**Twitter Data fetching Using Hadoop and Flume**

There arefollowing steps for apache flume installation and fetch Twitter data to store it in HDFS.

**Step:1** Download Flume and extract it with following command:

$ tar -xvzf apache-flume-1.9.0-bin.tar.gz

**Step:2** Download the flume-sources-1.0-SNAPSHOT.jar from following link:

<https://drive.google.com/file/d/0B_t6uqPmWadsdWJNQ0NjaXBUYUk/view?usp=sharing>

**Step:3** Paste flume-sources-1.0-SNAPSHOT.jar in flume lib folder

cp flume-sources-1.0-SNAPSHOT.jar $FLUME\_HOME/lib

**Step:4** Add it to the flume class path in the conf/flume-env.sh file and add the java home path also in the same.

export JAVA\_HOME=/usr/lib/jvm/java-11-openjdk-amd64

FLUME\_CLASSPATH=“/home/vaishali/apache-flume-1.9.0-bin/lib/flume-sources-1.0-SNAPSHOT.jar”

**Step:5** Download consumerKey, consumerSecret, accessTokenSecret from https://apps.twitter.com/ which can be accessed from your twitter developer account by creating a simple app.

**Step:6** Create flume-twitter.conf file in conf folder and paste given lines:

Graphical user interface, text

Description automatically generated

**Step:7** Add the .bashrc file

$ gedit ~/.bashrc

Graphical user interface, text

Description automatically generated

**Step:8** Run following command to reload new .bashrc file

$source ~/.bashrc

**Step:9** Rename these 3 files in lib folder of Flume. (All you need to do just change the extention of these files from .jar to .org)

twitter4j-core-3.0.3.jar twitter4j-media-support-3.0.3.jar twitter4j-stream-3.0.3.jar

to

twitter4j-core-3.0.3.org twitter4j-media-support-3.0.3.org twitter4j-stream-3.0.3.org

**Step:10** Make Directories in HDFS using following Command:

$ hdfs dfs –mkdir /user/flume/tweets

**Step:11** Run Flume and collect data into HDFS.

Starting Hadoop initially using command:

$ Start-all.sh

Now,

Start flume using the below command

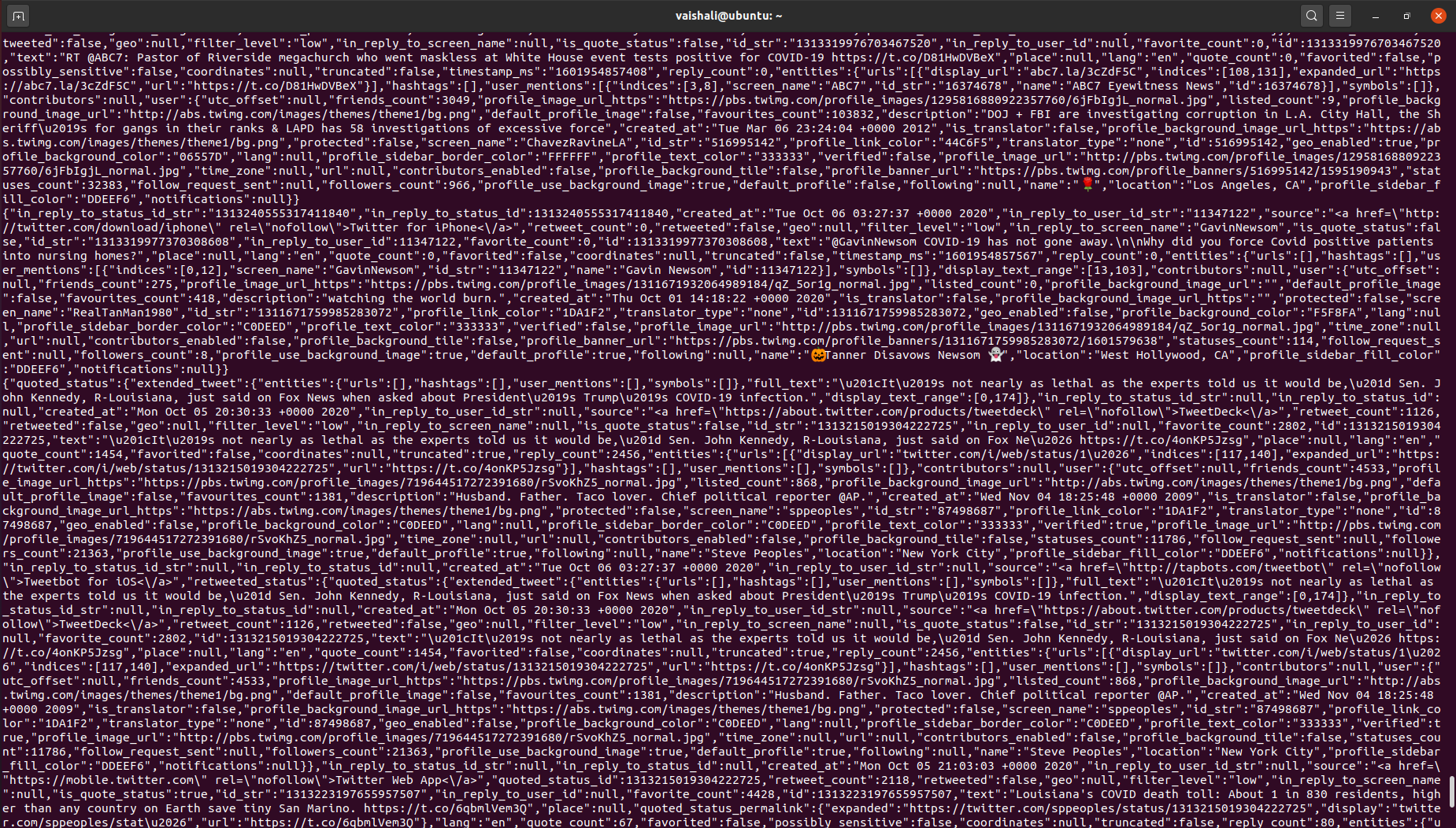
(Assuming you are inside ‘/home/vaishali/apache-flume-1.9.0-bin’)

bin/flume-ng agent -n TwitterAgent --conf ./conf/-f --conf-file conf/flume-twitter.conf -Dflume.root.logger=DEBUG,console

After a couple of minutes the Tweets should appear in HDFS.

**Step:12** To check the content of Twitter streamed data use the command:

$ hdfs dfs -cat /user/flume/tweets/txt\_file



This will show us the tweets done by people on covid-19.