*Big Data and Hadoop Development*

Session 11: Sqoop Flume

Assignment 3

**A C A D G I L D Page 1**

*Big Data and Hadoop Development*

**Table of Contents**

1. Introduction .......................................................................................................................................... 3

2. Objective ............................................................................................................................................... 3

3. Prerequisites: ........................................................................................................................................ 3

4. Associated Data Files ............................................................................................................................ 3

5. Problem Statement ............................................................................................................................... 3

6. Expected Output ................................................................................................................................... 3

7. Approximate Time to Complete Task ................................................................................................... 3

**A C A D G I L D Page 2**

*Big Data and Hadoop Development*

**1. Introduction**

In this assignment you need to perform the tasks given.

**2. Objective**

This assignment will help you to consolidate the concepts learnt in the session.

**3. Prerequisites**

None

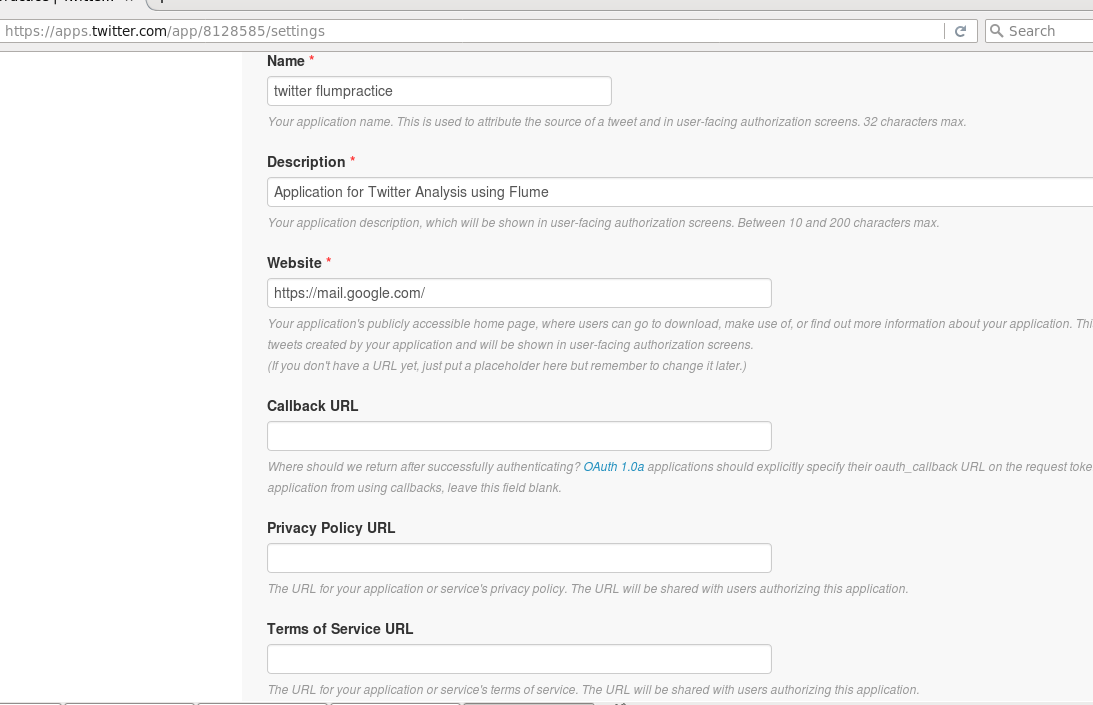
**4. Associated Data Files**

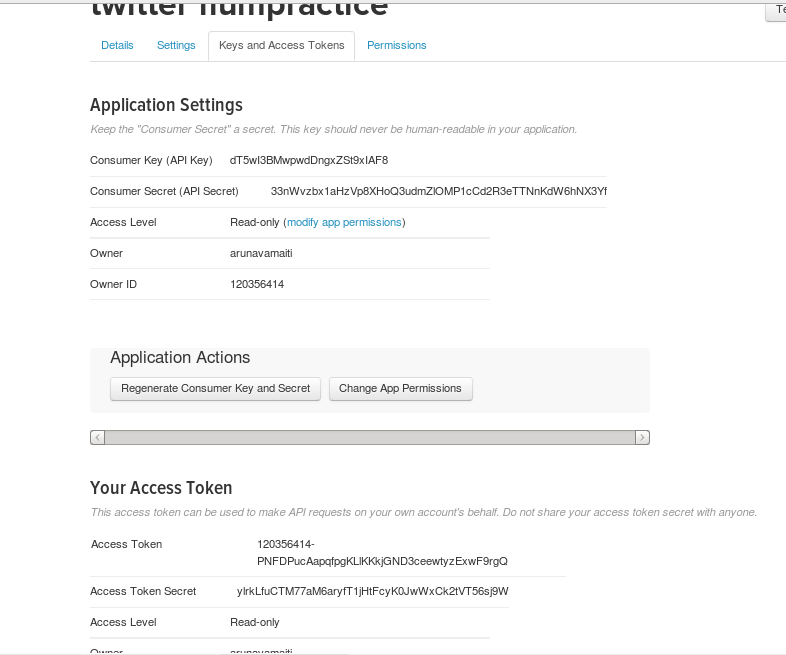
N/A

**5. Problem Statement**

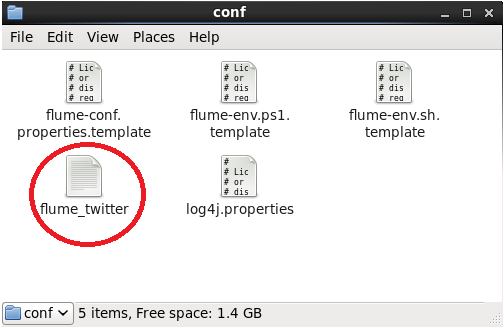
Create a flume agent that streams data from Twitter and stores in the HDFS.

* Login to Twitter Account
* Go to the [**https://apps.twitter.com/app**](https://apps.twitter.com/app)
* and click to the ‘**create new app**’ button at [**https://apps.twitter.com/app**](https://apps.twitter.com/app)
* Create a Twitter application by giving the appropriate details.

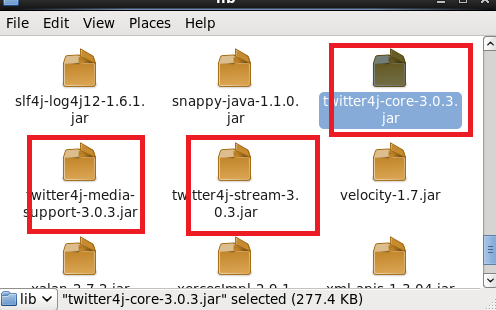




Now go to /**usr/local/flume/conf** (where the flume is installed )



need to make sure below jar are in lib folder



Now we have to write the configuration file for the Twitter Streaming. We use the newly created file for flume\_twitter**.**

* **Need to set Keys**
* **and need to set twitter Keywords**

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=uX0TWqkx0okYEjjqLzxIx6mD6

TwitterAgent.sources.Twitter.consumerSecret=rzHIs3TMJnADbZNvdGU7LQUo0kPxPISq3RGSLfqcBip39X5END

TwitterAgent.sources.Twitter.accessToken=559516596-yDA9xqOljo4CV32wSnqsx2BXh4RBIRKFxZGSZrPC

TwitterAgent.sources.Twitter.accessTokenSecret=zDxePILZitS5tIWBhre0GWqps0FIj9OadX8RZb6w8ZCwz

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/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

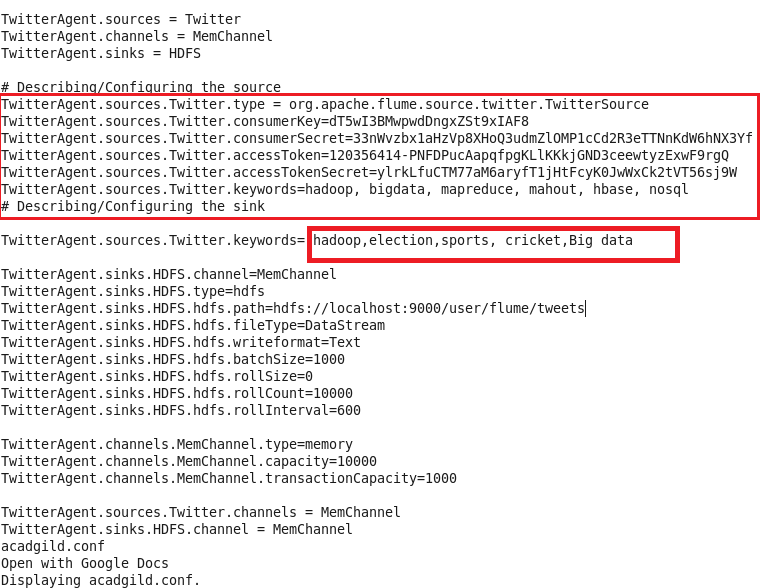
TwitterAgent.channels.MemChannel.type=memory

TwitterAgent.channels.MemChannel.capacity=10000

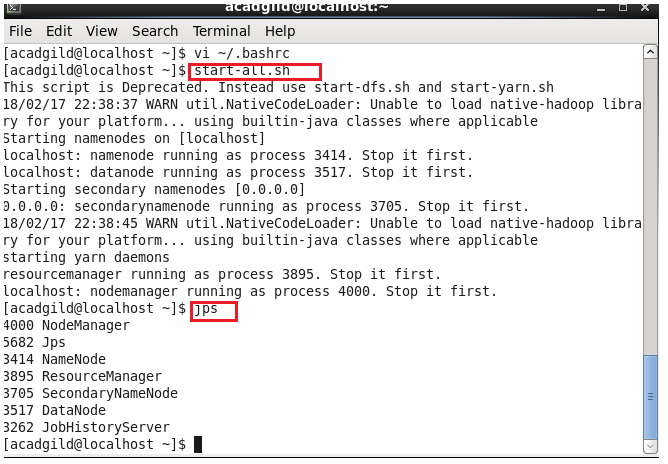
TwitterAgent.channels.MemChannel.transactionCapacity=1000

TwitterAgent.sources.Twitter.channels = MemChannel

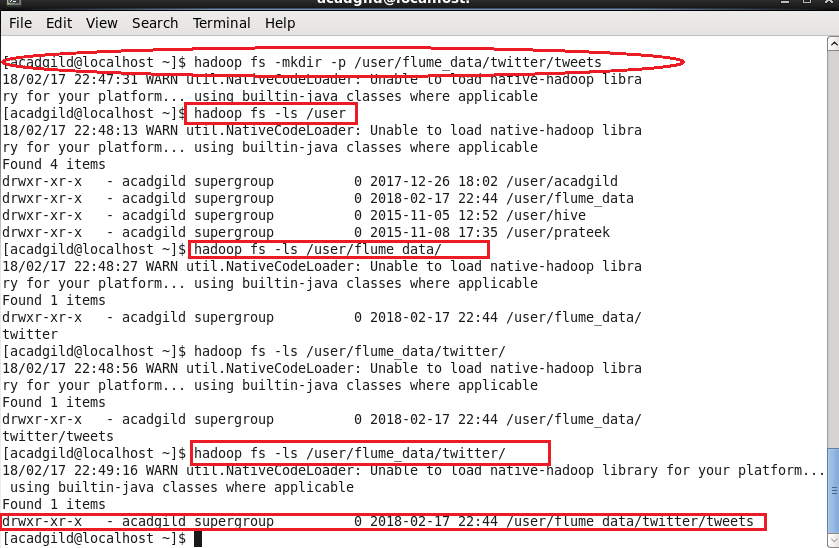
TwitterAgent.sinks.HDFS.channel = MemChannel



* Now we have to execute the flume agent and configuration file
* start all Hadoop daemons (**start-all.sh**).
* Check if they are running (**jps**)



* Create directory inside HDFS path, where the Twitter tweet data should be stored **/user/flume/twitter/tweets/**



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 <location of created/edited conf file>**

**and for stop ctrl+s**



**to verify the data**



output:

