Twitter Data Ingestion Using Flume

### Author:

**Vidur Nayyar**

# Content

Content 2

Pre-requisites 3

HDP 2.3.2 sandbox or higher 3

Maven 4

Java exe 4

Twitter Flume Source 5

Twitter Flume Configuration file 6

Hive DDL for Twitter Data 7

Reference 8

# Pre-requisites

Before we start we need to make sure that the following has been downloaded on your machine:

1. HDP 2.3.2 sandbox or higher
2. Maven
3. Java exe

## HDP 2.3.2 sandbox or higher

Download the latest HDP sandbox from the Hortonworks website. [[Hortonworks](http://hortonworks.com)]. You can download the HDP 2.3.2 sandbox from [HDP 2.3.2](http://hortonworks.com/products/hortonworks-sandbox/) . The installation guide can be downloaded from [Windows and Mac](http://hortonworks.com/wp-content/uploads/2015/07/Import_on_Vbox_7_20_2015.pdf).

Once you have downloaded the HDP sandbox, you can go to the vmbox to open the sandbox as a virtual machine. Please give the sandbox exactly half the amount of RAM memory that your computer has (this RAM allocated for the sandbox should not be less than 3 Gb). When you open the sandbox for the first time, you will need to change the password from Hadoop to a custom password.

You may log in to the sandbox by SSH into it. You will need to use the following script to ssh.

* ssh root@127.0.0.1 -p 2222

Use your local terminal/putty to provide in the new password, because the key might not be recognized in the virtual machine. (It might take a number of attempts to get this right. If you still are not able to login, please download the sandbox again.).

You may also get an error of the RSA key. In that case open the known key file in the .ssh directory . If you are using Mac and had added some other virtual machine on to the [127.0.0.1]2222 ip address you would need to open the .ssh directory by typing this in the terminal:

* Open .ssh

Once then open the known key file and delete the rsa key for the [127.0.0.1]2222 ip address.

**Setting the permissions for the sandbox**

You would need to set the permissions for the root directory by following the following steps:

* [hdfs@sandbox root]$ hdfs dfs -mkdir /user/root
* [hdfs@sandbox root]$ hdfs dfs -chown root:root /user/root
* [hdfs@sandbox root]$ hdfs dfs -chmod -R 777 /user/root
* [hdfs@sandbox root]$ hdfs dfs -chmod -R 777 /user/root

## Maven

Apache Maven is a software project management and comprehension tool. Based on the concept of a project object model (POM), Maven can manage a project's build, reporting and documentation from a central piece of information.

To know more about Maven, you can look for the "About Maven" section of the navigation of [Maven](https://maven.apache.org). This includes an in-depth description of [what Maven is](https://maven.apache.org/what-is-maven.html), a [list of some of its main features](https://maven.apache.org/maven-features.html), and a set of [frequently asked questions about what Maven is](https://maven.apache.org/general.html).

Maven is used to recompile jars that are otherwise corrupted. Maven would download all the imported libraries and link your java code to them. You will need to add the maven code of the libraries to the pop file for maven.

You would need to define your java directory for mavin to work, you can do that by the following script:

* export JAVA\_HOME=$(/usr/lib/jvm/java-1.7.0-openjdk.x86\_64/jre)

To download maven you need to do the following:

* sudo wget http://repos.fedorapeople.org/repos/dchen/apache-maven/epel-apache-maven.repo -O /etc/yum.repos.d/epel-apache-maven.repo
* sudo sed -i s/\$releasever/6/g /etc/yum.repos.d/epel-apache-maven.repo
* sudo yum install -y apache-maven

To check if maven has been downloaded do the following:

* mvn –version

Maven is used to recompile jars that are otherwise corrupted(you don’t need to do this, this needs to be done only if the jar file is corrupted. I will make sure that the jar file provided to you is never corrupted. In case it ever is you may recompile using this)

* put jar file into /usr/hdp/

## Java exe

Please refer the following link [Java for sandbox](http://tecadmin.net/steps-to-install-java-on-centos-5-6-or-rhel-5-6/)

# Twitter Flume Source

Before starting with the flume configuration file, I would advise you to create a folder in your sandbox to place the configuration file and all the files linked to twitter ingestion files.

You can do so by:

* mkdir TwitterData

Use the WinSCP to transfer the files that you have downloaded from the hexacorpadmin GitHub account.

Transfer all the contents into a new local directory named TwitterData in local.

You will find a ***TwitterFlume-Source-T4J*** JAR file. This jar file is the twitter flume source jar file.

You have to move that jar file into /usr/hdp/2.3.2.0-2950/flume/lib

once downloaded, you would need to move the downloaded jar file into the /usr/hdp/2.3.2.0-2950/flume/lib folder in your sandbox. You can do this by running the following command:

* cp TwitterFlume-Source-T4J /usr/hdp/2.3.2.0-2950/flume/lib

# Twitter Flume Configuration file

Before beginning with Flume Ingestion, it is necessary to update your system time according to the flume service so as to avoid any from Twitter in later steps (for Ex – during oozie workflow where you provided date as parameters).

* To do this we use the NTP (Network Time Protocol), which is the time protocol that most servers in the world use to keep in sync.
* Run the below steps in your terminal to install NTP
* yum install -y ntp
* service ntpd stop
* ntpdate pool.ntp.org
* service ntpd start

Now let’s start with the configuration file for flume.

Look for the **.conf** file in the downloaded repository file. This is the configuration file. The configuration file contains all the information about the fields you would need to populate. You will need to provide your credentials, that is the customer id, secret, access token and token secret. You can get these by following the steps given here <https://dev.twitter.com/oauth/overview/application-owner-access-tokens> you can make your own app from the following <https://apps.twitter.com>

Once you get the credentials, replace the code in the .conf file like the following :

* *TwitterAgent.sources.Twitter.consumerKey = pfc3OsyN88gAW4Px4gyHWTBF3*
* *TwitterAgent.sources.Twitter.consumerSecret = 6z5gmQ44GLyH5Y6s0qWIwqk0TPiL6bE8iboqOlzWTUv9Cod6WK*
* *TwitterAgent.sources.Twitter.accessToken = 2992840146-3kPFX7z6o0PxkRGvAFeSrD9f8px4JLX3IIefhdg*
* *TwitterAgent.sources.Twitter.accessTokenSecret = EGXeQoqbvI23kUotitnbl6ddMu8pvWYUmW3DTcxcanNYx*

You may change the path where the tweets will be dumped in the hdfs by changing the following code:

* TwitterAgent.sinks.HDFS.hdfs.path = hdfs://sandbox.hortonworks.com:8020/tmp/flume/twitter/hotel/%Y/%m/%d

By default the path is set to:

* hdfs://sandbox.hortonworks.com:8020/tmp/flume/twitter/hotel/%Y/%m/%d

Once you have done the above, you are all set to ingest the twitter data.

Now run the flume agent by executing the following command on your terminal:

* cd TwitterData
* /usr/bin/flume-ng agent --conf ./conf/ -f /root/TwitterData/TwitterFlume.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent

# Hive DDL for Twitter Data

You will notice that the twitter repository that you downloaded has two more folders these folders contain support files for sentiment analysis of the twitter data.

It is advisable to create a new folder in HDFS where you can put the support files for the twitter sentiment analysis. This can be done as follows:

* hadoop fs –mkdir /user/root/TwitterData

Now copy the supporting files, that is the dictionary and the location files into the new folder that you created in hdfs. Do the following.

* Hadoop fs –cp dictionary /user/root/TwitterData/dictionary
* Hadoop fs –cp time\_zone\_map /user/root/TwitterData/time\_zone\_map

Please make sure that these files are placed in these folders because these files are used by the hive table formation.

Note that the repository also contains a jar file for serde. It is very important to add this jar file , This needs to be done just once, that is the first time the code is executed. It can be done by:

* hive

Once the hive shell has started past the code given below. You have to copy the below JAR into the */tmp* directoryin hdfs and then run the below command in hive.

* ADD JAR json-serde-1.1.6-SNAPSHOT-jar-with-dependencies.jar;

Now you are all set to run the Hive DDL script given in the repository folder that you downloaded.

You might need to change the Location of the first table that is created, this table is the **tweets\_raw** table. The location of this table will correspond with the location of the ingested data. You may also need to change the location of the **dictionary** table that will correspond to the place where you have put the dictionary in hdfs. You may also need to change the location of the **time\_zone\_map** table that will correspond to the place where you have put the time\_zone\_map in hdfs.

Everything else that might need to be tweaked has been commented in the ddl file.

# Reference

1. <http://www.thecloudavenue.com/2013/03/analyse-tweets-using-flume-hadoop-and.html>
2. <https://gist.github.com/sebsto/19b99f1fa1f32cae5d00>
3. <https://github.com/cloudera/cdh-twitter-example>
4. <http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-sentiment-data/>
5. <https://github.com/cloudera/cdh-twitter-example/archive/master.zip>
6. <https://dev.twitter.com/oauth/overview/application-owner-access-tokens>
7. <https://github.com/vidur89/hadoop-repository/tree/master/twitter-repository>