

Task :

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

Go to Twitter account and create an application on Twitter apps

Save below Consumer Key, Consumer secret, Access Token, Access Token Secret and update into configuration file.

Status

Your application access token has been successfully generated. It may take a moment for changes you've made to reflect. [Refresh](#) if your changes are not yet indicated.

AcadgildGunjan

Test OAuth

[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

Application Settings

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

Consumer Key (API Key)	YhDqTZgsPnuW5yzyj45upxkPY
Consumer Secret (API Secret)	ZrKQ8fSMqjXkFLiugg6rN2yTbUcKtp5qRQMogSgK7zf17Ac4fi
Access Level	Read and write (modify app permissions)
Owner	GunjanArrora
Owner ID	851693541206106112

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	851693541206106112-eveP3PIw3xAhKQsXbN7Bg4sAP3dU3U6
Access Token Secret	8zrRxSc77JBQVrtBOnWnaigbQSpR5hqPS5kNCEptyKCo
Access Level	Read and write
Owner	GunjanArrora
Owner ID	851693541206106112

Token Actions

[Regenerate My Access Token and Token Secret](#) [Revoke Token Access](#)

➔ Check all jar files of twitter, whether they are available or not

Tip make sure below jars placed in \$FLUME_HOME/lib directory:

- ➔ *twitter4j-core-X.XX.jar*
- ➔ *twitter4j-stream-X.X.X.jar*
- ➔ *twitter4j-media-support-X.X.X.jar*

Configuration file: (Update Configuration file with latest key – values and put into VM-Flume path)

```
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=YhDqTZgsPnuW5yzyj45upxkP
Y
TwitterAgent.sources.Twitter.consumerSecret=ZrKQ8fSMqjXkFLiugg6rN
2yTbUcKtp5qRQMogSgK7zf17Ac4fi
TwitterAgent.sources.Twitter.accessToken=851693541206106112-
eveP3Plw3xAhKQsXbN7Bg4sAP3dU3U6
TwitterAgent.sources.Twitter.accessTokenSecret=
8zrRxSc77JBQVrtBOnWnaigbQSpR5hqPS5kNCEpctyKCo
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce,
mahout, hbase, nosql
# Describing/Configuring the sink

TwitterAgent.sources.Twitter.keywords= hadoop,election,sports,
cricket,Big data,gunjan,example

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/user/flum
e/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
```

Create directory on hdfs with names tweets.

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -mkdir -p /user/flume/tweets/
18/05/05 18:28:01 WARN util.NativeCodeLoader: Unable to load native hadoop lib
```

```

drwxr-xr-x - acadgild supergroup 0 2018-02-05 14:30 /user/flume
[acadgild@localhost ~]$ hadoop fs -ls /user/flume
18/05/05 10:22:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - acadgild supergroup 0 2018-05-05 10:20 /user/flume/tweet
s

```

For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

```

[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/acadgildTwitterFlume.conf

```

The above command will start fetching data from Twitter and streams it into the HDFS given path.

Streaming:

```

, code=-1, retryAfter=-1, rateLimitStatus=null, version=3.0.3}
7) at twitter4j.internal.http.HttpClientImpl.request(HttpClientImpl.java:17
ava:61) at twitter4j.internal.http.HttpClientWrapper.request(HttpClientWrapper.j
89) at twitter4j.internal.http.HttpClientWrapper.get(HttpClientWrapper.java:
6) at twitter4j.TwitterStreamImpl.getSampleStream(TwitterStreamImpl.java:17
at twitter4j.TwitterStreamImpl$4.getStream(TwitterStreamImpl.java:164)
at twitter4j.TwitterStreamImpl$TwitterStreamConsumer.run(TwitterStreamIm
pl.java:462)
18/05/05 10:26:57 INFO twitter4j.TwitterStreamImpl: Waiting for 160000 milliseco
nds

```

```

twitter.TwitterSource: Processed 100 docs
twitter.TwitterSource: Processed 200 docs
twitter.TwitterSource: Processed 300 docs
twitter.TwitterSource: Processed 400 docs
twitter.TwitterSource: Processed 500 docs

```

Once, the tweet data started streaming it into the given HDFS path we can use 'Ctrl+c' command to stop the streaming process

⇒ To check the contents of the tweet data we can use the following command:`hadoop fs -ls /user/flume/tweets`

```

INFO hdfs.BucketWriter: Creating hdfs://localhost:9000/user/flume/tweets/FlumeData.1523160169655.

```

Use the 'cat' command to display the tweet data inside the /user/flume/tweets/FlumeData.1523160169655 path using

hadoop dfs -cat /user/flume/tweets /FlumeData.1523160169655

```
licable
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int","null"]},{"name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"user_statuses_count","type":["int","null"]},{"name":"user_followers_count","type":["int","null"]},{"name":"user_name","type":["string","null"]},{"name":"user_screen_name","type":["string","null"]},{"name":"created_at","type":["string","null"]},{"name":"text","type":["string","null"]},{"name":"retweet_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]},{"name":"in_reply_to_user_id","type":["long","null"]},{"name":"source","type":["string","null"]},{"name":"in_reply_to_status_id","type":["long","null"]},{"name":"media_url_https","type":["string","null"]},{"name":
```