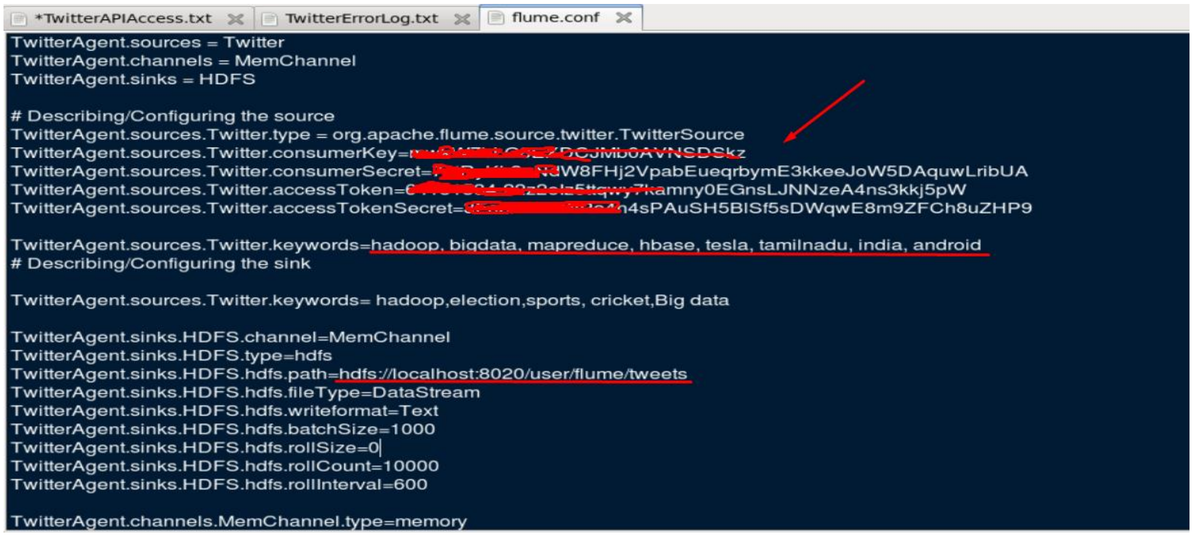
Session 12:
Oozie and Flume
Assignment 1

TASK:

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

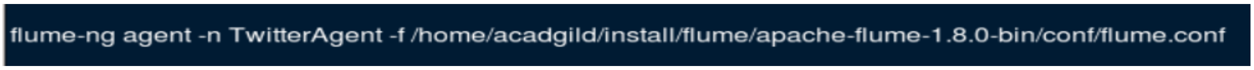
EXECUTION:

- As mentioned in the link, created new Twitter app, generated both consumer key & secret and access token and secret.
- Then configured the key and secret in flume.conf file in local file system. Also updated some of the keywords to search in twitter tweets. I have masked the key and secret in the below screenshot for the matter of privacy.
- Then configured the hdfs file system path where the output file to be stored.



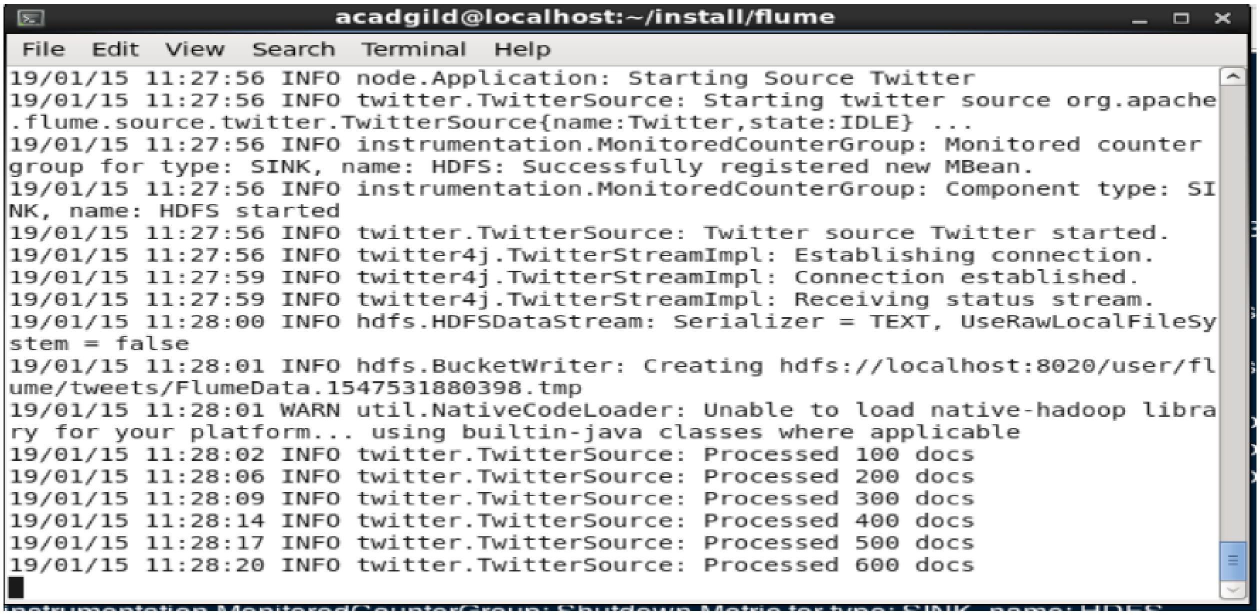
```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey=[redacted]
TwitterAgent.sources.Twitter.consumerSecret=[redacted]W8FHj2VpabEueqrbymE3kkeeJoW5DAquwLribUA
TwitterAgent.sources.Twitter.accessToken=[redacted]amny0EGnsLJNNzeA4ns3kkj5pW
TwitterAgent.sources.Twitter.accessTokenSecret=[redacted]n4sPAuSH5BISf5sDWqwE8m9ZFCh8uZHP9

TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, hbase, tesla, tamilnadu, india, android
# 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:8020/user/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
```

- Then executed the flume command to process twitter tweets and store the output in mentioned HDFS file system path,



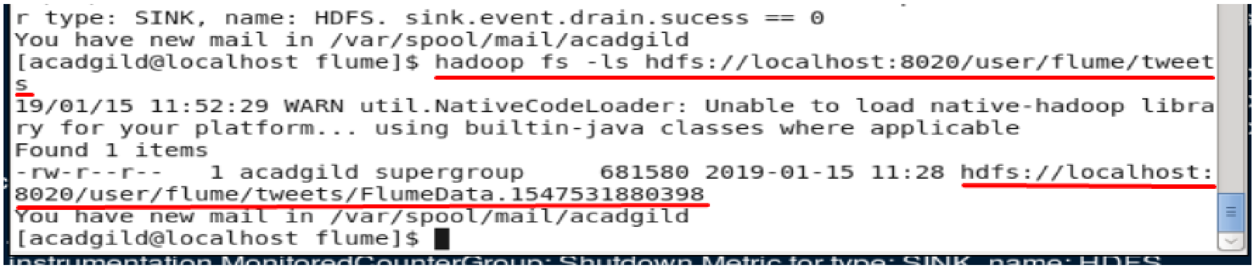
```
flume-ng agent -n TwitterAgent -f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/flume.conf
```

- Below is the execution sequence of command, in which it established the connection with Twitter API and creating output file in HDFS file system. Then tweets are getting pulled, processed from twitter source.



acadgild@localhost:~/install/flume
File Edit View Search Terminal Help

```
19/01/15 11:27:56 INFO node.Application: Starting Source Twitter
19/01/15 11:27:56 INFO twitter.TwitterSource: Starting twitter source org.apache
.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} ...
19/01/15 11:27:56 INFO instrumentation.MonitoredCounterGroup: Monitored counter
group for type: SINK, name: HDFS: Successfully registered new MBean.
19/01/15 11:27:56 INFO instrumentation.MonitoredCounterGroup: Component type: SI
NK, name: HDFS started
19/01/15 11:27:56 INFO twitter.TwitterSource: Twitter source Twitter started.
19/01/15 11:27:56 INFO twitter4j.TwitterStreamImpl: Establishing connection.
19/01/15 11:27:59 INFO twitter4j.TwitterStreamImpl: Connection established.
19/01/15 11:27:59 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
19/01/15 11:28:00 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSy
stem = false
19/01/15 11:28:01 INFO hdfs.BucketWriter: Creating hdfs://localhost:8020/user/fl
ume/tweets/FlumeData.1547531880398.tmp
19/01/15 11:28:01 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
19/01/15 11:28:02 INFO twitter.TwitterSource: Processed 100 docs
19/01/15 11:28:06 INFO twitter.TwitterSource: Processed 200 docs
19/01/15 11:28:09 INFO twitter.TwitterSource: Processed 300 docs
19/01/15 11:28:14 INFO twitter.TwitterSource: Processed 400 docs
19/01/15 11:28:17 INFO twitter.TwitterSource: Processed 500 docs
19/01/15 11:28:20 INFO twitter.TwitterSource: Processed 600 docs
```



```
r type: SINK, name: HDFS. sink.event.drain.sucess == 0
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost flume]$ hadoop fs -ls hdfs://localhost:8020/user/flume/tweet
s
19/01/15 11:52:29 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup     681580 2019-01-15 11:28 hdfs://localhost:
8020/user/flume/tweets/FlumeData.1547531880398
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost flume]$ 
```