\_\_\_\_\_\_\_\_\_GETTING TWEETS ON HDFS APACHE FLUME \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1--) Sign in twitter account and visit : apps.twitter.com

The first step is to create an application in https://apps.twitter.com and then generate the corresponding keys

2--) Assuming that Hadoop has already been installed and configured, the next step is download Flume and extract it to any folder.

: wget http://archive.apache.org/dist/flume/1.4.0/apache-flume-1.4.0-bin.tar.gz

: tar -zxvf apache-flume-1.4.0-bin.tar.gz

3--) Download the flume-sources-1.0-SNAPSHOT.jar and add it to the flume class path as shown below in the conf/flume-env.sh.template file

: wget http://files.cloudera.com/samples/flume-sources-1.0-SNAPSHOT.jar

: nano apache-flume-1.4.0-bin/conf/flume-env.sh.template

add flume classpath :-- FLUME\_CLASSPATH="/home/ubuntu/flume-sources-1.0-SNAPSHOT.jar"

4--) Rename flume-env.sh.template to flume-env.sh

sudo mv /home/ubuntu/apache-flume-1.4.0-bin/conf/flume-env.sh.template /home/ubuntu/apache-flume-1.4.0-bin/conf/flume-env.sh

5--) The jar contains the java classes to pull the Tweets and save them into HDFS.

make file named flume.conf in /home/ubuntu/apache-flume-1.4.0-bin/conf paste all the agents (flume, memory and hdfs) defined as below in flume.conf

:

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = YOUR\_CONSUMER\_KEY

TwitterAgent.sources.Twitter.consumerSecret = YOUR\_CONSUMER\_SECRET

TwitterAgent.sources.Twitter.accessToken = YOUR\_ACCESS\_TOKEN

TwitterAgent.sources.Twitter.accessTokenSecret = YOUR\_ACCESS\_TOKEN\_SECRET

TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientist, business intelligence, mapreduce

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://master:8020/twitter/data

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 = 100

6--) The consumerKey, consumerSecret, accessToken and accessTokenSecret have to be replaced with those obtained from https://dev.twitter.com/apps. And, TwitterAgent.sinks.HDFS.hdfs.path should

point to the NameNode and the location in HDFS where the tweets will go to.The TwitterAgent.sources.Twitter.keywords value can be modified to get the tweets for some other topic like football,

criket etc.

7--) Start flume using the below command

: cd /home/ubuntu/apache-flume-1.4.0-bin/bin

: ./flume-ng agent --conf /home/ubuntu/apache-flume-1.4.0-bin/conf/ -f /home/ubuntu/apache-flume-1.4.0-bin/conf/flume.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent