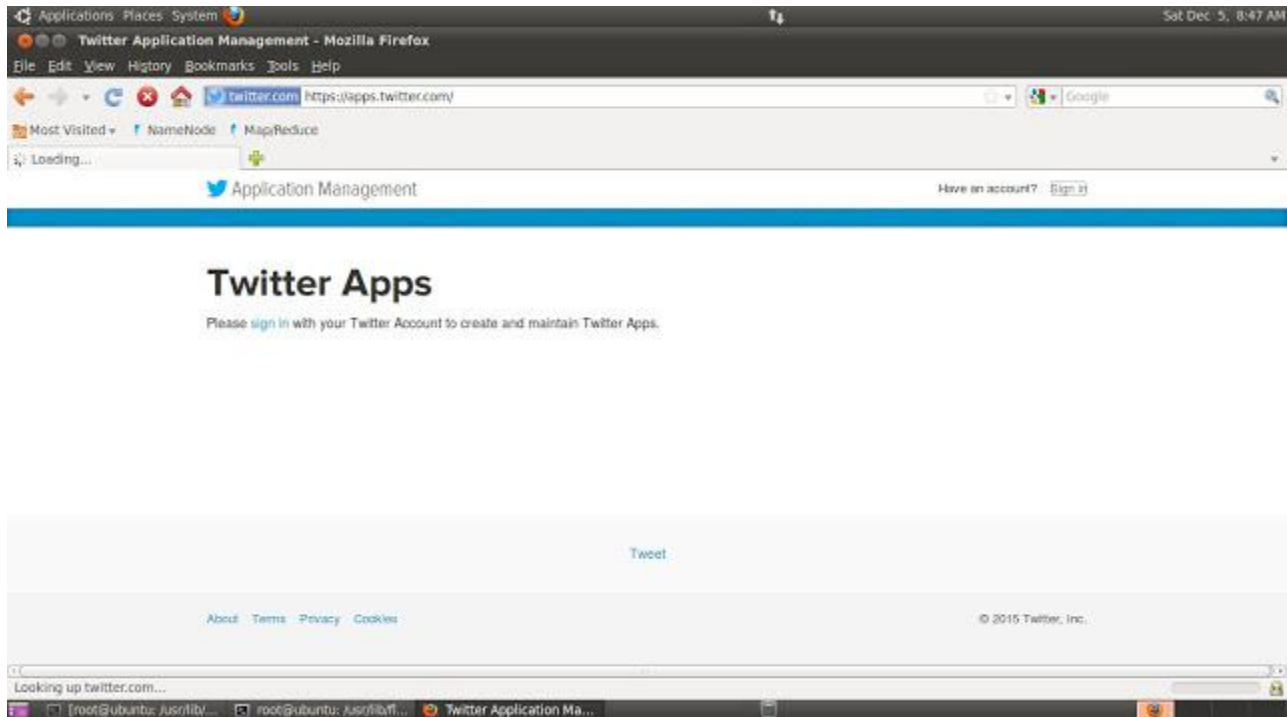
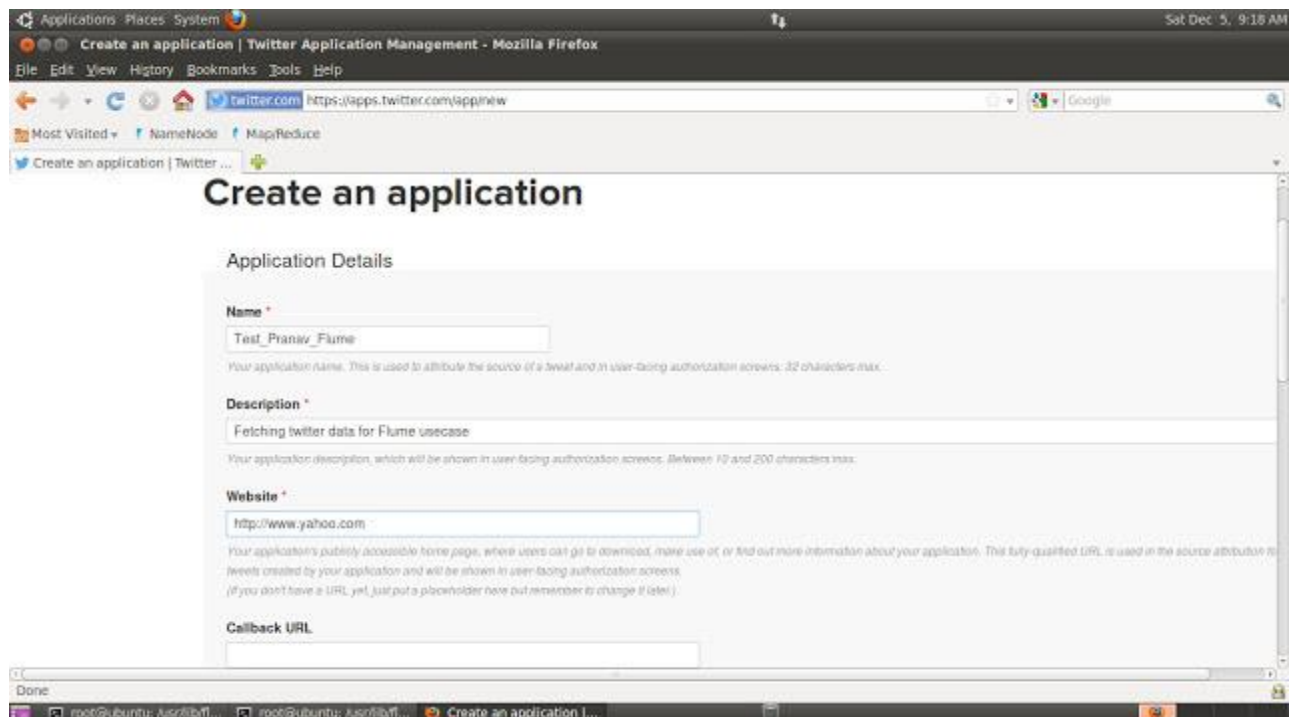
Sign In, the twitter account.
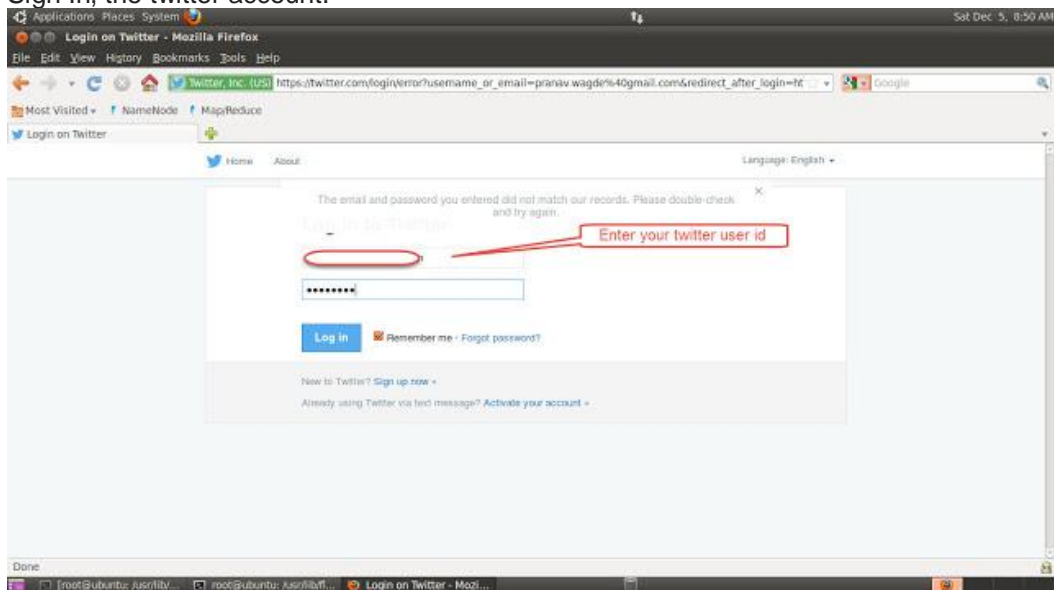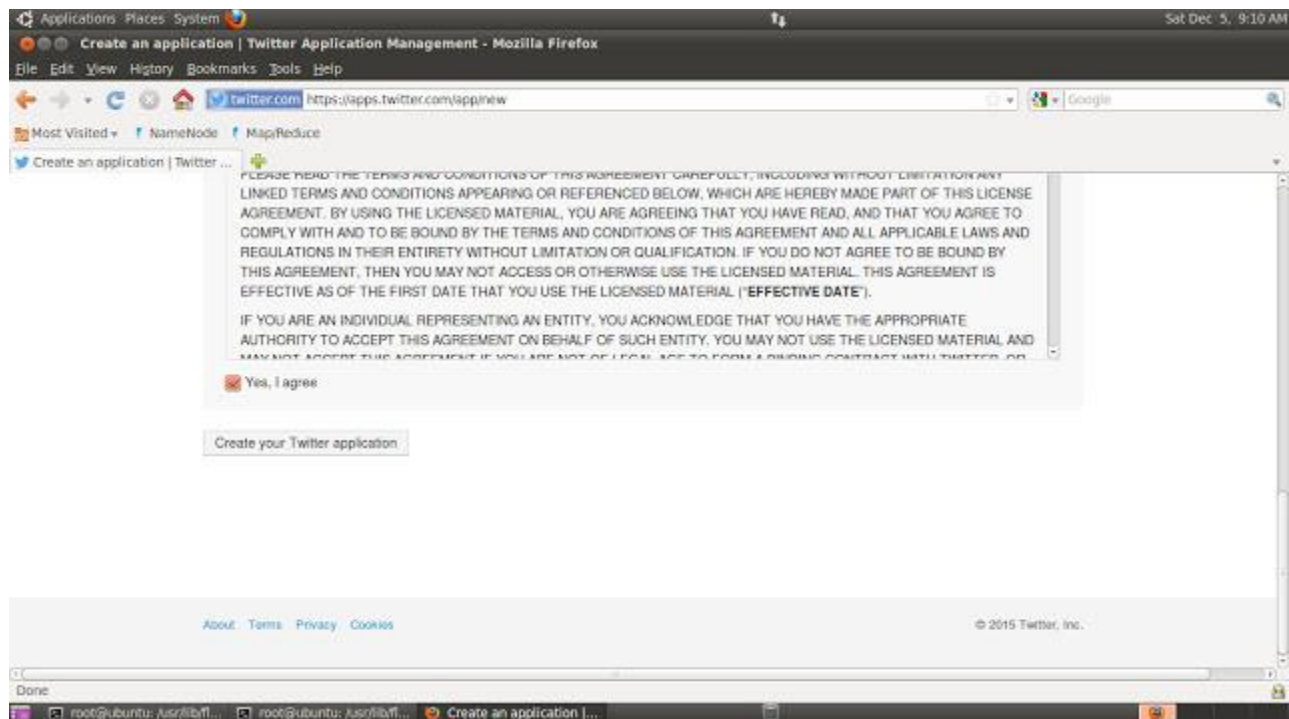
Click on Create New App

Give the details as shown below. **(Make sure you will give unique name to the application)**

Note: Website should be fully qualified URL.
e.g., instead of google.com, should give http://www.google.com

Tick "Yes, I agree" and click Create your Twitter application.

Application will be created as below

Application Management

RameshBatch84

Details | Settings | Keys and Access Tokens | Permissions

flume exmaple
https://www.kellytechnologies.com

Organization

Information about the organization or company associated with your application. This information is optional.
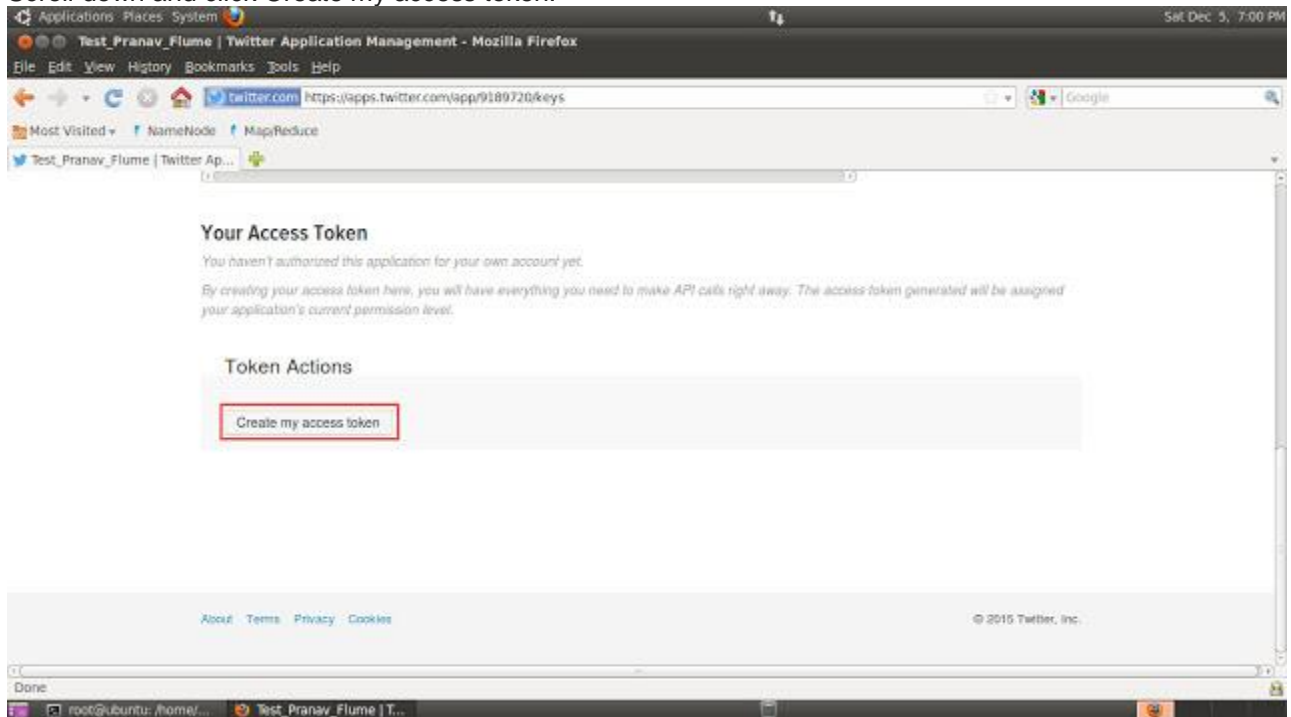
Organization   None

Organization website   None

Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level   Read and write (modify app permissions)

Click on the created application and select tab "Keys and Access Tokens".

Scroll down and click Create my access token.

If the access token successfully created, you will get below message as Status.

These 4 highlighted details in above screenshot, we need as a part of flume configuration file(mentioned in the below screenshot).

We could place the flume.conf file at the $flume_home/conf/flume.conf

Add the details into flume.conf as below.
CosumerKey == Consumer Key from Twitter
ConsumerSecret == Consumer Secret from Twitter
accessToken == Access Token from twitter

accessTokenSecret == Access Token Secret from twitter

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=GK290CbplPqBMiM8ZwAQgmY5i
TwitterAgent.sources.Twitter.consumerSecret=URDN1PMwjDjsb5phFo9XJyKU0Y4TubBhu7Q9LJSTgIDdZ0ukhn
TwitterAgent.sources.Twitter.accessToken=277481794-aMWA5069hTBKFbdtee6mdIOegoKJcP29rh4zkbYw
TwitterAgent.sources.Twitter.accessTokenSecret= QQ0GJNcCygEzAZml21hPKi2F0u0NdpaCEVgbv11YAL8Cf
TwitterAgent.sources.Twitter.keywords=hadoop,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:8020/user/flume/tweets1
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
~
~
~
~
~
```

Run the below command to fetch the data from twitter.

If everything goes right, we can see the below output, on HDFS location, mentioned in the flume.conf file.

Error Rectifications

While executive the Flume command, user may get the following error.

Error says that Ensure that you have set the valid consumer key/secret, access token/secret and system clock is in sync.

Resolution 1: First of all, check the access key/secret and access tokens/secret are correct as per twitter's application values.

Resolution 2: There may be a chance that the Host OS and the guest OS has difference in timezone, e.g., My Host OS Windows 8 has time set as per India (IST) and Guest OS has US time zone.

To resolve the timezone problem, do the following.

From root user login
Stop ntp service
$ service ntp stop => This will stop the ntp service
$ ntpdate ntp.ubuntu.com => This will update the Guest OS time same as Host OS
$ service ntp start => This will start the ntp service

Re-execute the Flume command...