Assignment Day 12

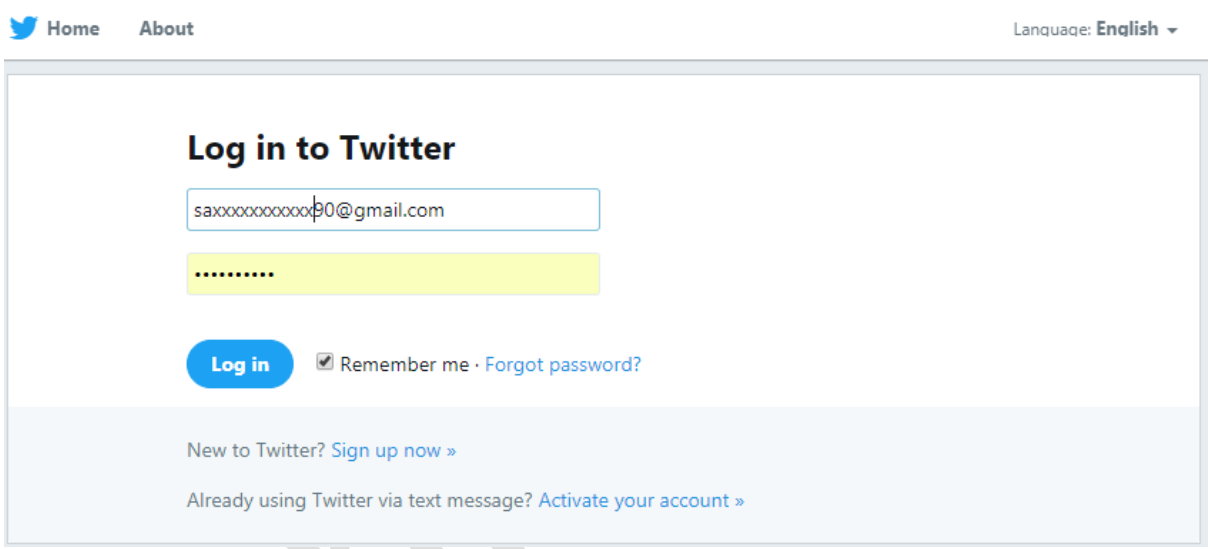
Task 1:

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

Ans:

Step 1:

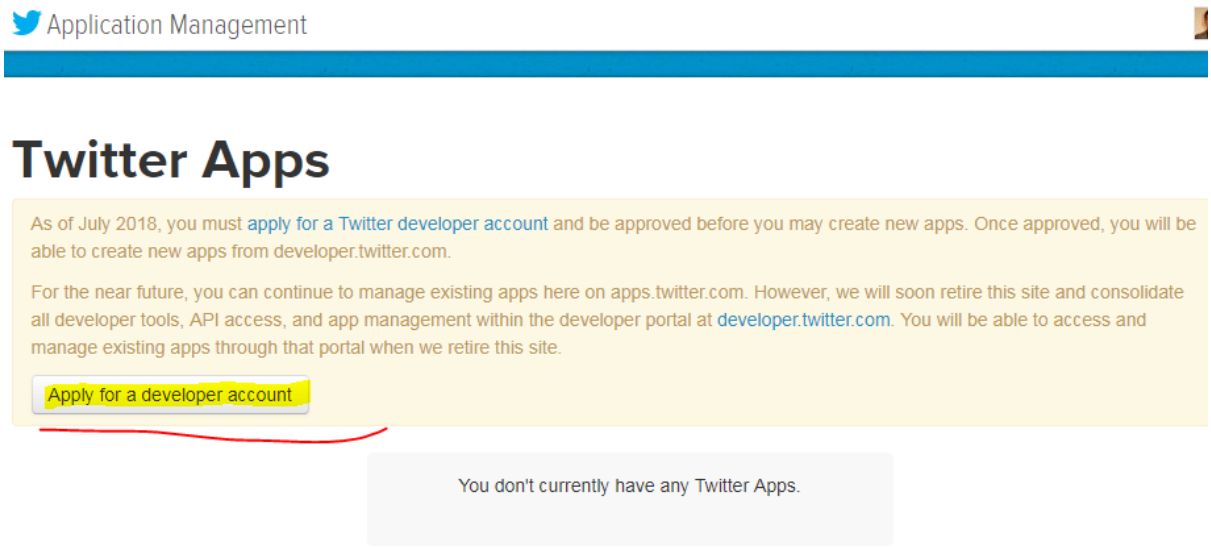
Login to the twitter account



Step 2:

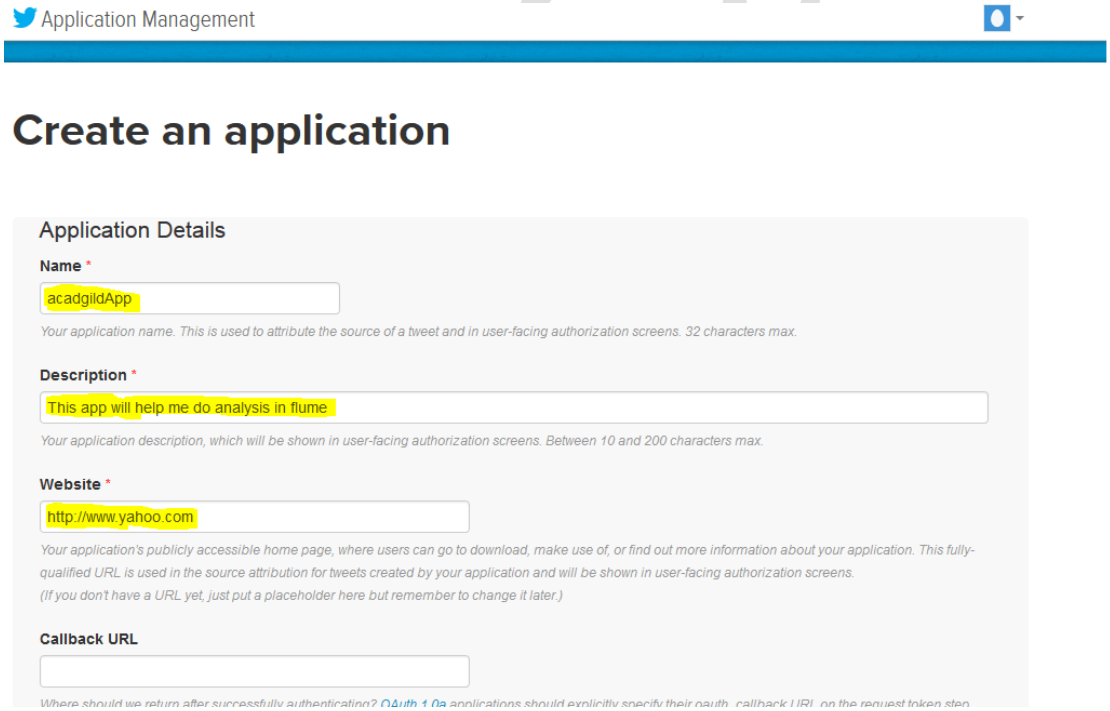
Go to the following link and click the 'Apply for Developer Account' button.

https://apps.twitter.com/app



Step 3:

Enter all the fields as below:



Step 4:

Accept the Developer agreement.

regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Developer Agreement

Effective: May 18, 2015.

This Twitter Developer Agreement ("**Agreement**") is made between you (either an individual or an entity, referred to herein as "**you**") and Twitter, Inc. and Twitter International Company (collectively, "**Twitter**") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("**EFFECTIVE DATE**").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT

☐ Yes, I agree

Create your Twitter application

Note: After few weeks your developer account would be approved.

Step 5:

Create a new flume.conf file & copy the Flume configuration code from the below link and paste it in the newly created file flume.conf.

https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk

Step 6:

You would receive consumerKey, consumerSecret, accessToken, accessTokenSecret from twitter once developer account is approved.

Copy these four values within **flume.conf** file as highlighted below.

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=DCjUjRSucocyREIvZQa6VJ5AP
TwitterAgent.sources.Twitter.consumerSecret=x1D1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj
TwitterAgent.sources.Twitter.accessToken=797943092-wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYglETi2sOA3
TwitterAgent.sources.Twitter.accessTokenSecret=ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsaI4In3lNVTAJJU
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink

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

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/home/acadgild/Desktop/TestHadoop/flume/tweets
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=1000

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Step 7:

Within the same flume.conf file enter the keywords that you want to search the tweets on twitter against the key **TwitterAgent.sources.Twitter.keywords**.

TwitterAgent.sources.Twitter.keywords= hadoop, bigdata, mapreduce, mahout, hbase, nosql

Step 8:

Create a new directory tweets which would store tweets stream by flume agent on to HDFS:

hadoop fs -mkdir -p /hadoopdata/flume/tweets

```
[acadgild@10 tweets]$ hadoop fs -mkdir -p /hadoopdata/flume/tweets
18/08/19 18:32:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicab
le
[acadgild@10 tweets]$ hadoop fs -ls /hadoopdata/flume/
18/08/19 18:32:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicab
le
Found 1 items
drwxr-xr-x   - acadgild supergroup          0 2018-08-19 18:32 /hadoopdata/flume/tweets
[acadgild@10 tweets]$
```

Step 9:

Mention the newly created directory path into the flume.conf as shown below:

TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/hadoopdata/flume/tweets

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=DCjUjRSucocyREIvZQa6VJ5AP
TwitterAgent.sources.Twitter.consumerSecret=xlD1nQkXJHAhghTztK6519I7U9Taq4WLl8fRqa9UUm5DCwYDVj
TwitterAgent.sources.Twitter.accessToken=797943092-wcNt3mgrbPiHYhEZ2K9RjWvjs3zAlYg1ETi2sOA3
TwitterAgent.sources.Twitter.accessTokenSecret=ohm8hds3X1d2S0JWsOaAu3HlpTjYvSsaI4In3lNVTAJJU
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql
# Describing/Configuring the sink

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

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/hadoopdata/flume/tweets
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=1000

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Note: Make sure all the daemons are started:

```
[acadgild@10 tweets]$ jps
29696 DataNode
30337 NodeManager
29571 NameNode
30228 ResourceManager
29973 SecondaryNameNode
31386 HMaster
5771 Jps
31484 HRegionServer
31294 HQuorumPeer
[acadgild@10 tweets]$
```

Step 10:

For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.

flume-ng agent -n TwitterAgent -f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/flume.conf

```
[acadgild@10 ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/flume.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop) for HDFS access
Info: Including HBASE libraries found via (/home/acadgild/install/hbase/hbase-1.2.6/bin/hbase) for HBASE access
Info: Including Hive libraries found via (/home/acadgild/install/hive/apache-hive-2.3.2-bin) for Hive access
+ exec /usr/java/jdk1.8.0_151/bin/java -Xmx20m -cp '/home/acadgild/install/flume/apache-flume-1.8.0-bin/lib/*:/home/acadgild/install/hadoop/hadoo
p-2.6.5/etc/hadoop/:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/
common/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/lib/*:/home/ac
adgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/lib/*:/home/acadgild/install/
hadoop/hadoop-2.6.5/share/hadoop/yarn/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/*:/home/acadgild/install/hadoop/had
oop-2.6.5/share/hadoop/mapreduce/*:/home/acadgild/install/hadoop/hadoop-2.6.5/contrib/capacity-scheduler/*.jar:/home/acadgild/install/hbase/hbase
-1.2.6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/hbase/hbase-1.2.6:/home/acadgild/install/hbase/hbase-1.2.6/lib/activation
-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopalliance-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-i18n-2.0.0-M15.ja
r:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.
0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-util-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/ac
adgild/install/hbase/hbase-1.2.6/lib/avro-1.7.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/insta
ll/hbase/hbase-1.2.6/lib/commons-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install
/hbase/hbase-1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-collections-3.2.2.jar:/home/acadgild/install/hb
ase/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configuration-1.6.jar:/home/acadgild/install/
hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-digester-1.8.jar:/home/acadgild/install/hbas
e/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-httpclient-3.1.jar:/home/acadgild/install/hbase/hbase-1
.2.6/lib/commons-io-2.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/common
s-logging-1.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math3-3.
1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/disruptor-3.3.0.jar:/home/a
cadgild/install/hbase/hbase-1.2.6/lib/findbugs-annotations-1.3.9-1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guava-12.0.1.jar:/home/acadgi
ld/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-servlet-3.0.jar:/home/acadgild/install/hbase/hb
ase-1.2.6/lib/hadoop-annotations-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-auth-2.5.1.jar:/home/acadgild/install/hbase/hbase-
1.2.6/lib/hadoop-client-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/l
ib/hadoop-hdfs-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-app-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.
2.6/lib/hadoop-mapreduce-client-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-core-2.5.1.jar:/home/acadgi
ld/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-jobclient-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client
-shuffle-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-api-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yar
n-client-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-
yarn-server-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-annotations-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/li
b/hbase-annotations-1.2.6-tests.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-client-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/
lib/hbase-common-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-common-1.2.6-tests.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib
/hbase-examples-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-external-blockcache-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2
```

The Streaming starts.

To stop streaming press ctrl c.

Step 11:

To check the contents of the tweet go to the output directory at hdfs:

hadoop fs -ls /hadoopdata/flume/tweets

********** **End** **********