

Sqoop Flume

Assignment

Student Name: Subham Vishal

Course: Big Data Hadoop & Spark Training

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

Contents

ntroduction	2
roblem Statement	2
rerequisite	2
tep1:	3
tep2:	3
tep 3:	
tep4:	e
tep5:	
tep6:	7
ten7:	۶

ACADGILD



Introduction

In this assignment, we are going to streams data from twitter and stores into HDFS and the screen shots are shared.

Problem Statement

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

Prerequisite

To stream data to our database from twitter we should have the following pre-requisites.

- Twitter account
- Hadoop cluster

Make sure you have below jars placed in your \$FLUME_HOME/lib/conf directory:

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

```
[acadgild@localhost lib]$ ls -l | grep twitter
-rw-r--r-- 1 acadgild acadgild 14733 May 11 2015 flume-twitter-source-1.6.0.jar
-rw-r--r-- 1 acadgild acadgild 284077 Aug 23 2014 twitter4j-core-3.0.3.jar
-rw-r--r-- 1 acadgild acadgild 27698 Aug 26 2014 twitter4j-media-support-3.0.3.jar
-rw-r--r-- 1 acadgild acadgild 56307 Aug 23 2014 twitter4j-stream-3.0.3.jar
[acadgild@localhost lib]$
```

If the above prerequisites are available we can move to our further step.



Step1:

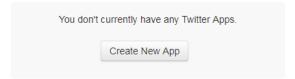
Login to the twitter account,

Step2:

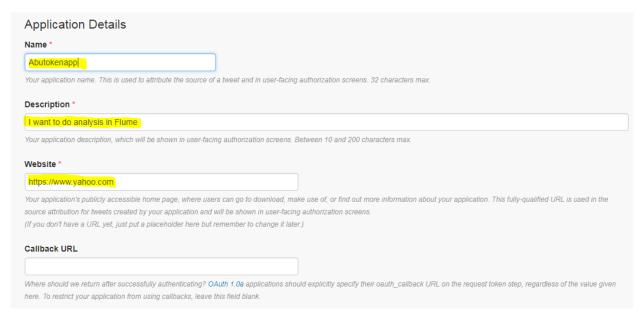
Go to the following link and click the 'create new app' button.

https://apps.twitter.com/app

Twitter Apps



Providing necessary details,



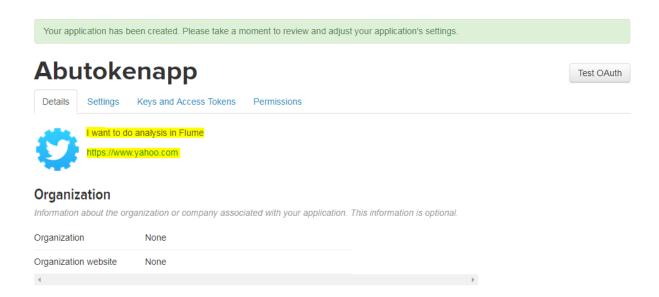
Accept the developer agreement and select the 'create your Twitter application' button'





	BIG DATA DEVELOPER	ACADGILI
	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 here. To restrict your application from using callbacks, leave this field blank.	ie value given
	Developer Agreement ✓ Yes, I have read and agree to the Twitter Developer Agreement.	
(Create your Twitter application	

Select the 'Keys and Access Token' tab.



Copy the consumer key and the consumer secret code, Scroll down further and select the 'create my access token' button.



Access Token Secret

Access Level

Owner

Owner ID

Abutokenapp Test OAuth Keys and Access Tokens **Application Settings** Keep the "Consumer Secret" a secret. This key should never be human-readable in your application. Consumer Key (API Key) DCjUjRSucocyREIvZQa6VJ5AP Consumer Secret (API Secret) x1D1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj Read and write (modify app permissions) Owner abarajitansa 797943092 Owner ID **Application Actions** Regenerate Consumer Key and Secret Change App Permissions 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** Create my access token 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 797943092-

Now, you will receive a message stating "that you have successfully generated your application access token".

wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYg1ETi2sOA3

ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsal4ln3lNVTAJJU

Read and write

abarajitansa

797943092



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.

Copy the Access Token and Access token Secret code.

```
Consumer Key (API Key) DCjUjRSucocyREIvZQa6VJ5AP

Consumer Secret (API Secret) x1D1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj

Access Token 797943092-wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYg1ETi2sOA3

Access Token Secret ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsal4In3INVTAJJU
```

Step 3:

Copy the Flume configuration code from the below link and paste it in the newly created file in the location,

/home/acadgild/apache-flume-1.6.0-bin/conf/flume_twitter.conf

https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWINidkk

Update the newly created file with twitter **api** keys like consumer key, Consumer token, Access token and the access token secret code and with the **key words**.

```
# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=DCjUjRSucocyREIvZQa6VJ5AP
TwitterAgent.sources.Twitter.consumerSecret=x1DlnQkXJHAhghTztK651917U9Taq4WL18fRqa9UUm5DCwYDVj
TwitterAgent.sources.Twitter.accessToken=797943092-wcNt3mgrbPiHYhEZK9RjWvjs3zAlYg1ETi2SOA3
TwitterAgent.sources.Twitter.accessToken=67979431431NVTAJJU
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink
TwitterAgent.sources.Twitter.keywords= hadoop,election,sports, cricket,Big data
```

Step4:

4.1 start all Hadoop daemons

ACADGILD



```
[acadgild@localhost lib]$ jps
3234 NodeManager
2819 DataNode
3125 ResourceManager
4661 Main
2712 NameNode
4315 HMaster
4107 RunJar
6172 Jps
[acadgild@localhost lib]$
```

Step5:

Create a new directory inside HDFS path, where the Twitter tweet data should be stored.

Hadoop dfs -mkdir /user/acadgild/hadoop/tweets

Step6:

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

flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.6.0-bin/conf/flume_twitter.conf

```
6172 Jps
[acadgild@localhost lib]$ flume-ng agent -n TwitterAgent -f /home/acadgild/apache-flume-1.6.0-bin/conf/flume_twitter.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/hadoop-2.7.2/bin/hadoop) for HDFS access
```

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





```
17/11/30 10:12:30 INFO hdfs.HDFSDataStream: Serializer = TEXĪ, UseRawLocalFileSystem = false
17/11/30 10:12:30 INFO hdfs.BucketWriter: creating hdfs://localhost:9000/user/acadgild/hadoop/tweets/FlumeData.1512010950366.tmp
Java HotSpot(TM) Client VM warning: You have loaded library /home/acadgild/hadoop.27.ib/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the stack guard now.
17/11/30 10:12:31 WARN util.MativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/11/30 10:12:31 INFO twitter.TwitterSource: Processed 100 docs
17/11/30 10:12:31 INFO twitter.TwitterSource: Processed 200 docs
17/11/30 10:12:32 INFO twitter.TwitterSource: Processed 300 docs
17/11/30 10:12:42 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:41 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:41 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:42 INFO twitter.TwitterSource: Processed 500 docs
17/11/30 10:12:45 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:53 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:55 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:12:55 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed 400 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Torcessed 1,000 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Torcessed 1,000 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Torcessed 1,000 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Run took 32 seconds and processed:
17/11/30 10:13:00 INFO twitter.TwitterSource: Run took 32 seconds and processed:
17/11/30 10:13:00 INFO twitter.TwitterSource: Run took 32 seconds and processed:
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed 1,100 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed 1,000 docs
17/11/30 10:13:00 INFO twitter.TwitterSource: Processed
```

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

Step7:

To check the contents of the tweet data we can use the following command:

hadoop fs -cat /user/acadgild/hadoop/tweets/FlumeData.1512016950366

```
| Jacob | Machago | S. - cat | Amer/scadgild/hadoop/twest/Flumbate.15120050306 |
Jave notSpect | Mr. Clear M. Warring; You have loaded library / home/acadgild/hadoop-2.7.2/ib/mative/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will try to fix the library with 'executed. < clifficer, or link it with '-r, neconcetack'.

17/11/20 10:18:15 AMB until mativeCodeLoader: Unable to load native-hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18:15 AMB until mativeCodeLoader: Unable to load native-hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18:15 AMB until mativeCodeLoader: Unable to load native-hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18:15 AMB until mativeCodeLoader: Unable to load native-hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... using bultim: jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your platform... unit jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for your jew classes where mphicable (17/11/20 10:18) and the load native hadoop library for
```

We can observe from the above image that we have successfully fetched twitter data into our HDFS cluster directory using Flume.