

BIG DATA
DEVELOPMENT

ACADGILD

Session 11: Sqoop Flume

Assignment 3

Big Data and Hadoop Development

Problem Statement

Create a flume agent that streams data from Twitter and stores in the HDFS.

Step 1

Login to given link <https://apps.twitter.com/> and sign in.

Step 2

Click on the **Create New App** button.

Below given window will pop-up

Create an application

Application Details

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

Description *
This app will help to analyse tweet data
Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *
http://www.google.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 successful authentication? 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.

Developer Agreement
☒ Yes, I have read and agree to the [Twitter Developer Agreement](#).

Big Data and Hadoop Development

Application Details

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

Description *
This app will help to analyse tweet data
Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *
http://www.wipro.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.)

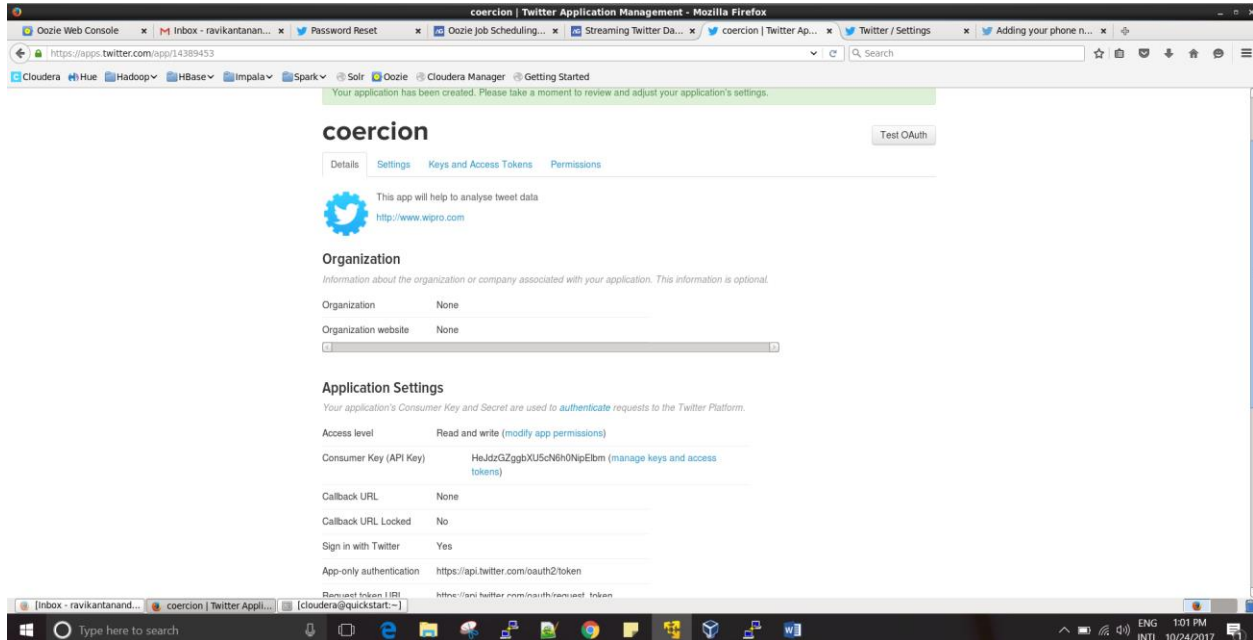
Callback URL
Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Developer Agreement
☒ Yes, I have read and agree to the [Twitter Developer Agreement](#).

Step 3

Fill in the details, accept the **Developer Agreement** when finished, click on the **Create your Twitter application button** which is at the bottom of the page. If everything goes fine, an App will be created with the given details as shown below.

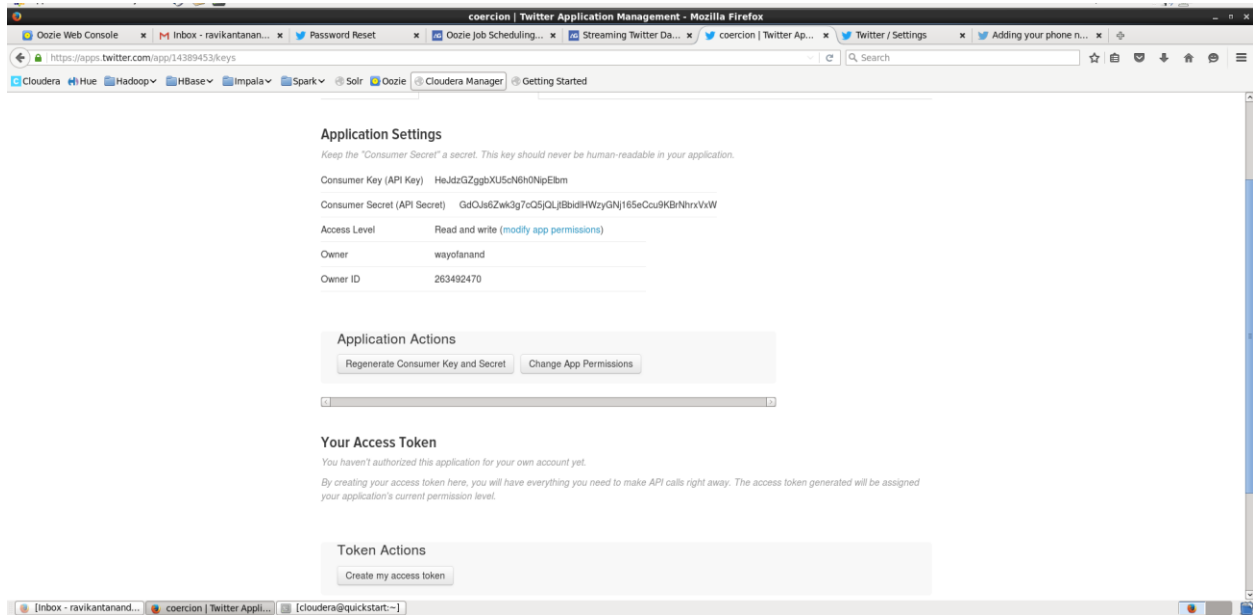
Big Data and Hadoop Development



Step 4

Under **keys and Access Tokens** tab at the bottom of the page, you can observe a button named **Create my access token**. Click on it to generate the access token.

Big Data and Hadoop Development



Consumer Key (API Key) HeJdzGZggbXU5cN6h0NipElbm

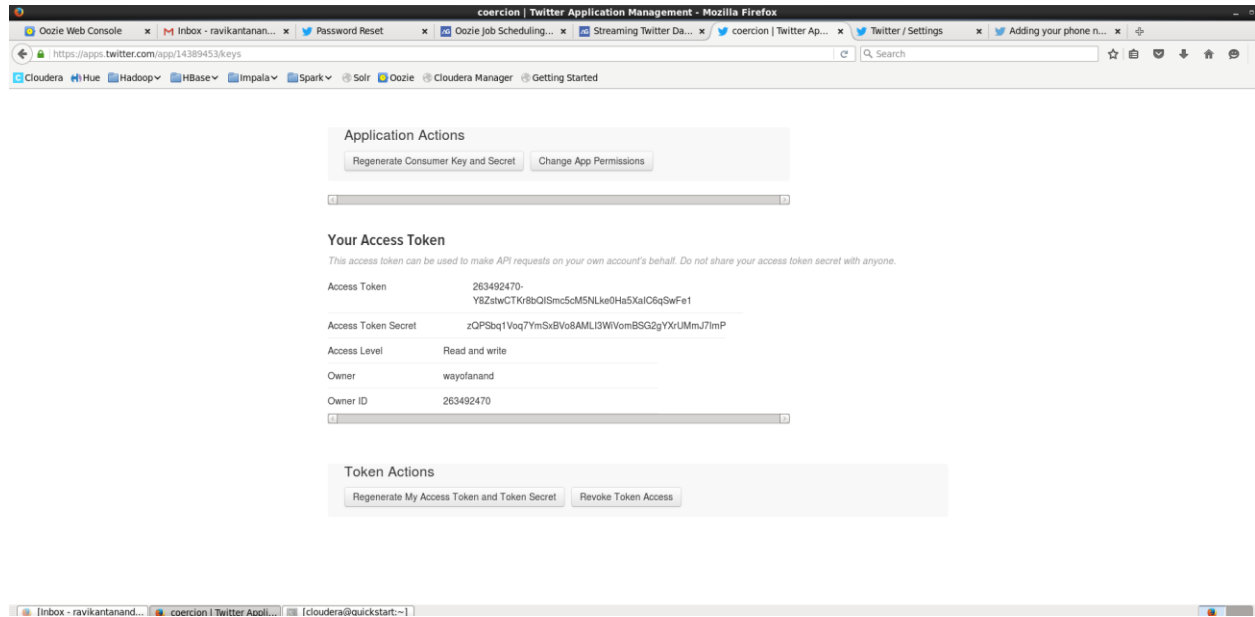
Consumer Secret (API Secret)

GdOJs6Zwk3g7cQ5jQLjBbidlHWzyGNj165eCcu9KBrNhrxVxW

Step 5

Finally, click on the **Test OAuth** button which is on the right side top of the page. This will lead to a page which displays your **Consumer key**, **Consumer secret**, **Access token**, and **Access token secret**. Copy these details. These are useful to configure the agent in Flume

Big Data and Hadoop Development



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 263492470-Y8ZstwCTKr8bQISmc5cM5NLke0Ha5XaIC6qSwFe1

Access Token Secret zQPSbq1Voq7YmSxBVo8AMLI3WiVomBSG2gYXrUMmJ7ImP

Access Level Read and write

Owner wayofanand

Owner ID 263492470

Big Data and Hadoop Development

```
[acadgild@localhost ~]$ hadoop fs -mkdir /user/flume/tweets/
```

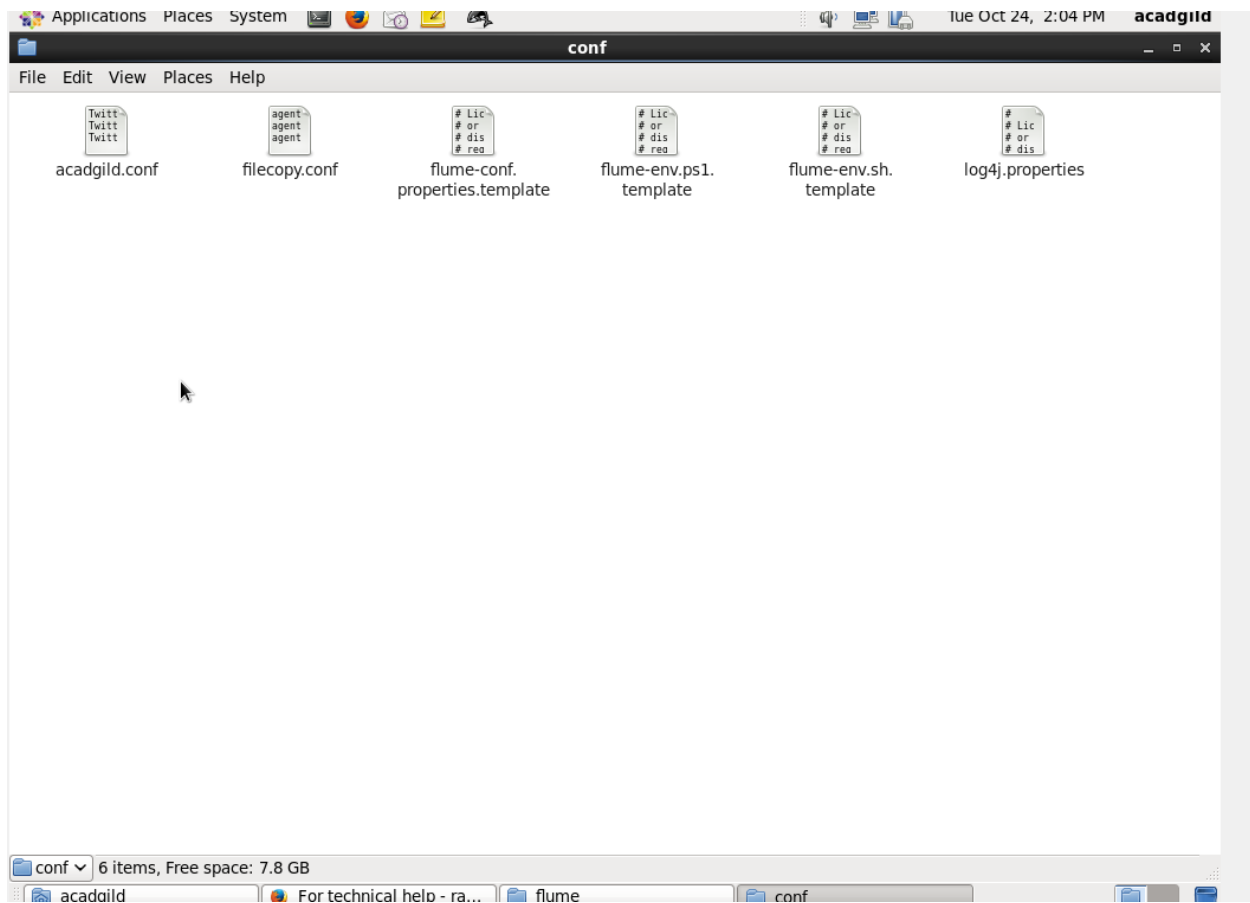
```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/
```

```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/
```

```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/flume/conf/acadgild.conf
```

```
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/flume/conf/acadgild.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop) for HDFS access
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/usr/local/hbase/Bin/hbase) for HBASE access
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
+ exec /usr/local/java/bin/java -Xmx20m -cp "/usr/local/flume/lib/*:/usr/local/hadoop-2.6.0/contrib/capacity-scheduler/*.jar:/usr/local/hadoop-2.6.0/etc/hadoop:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/activation-1.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/apachedes-118n-2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/apachedes-kerberos-codec-2.0.0-M15.jar:/usr/local/hadoop-p-2.6.0/share/hadoop/common/lib/api-asnl-api-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/asml-sm-3.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/avro-1.7.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-codec-1.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-collections-3.2.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-compress-1.4.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-digester-1.8.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-el-1.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-httpclient-3.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-io-2.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-logging-1.1.3.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-net-3.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib curator-client-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib curator-framework-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib curator-recipes-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib gson-2.2.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/guava-11.0.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/hadoop-annotations-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/hadoop-auth-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/hamcrest-core-1.3.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/htrace-core-3.0.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/httpclient-4.2.5.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/httpcore-4.2.5.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-jaxrs-1.9.13.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jasper-compiler-5.23.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jasper-runtime-5.23.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jersey-core-1.
```

[Big Data and Hadoop Development](#)



Big Data and Hadoop Development

```
acadgild.conf x
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=HeJdzGZggbXU5cN6h0NipElbm
TwitterAgent.sources.Twitter.consumerSecret=Gd0Js6Zwk3g7cQ5jQLjtBbidlHWzyGNj165eCcu9KBrNhrxVxw
TwitterAgent.sources.Twitter.accessToken=263492470-Y8ZstwCTKr8bQISmc5cM5NLke0Ha5XaIC6qSwFe1
TwitterAgent.sources.Twitter.accessTokenSecret=zQPSbqlVoq7YmSxBVo8AMLi3WiVomBSG2gYXRUMmJ7Imp
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, 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:9000/user/flume/tweetss
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

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