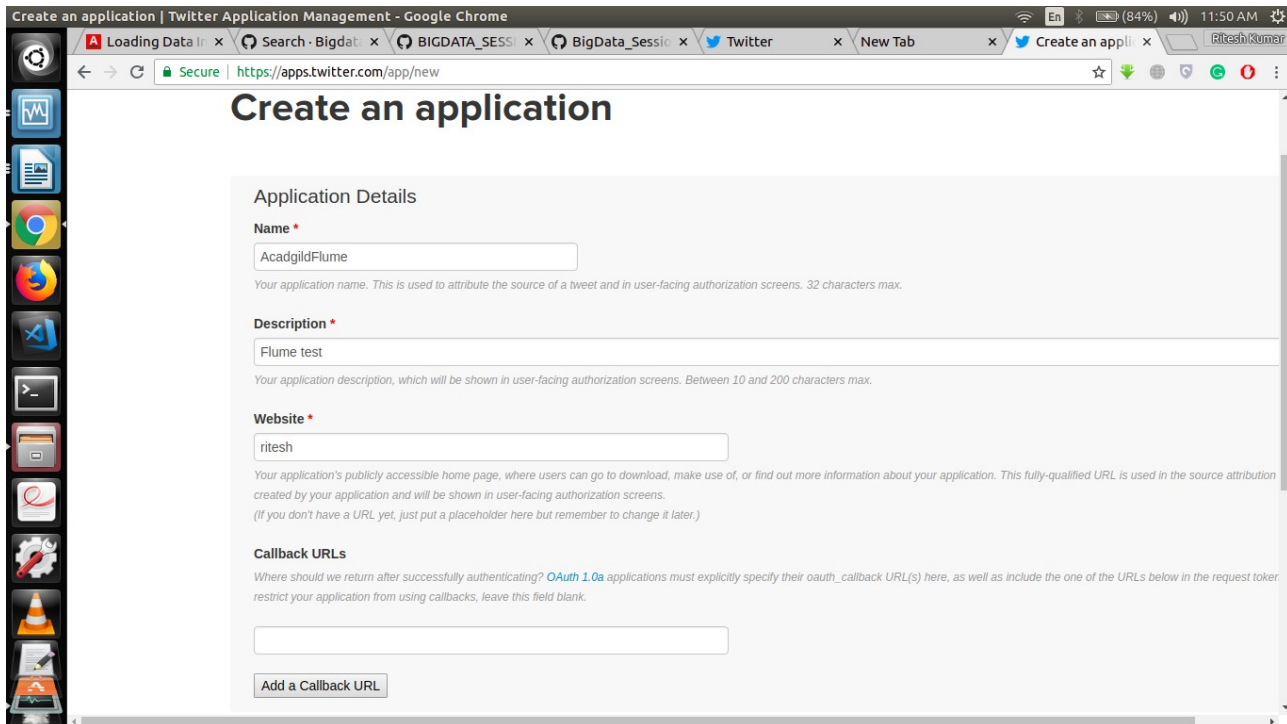


Bigdata Assignment 4.5

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

Solution :

- Login to twitter
- Go to “<https://apps.twitter.com/app>” and create new app.
- Fill the details as hown in the below screenshot



Create an application | Twitter Application Management - Google Chrome

Secure | <https://apps.twitter.com/app/new>

Create an application

Application Details

Name *

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

Description *

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

Website *

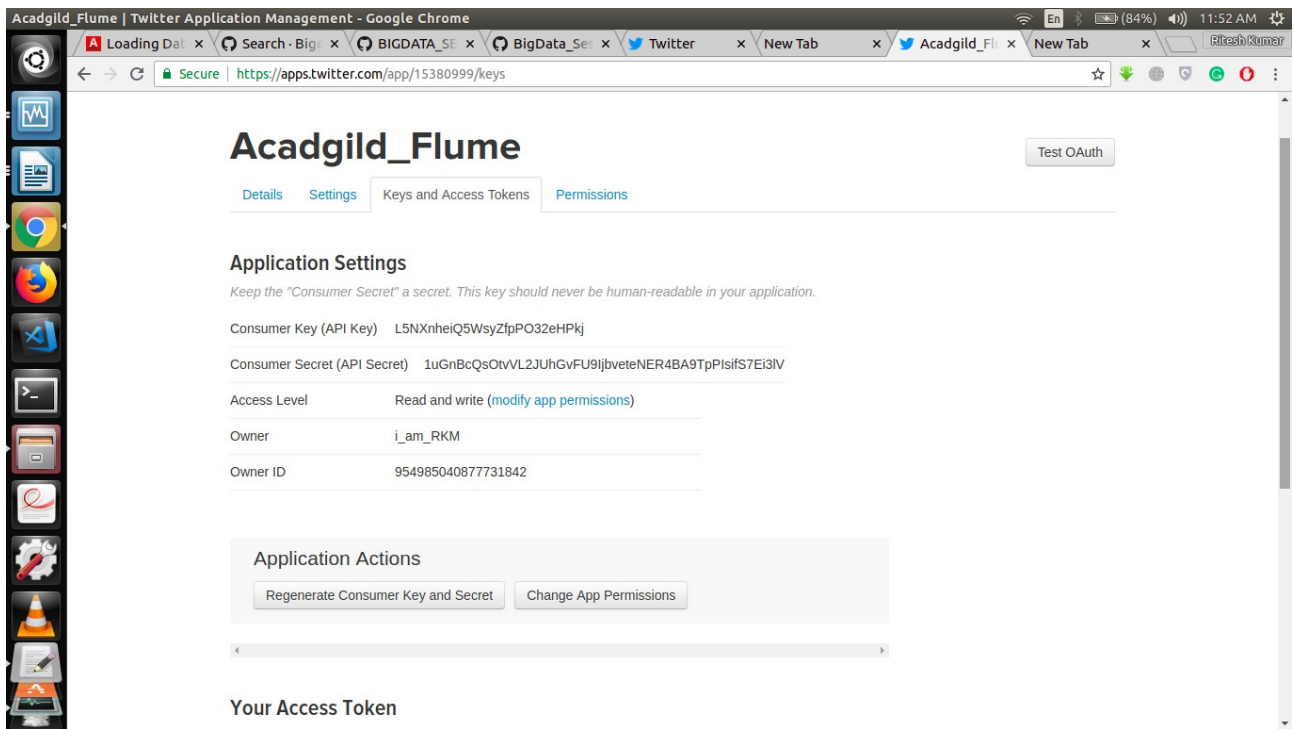
Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution created by your application and will be shown in user-facing authorization screens.
(If you don't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

Callback URLs

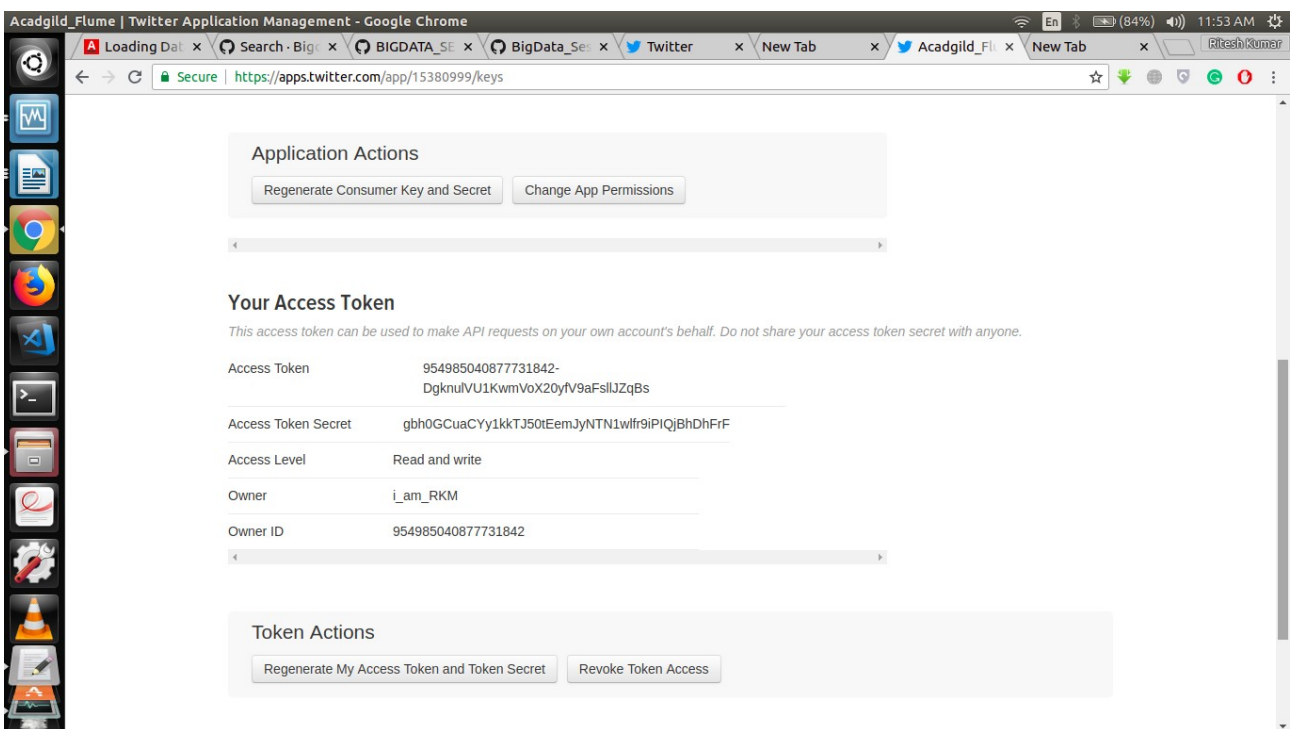
Where should we return after successfully authenticating? OAuth 1.0a applications must explicitly specify their oauth_callback URL(s) here, as well as include the one of the URLs below in the request token to restrict your application from using callbacks, leave this field blank.

[Add a Callback URL](#)

- Go to “Keys and Access Tokens” tab and copy the “Consumer Key” and “Consumer Secrete Code”.



- Click on the “Create My Access Token” button.
- Copy “Access Token” and Access Token Secret”.



- Copy the tokens in the flume configuration file where we have set the sources , search terms , channels and sink

```
acadgild.conf (~/.RITESH/4.5) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
acadgild.conf
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=LSNXnheiQ5WsyZfpP032eHPkj
TwitterAgent.sources.Twitter.consumerSecret=luGnBcQs0tvVL2JUhGvFU9IjbveteNER4BA9TpPisifs7Ei3lv
TwitterAgent.sources.Twitter.accessToken=95498504087773184J-DgknulVU1KwmVoX20yfV9aFslLJZqBs
TwitterAgent.sources.Twitter.accessTokenSecret=gbh0GCuaCyyIkktJ50tEemJyNTN1wlfr9iPIQjBhDhFrF
TwitterAgent.sources.Twitter.keywords=hadoop
# Describing/Configuring the sink

TwitterAgent.sources.Twitter.keywords= hadoop

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/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=1000

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

- Created a folder 'tweets' in hdfs.

hadoop fs -mkdir /tweets

```
acadgild@localhost:~/.RITESH/4.5
File Edit View Search Terminal Help
[acadgild@localhost 4.5]$ hadoop fs -mkdir /tweets
18/06/19 11:59:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost 4.5]$ hadoop fs -ls /
18/06/19 11:59:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 6 items
drwxr-xr-x - acadgild supergroup          0 2018-06-19 10:14 /hbase
drwxr-xr-x - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
-rw-r--r-- 1 acadgild supergroup    26204 2018-06-19 10:07 /student.txt
drwxrwx--- - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x - acadgild supergroup          0 2018-06-19 11:59 /tweets
drwxr-xr-x - acadgild supergroup          0 2018-02-09 14:50 /user
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost 4.5]$
```

- Then gave flume command to get the tweets to the hdfs using flume streaming.

flume ng-agent -n TwitterAgent -f acadgild.conf

```
acadgild@localhost:~/RITESH/4.5
File Edit View Search Terminal Help

[acadgild@localhost 4.5]$ flume-ng agent -n TwitterAgent -f acadgild.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.2.6/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/home/acadgild/install/hive/apache-hive-2.3.2-bin) for Hive access
+ exec /usr/java/jdk1.8.0_151/bin/java -Xmx20m -cp /home/acadgild/install/flume/apache-flume-1.8.0-bin/lib/*:/home/acadgild/
install/hadoop/hadoop-2.6.5/etc/hadoop/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/*:/home/acadgild/i
install/hadoop/hadoop-2.6.5/share/hadoop/common/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/*:/home/acadgild/
install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/*:/home/acad
gild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/*:/home
/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/ma
preduce/*:/home/acadgild/install/hadoop/hadoop-2.6.5/contrib/capacity-scheduler/*:/home/acadgild/install/hbase/hbase-1.2.
6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/hbase/hbase-1.2.6:/home/acadgild/install/hbase/hbase-1.2.
6/lib/activation-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopalliance-1.0.jar:/home/acadgild/install/hbase/hbase-1
.2.6/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home
/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-util-1.0.
0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/avro-1.7.4.ja
r:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commo
ns-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install/hbase/hba
se-1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-collections-3.2.2.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configuration-1
.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/co
mmons-digester-1.8.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hbase-1.2
.6/lib/commons-httpclient-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-io-2.4.jar:/home/acadgild/install/hbas
e/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-logging-1.2.jar:/home/acadgild/in
stall/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math3-3.1.1.jar:/home/a
cadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/disruptor-3.3.0.jar:/h
ome/acadgild/install/hbase/hbase-1.2.6/lib/finbugs-annotations-1.3.9-1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guav
a-12.0.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-se
rvlet-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-annotations-2.5.1.jar:/home/acadgild/install/hbase/hbase-1
.2.6/lib/hadoop-auth-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-client-2.5.1.jar:/home/acadgild/install/hba
se/hbase-1.2.6/lib/hadoop-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-hdfs-2.5.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-app-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapre
duce-client-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-core-2.5.1.jar:/home/acadgi
ld/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-jobclient-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/had
oop-mapreduce-client-shuffle-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-api-2.5.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/hadoop-yarn-client-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-common-2
.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-server-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1
```

```
acacgild@localhost:~/RITESH/4.5
File Edit View Search Terminal Help

ter, state: IDLE} ...
18/06/19 12:05:32 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfull
y registered new MBean.
18/06/19 12:05:32 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/06/19 12:05:32 INFO twitter.TwitterSource: Twitter source Twitter started.
18/06/19 12:05:32 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/06/19 12:05:38 INFO twitter4j.TwitterStreamImpl: Connection established.
18/06/19 12:05:38 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/06/19 12:05:38 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/06/19 12:05:39 INFO hdfs.BucketWriter: Creating hdfs://localhost:8029/tweets/FlumeData.1529390138973.tmp
18/06/19 12:05:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
18/06/19 12:05:42 INFO twitter.TwitterSource: Processed 100 docs
18/06/19 12:05:45 INFO twitter.TwitterSource: Processed 200 docs
18/06/19 12:05:49 INFO twitter.TwitterSource: Processed 300 docs
18/06/19 12:05:53 INFO twitter.TwitterSource: Processed 400 docs
18/06/19 12:05:56 INFO twitter.TwitterSource: Processed 500 docs
18/06/19 12:05:59 INFO twitter.TwitterSource: Processed 600 docs
18/06/19 12:06:03 INFO twitter.TwitterSource: Processed 700 docs
^C18/06/19 12:06:04 INFO lifecycle.LifecycleSupervisor: Stopping lifecycle supervisor 10
18/06/19 12:06:04 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider stopping
18/06/19 12:06:04 INFO twitter.TwitterSource: Twitter source Twitter stopping...
18/06/19 12:06:04 WARN twitter4j.TwitterStreamImpl: Stream already closed.
18/06/19 12:06:04 INFO twitter.TwitterSource: Twitter source Twitter stopped.
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel stopped
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.st
art.time == 1529390132554
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.st
op.time == 1529390164960
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.ca
pacity == 10000
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.cu
rrent.size == 0
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.ev
ent.put.attempt == 24
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.ev
ent.put.success == 24
18/06/19 12:06:04 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: CHANNEL, name: MemChannel. channel.ev
ent.take.attempt == 26
```

As it is visible in the screenshot , tweets are streaming into hdfs using flume from twitter source.

- Stopped the process after time using 'ctrl+d'

- Checked the tweets folder in hdfs and its contents

hadoop fs -ls /tweets

hadoop fs -cat /tweets/FlumeData.1529390138973



- As it is visible in the below screenshot , tweets are collected in the tweets folder in the hdfs.



The configuration file for flume (Directly can't upload in github due to safety)

acadgild.conf

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=L5NXnheiQ5Wsy
ZfpPO32eHPkj
TwitterAgent.sources.Twitter.consumerSecret=1uGnBcQsOtv
VL2JUHGvFU9IjbveteNER4BA9TpPIsifS7Ei3lV
TwitterAgent.sources.Twitter.accessToken=9549850408777318
42-DgknulVU1KwmVoX20yfV9aFslIJZqBs
TwitterAgent.sources.Twitter.accessTokenSecret=gbh0GCuaC
Yy1kkTJ50tEemJyNTN1wlfr9iPIQjBhDhFrF
TwitterAgent.sources.Twitter.keywords=hadoop
Describing/Configuring the sink**

TwitterAgent.sources.Twitter.keywords= hadoop

**TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/twe
ets
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
0**

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