Lab twitter data2

```
=====================================================================================================================
```

```
=====================================================================================================================

Use this command reference file to copy and paste text for your lab.

=====================================================================================================================
Instructions:This document explains how to stream  twitter data through flume.


#Download twitter jar through  below  command:
git clone https://github.com/git-bigdatatrunk/Big-Data-Internship-Program-DataIngestion-Sqoop-and-Flume.git

ls

cd Big-Data-Internship-Program-DataIngestion-Sqoop-and-Flume/twitter_jar/

#Copy all below jar on /usr/lib/flume-ng/lib/  directory through cp command

sudo cp flume-sources-1.0-SNAPSHOT.jar /usr/lib/flume-ng/lib
sudo cp twitter4j-core-3.0.3.jar /usr/lib/flume-ng/lib
sudo cp twitter4j-media-support-3.0.3.jar /usr/lib/flume-ng/lib
sudo cp twitter4j-stream-3.0.3.jar /usr/lib/flume-ng/lib
sudo cp flume-ng-core-1.7.0.jar /usr/lib/flume-ng/lib

#Copy twitter.conf file on /etc/flume-ng/conf directory through below command
sudo cp twitter.conf /etc/flume-ng/conf

#Open new terminal and go to below directory
cd /etc/flume-ng/conf

#Open new terminal and Check all jar available /usr/lib/flume-ng/lib/ directory
cd /usr/lib/flume-ng/lib/

#Create twitter app and got the key
#Login to https://apps.twitter.com/ and creted consumer key, consumer secret(api secret),Access token and Access token scret.

#Go  to root directory
su root
cloudera
cd /etc/flume-ng/conf
ls
mv flume-env.sh.template flume-env.sh
gedit flume-env.sh
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License.  You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# If this file is placed at FLUME_CONF_DIR/flume-env.sh, it will be sourced
# during Flume startup.

# Enviroment variables can be set here.

export JAVA_HOME=/usr/java/jdk1.7.0_67-cloudera


# Give Flume more memory and pre-allocate, enable remote monitoring via JMX
export JAVA_OPTS="-Xms100m -Xmx2000m -Dcom.sun.management.jmxremote"

# Note that the Flume conf directory is always included in the classpath.
FLUME_CLASSPATH=" /usr/lib/flume-ng/lib/*"


#Twitter.conf file available on /etc/flume-ng/conf directory
gedit twitter.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=<consumerKey>
TwitterAgent.sources.Twitter.consumerSecret=<consumerSecret>
TwitterAgent.sources.Twitter.accessToken=<accessToken>
TwitterAgent.sources.Twitter.accessTokenSecret= <accessTokenSecret>

TwitterAgent.sources.Twitter.keywords= hadoop,election,sports,cricket,Big data,News

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/user/cloudera/twitter_data/%Y-%m-%d-%H-%M
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=100000
TwitterAgent.channels.MemChannel.transactionCapacity=1000

#Run the below command and stream the twitter data

flume-ng agent -c /etc/flume-ng/conf -f /etc/flume-ng/conf/twitter.conf -n TwitterAgent -Dflume.root.logger=INFO,console
#Check twitter data in terminal:

hadoop fs -ls hdfs://localhost:8020/user/cloudera/

#Twitter data stored on /user/cloudera/twitter_data/ directory

hadoop fs -ls hdfs://localhost:8020/user/cloudera/twitter_data/
*******************************************************************
#login to flume conf directory
su roo
Password-cloudera
cd /etc/flume-ng/conf
 cat > TwitterDataAvroSchema.avsc

{"type":"record",
 "name":"Doc",
 "doc":"adoc",
 "fields":[{"name":"id","type":"string"},
           {"name":"user_friends_count","type":["int","null"]},
           {"name":"user_location","type":["string","null"]},
           {"name":"user_description","type":["string","null"]},
           {"name":"user_statuses_count","type":["int","null"]},
           {"name":"user_followers_count","type":["int","null"]},
           {"name":"user_name","type":["string","null"]},
           {"name":"user_screen_name","type":["string","null"]},
           {"name":"created_at","type":["string","null"]},
           {"name":"text","type":["string","null"]},
           {"name":"retweet_count","type":["long","null"]},
           {"name":"retweeted","type":["boolean","null"]},
           {"name":"in_reply_to_user_id","type":["long","null"]},
           {"name":"source","type":["string","null"]},
           {"name":"in_reply_to_status_id","type":["long","null"]},
           {"name":"media_url_https","type":["string","null"]},
           {"name":"expanded_url","type":["string","null"]}
          ]
}


cat > avrodataread.hql

drop table tweets;
CREATE TABLE tweets
  ROW FORMAT SERDE
     'org.apache.hadoop.hive.serde2.avro.AvroSerDe'
  STORED AS INPUTFORMAT
     'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'
  OUTPUTFORMAT
     'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'
  TBLPROPERTIES ('avro.schema.url'='file:///etc/flume-ng/conf/TwitterDataAvroSchema.avsc') ;

LOAD DATA INPATH '/user/cloudera/twitter_data/*/FlumeData.*' OVERWRITE INTO TABLE tweets;

hive - f avrodataread.hql

#Login to hive terminal through below command

hive

desc tweets;
```