

TASK

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

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

- Twitter account
- Hadoop cluster

Since, both prerequisites are available so moving to following steps:

Step 1: Created Twitter App by logging into my twitter account and from there I copied below keys:

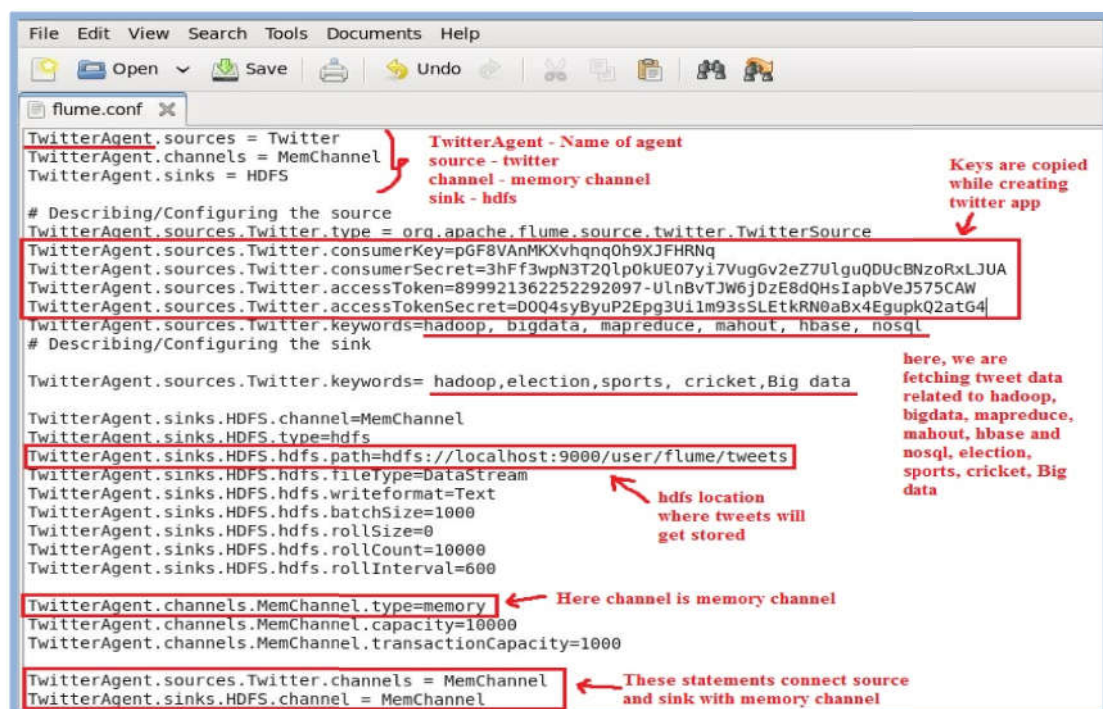
```
Cosumer Key (API Key) = pGF8VAnMKXvhqngOh9XJFHRNq
Consumer Secret (API Secret) = 3hFf3wpN3T2QlpOkUEO7yi7VugGv2eZ7UlgUQDUcBNzoRxLJUA
Access Token = 899921362252292097-UlnBvTJW6jDzE8dQHsIapbVeJ575CAW
Access Token Secret = DOQ4syByuP2Epg3UiIm93sSLEtkRN0aBx4EgupkQ2atG4
```

Step 2: Since here in this task we have to stream and analyse twitter data, therefore below jar files are checked inside `/$FLUME_HOME/lib` [i.e. `/usr/local/flume/lib`] directory:

```
[acadgild@localhost lib]$ ls -lrt t*.jar
-rw-r--r--. 1 acadgild acadgild 56307 Aug 23 2014 twitter4j-stream-3.0.3.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
```

Above screenshot shows that all three required jars are present inside specified directory.

Step 3: Created `flume.conf` file [using `gedit flume.conf` command] inside `/usr/local/flume/conf` with above specified twitter app keys:



Step 4: In new terminal started all the Hadoop daemons before running the flume command to fetch the twitter data. Using 'jps' command checked whether all required daemons have started or not.

```
[acadgild@localhost ~]$ $HADOOP_HOME/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/08/22 15:35:25 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-aca
dgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/had
oop-acadgild-secondarynamenode-localhost.localdomain.out
17/08/22 15:36:30 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-
resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-ac
adgild-nodemanager-localhost.localdomain.out
[acadgild@localhost ~]$ jps
5200 Jps
5140 NameNode
4600 NameNode
4698 DataNode
5038 ResourceManager
4863 SecondaryNameNode
[acadgild@localhost ~]$
```

Required daemons started

Step 5: Created new directory inside HDFS path, where the Twitter's tweet data would be stored:

```
[acadgild@localhost ~]$ hadoop fs -mkdir -p /user/flume/tweets
17/08/22 15:38:58 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[acadgild@localhost ~]$ hadoop fs -ls /user
17/08/22 15:39:16 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 9 items
drwxr-xr-x - acadgild supergroup 0 2017-08-22 02:12 /user/acadgild
drwxr-xr-x - acadgild supergroup 0 2017-08-22 15:39 /user/flume
drwxr-xr-x - acadgild supergroup 0 2017-08-21 23:30 /user/hive
drwxr-xr-x - acadgild supergroup 0 2017-08-22 13:45 /user/my_hive_tab
le
drwxr-xr-x - acadgild supergroup 0 2017-08-13 00:11 /user/my_pig_stuf
f
drwxr-xr-x - acadgild supergroup 0 2017-08-22 13:08 /user/my_sqoop
drwxr-xr-x - acadgild supergroup 0 2017-08-20 23:29 /user/oozie
drwxr-xr-x - acadgild supergroup 0 2015-11-08 17:35 /user/prateek
-rw-r--r-- 1 acadgild supergroup 26204 2017-08-21 13:34 /user/student.txt
[acadgild@localhost ~]$ hadoop fs -ls /user/flume
17/08/22 15:39:32 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x - acadgild supergroup 0 2017-08-22 15:39 /user/flume/tweet
$
[acadgild@localhost ~]$
```

Step 6: Used below command to fetch the twitter's tweet data into the HDFS cluster path:

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

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

```
17/08/22 15:54:00 INFO twitter4j.TwitterStreamImpl: Waiting for 250 milliseconds
17/08/22 15:54:01 INFO twitter4j.TwitterStreamImpl: Establishing connection.
17/08/22 15:54:03 INFO twitter4j.TwitterStreamImpl: Connection established.
17/08/22 15:54:03 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
17/08/22 15:54:07 INFO twitter.TwitterSource: Processed 800 docs
17/08/22 15:54:12 INFO twitter.TwitterSource: Processed 900 docs
17/08/22 15:54:16 INFO twitter.TwitterSource: Processed 1,000 docs
17/08/22 15:54:16 INFO twitter.TwitterSource: Total docs indexed: 1,000, total s
kipped docs: 0
17/08/22 15:54:16 INFO twitter.TwitterSource:      19 docs/second
17/08/22 15:54:16 INFO twitter.TwitterSource: Run took 52 seconds and processed:
17/08/22 15:54:16 INFO twitter.TwitterSource:      0.005 MB/sec sent to index
17/08/22 15:54:16 INFO twitter.TwitterSource:      0.262 MB text sent to index
17/08/22 15:54:16 INFO twitter.TwitterSource: There were 0 exceptions ignored:
17/08/22 15:54:23 INFO twitter.TwitterSource: Processed 1,100 docs
```

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

Step 7: To check whether file containing tweet data got created successfully inside hdfs or not, used the following command:

```
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
17/08/22 15:46:37 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--  1 acadgild supergroup    1815770 2017-08-22 15:46 /user/flume/tweet
s/FlumeData.1503396832990
[acadgild@localhost ~]$
```

Twitter data got stored inside this file

Step 8: Below **cat** command displays the tweet data inside the file created in above step:

```
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.1503396832990
0
```

しかし直後、1人の女性が全員を殺害！

Tweet Data

=>https://t.co/q3BE2shsVQ

地獄へ突き落とした理由がヤバイ...

https://t.co/vsPSX2Z641a href="https://twitter.com/" rel="nofollow">最新の話題
Tweethttps://pbs.twimg.com/media/DHxJsqMXUAEfVAB.jpghttps://twitter.com
/livenewstweet/status/899817008417562624/photo/1\$89993820872053145703

FranceB#Des
perateHousewives #PrisonBreak076ueule d'angeKatris Prideen(2017-08-22T15:46:03
Z0RT @proportionnelle: Sérieusement @AUCHAN France ? https://t.co/9ubHUtA2wk00
a href="http://twitter.com/download/android" rel="nofollow">Twitter for Androi
dhttps://pbs.twimg.com/media/DHxJsqMXUAEfVAB.jpghttps://twitter.com/prop
ortionnelle/status/899675187737710593/photo/10_09'00+J]0000
[acadgild@localhost ~]\$