# Session 12:
# Oozie and Flume
# Assignment 1

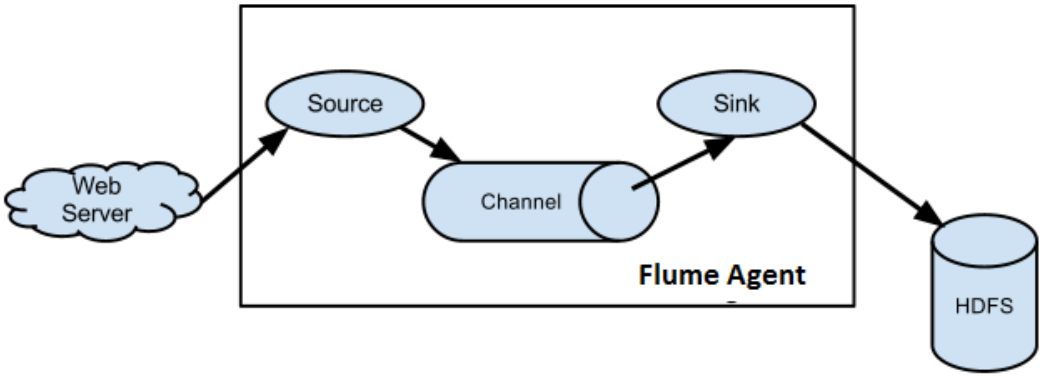

- Prachi Mohite

Apache Flume is a Hadoop ecosystem component used to collect, aggregate and moves a large amount of log data from different sources to a centralized data store. It is an open source component which is designed to locate and store the data in a distributed environment and collects the data as per the specified input key(s).

Flume is composed of the following components. **Flume Event:** It is the main unit of the data that is transported inside the **Flume** (Typically a single log entry). It contains a payload of the byte array that is to be transported from the source path to the destination path which could be accompanied by optional
headers. A Flume
event will be in the following

Header → Payload

**Flume Agent:** Is an independent Java virtual machine daemon process which receives the data (events) from clients and transports to the subsequent destination (sink or agent).




**Source:** Is the component of Flume agent which receives data from the data generators say, twitter, facebook, weblogs from different sites and transfers this data to one or more channels in the form of Flume event. The external source sends data to Flume in a format that is recognized by the target Flume source. Example, an Avro Flume source can be used to receive Avro data from Avro clients or other Flume agents in the flow that send data from an Avro sink, or the Thrift Flume source will receive data from a Thrift sink, or a Flume Thrift RPC client or Thrift Clients are written in any language generated from the Flume thrift protocol.

**Channel:** Once, the Flume source receives an Event, it stores this data into one or more channel and buffers them till they are consumed by sinks. It acts as a bridge between the source and sinks. These channels are implemented to handle any number of sources and sinks.

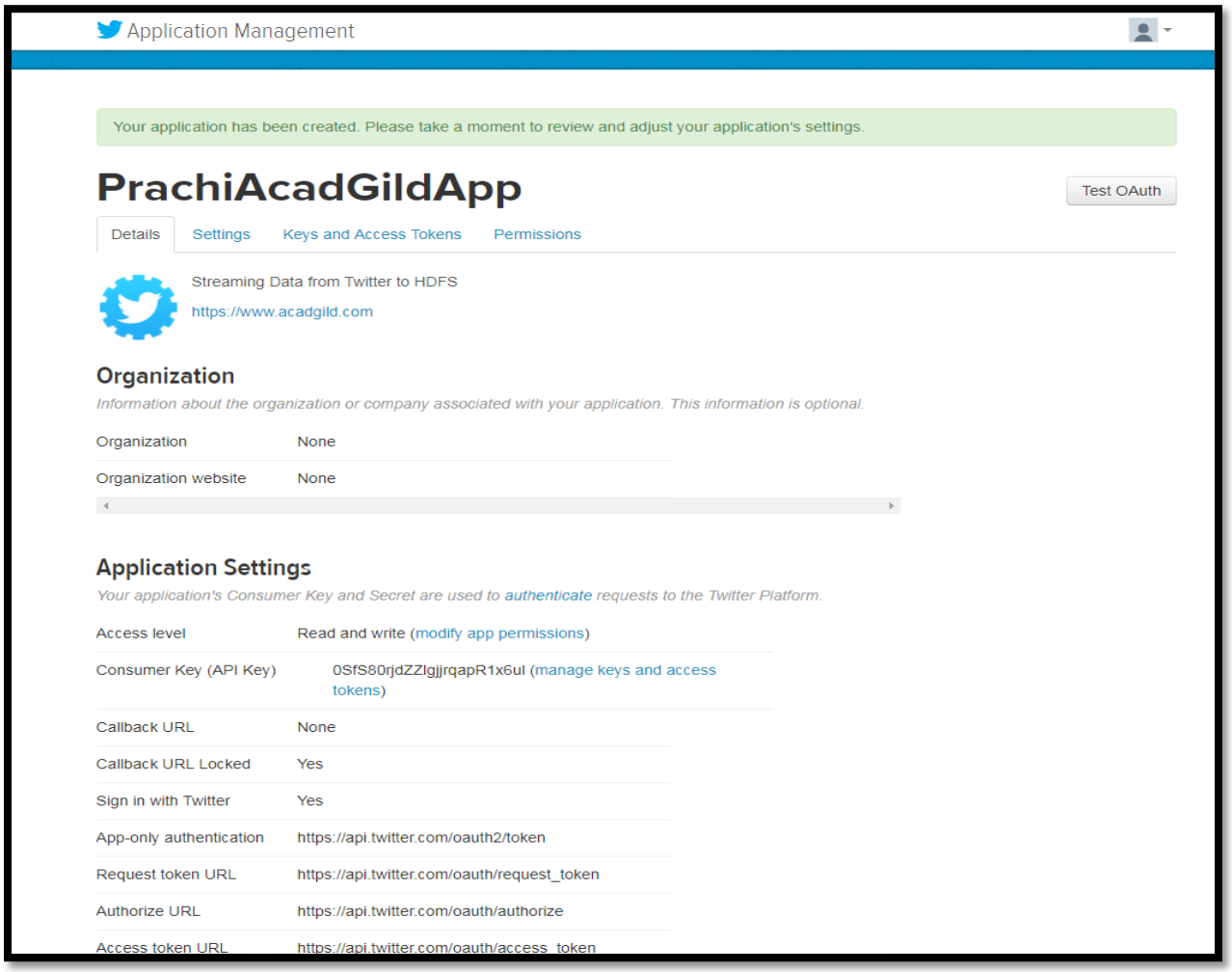
Task :
Create a flume agent that streams data from Twitter and stores in the HDFS.

**Streaming Twitter Data** To stream data to our database from twitter we should have the following pre-requisites.

- Twitter account
- Hadoop cluster

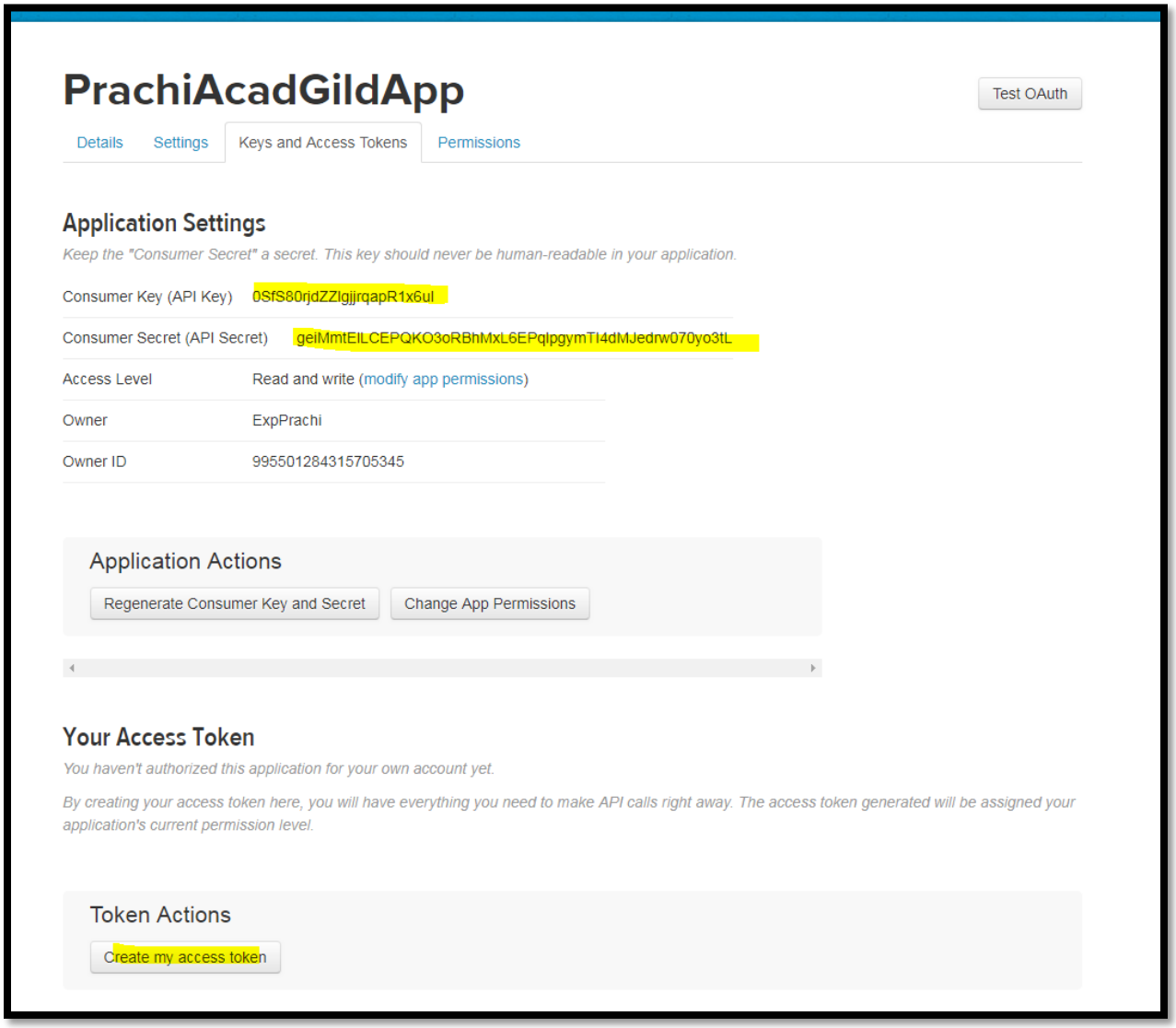
If both prerequisites are available we can move to our further step. **Step 1:** Login to the twitter account

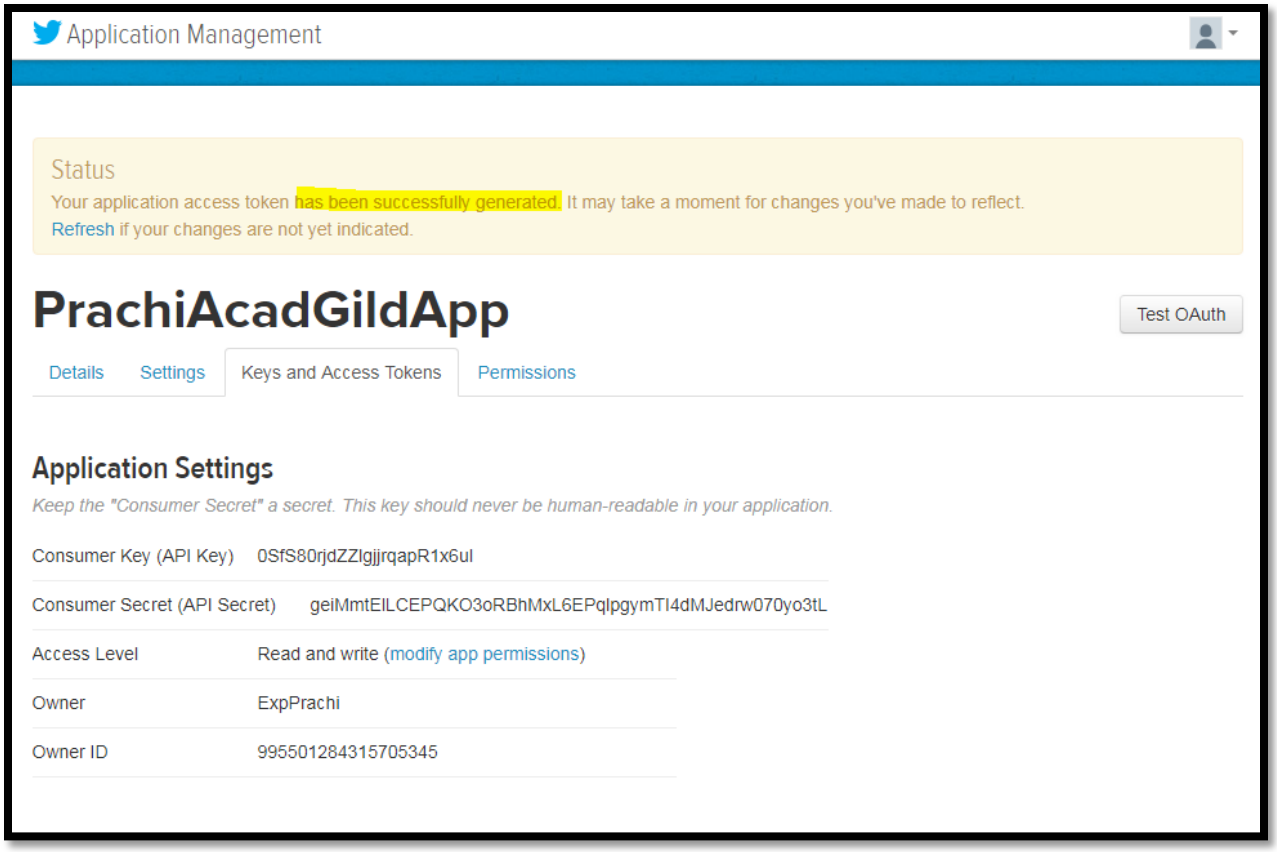
1. Login to Twitter Account
2. Go to the following link and click the 'create new app' button. https://apps.twitter.com/app
3. **Complete all the necessary steps required and 'Create new Twitter Application'**

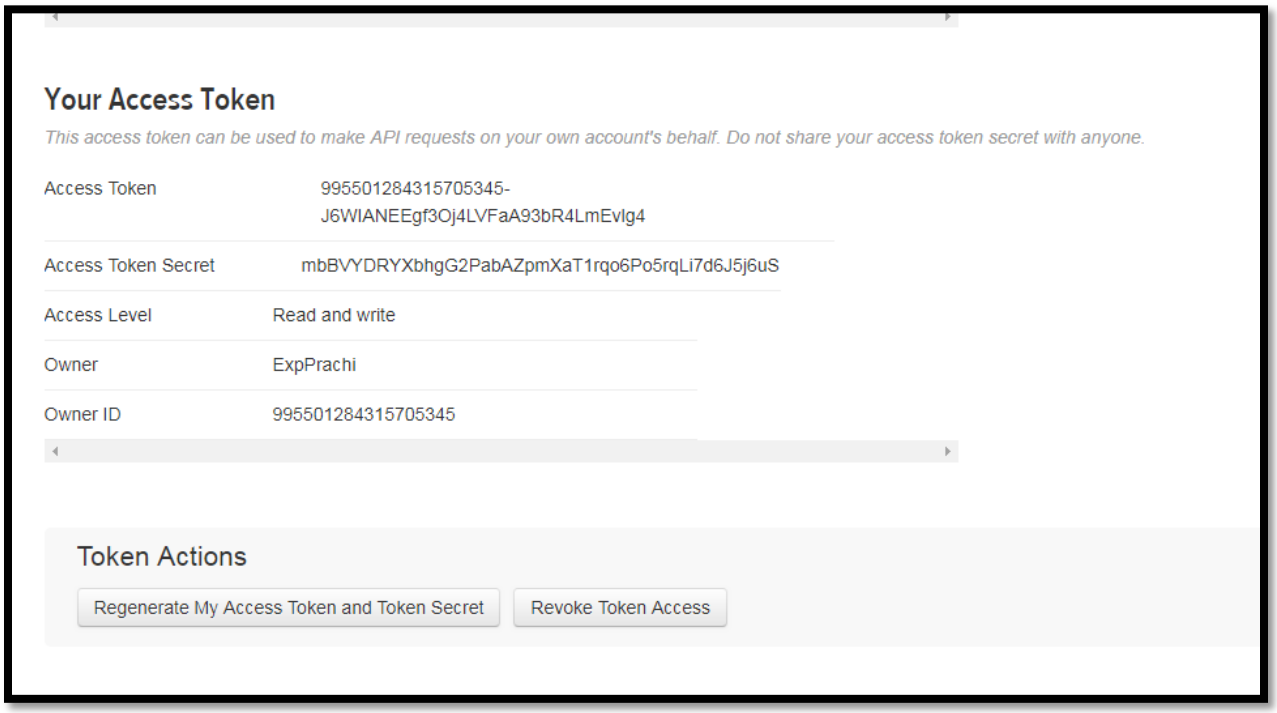


4. Select the ‘Keys and Access Token’ tab.

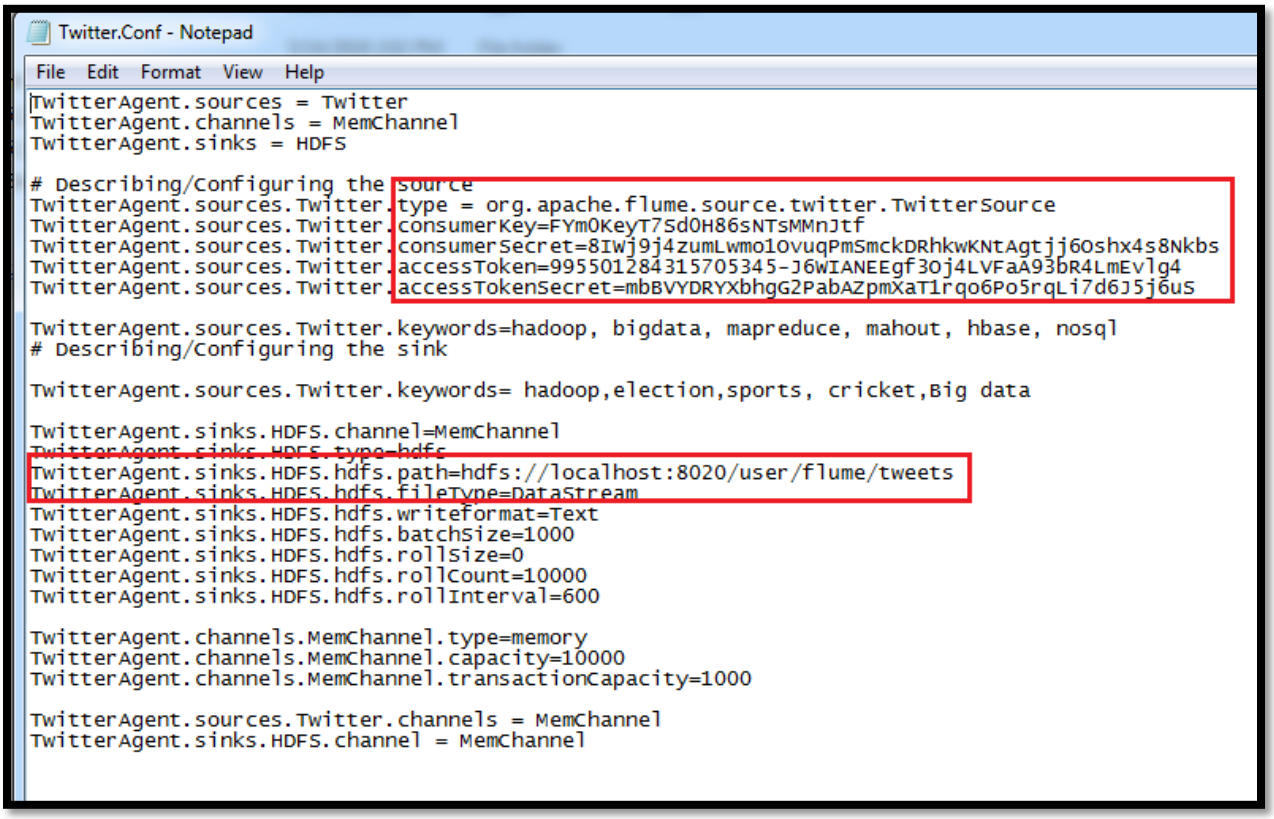
   Copy the consumer key and the consumer secret code.
   Scroll down further and select the ‘create my access token’ button.



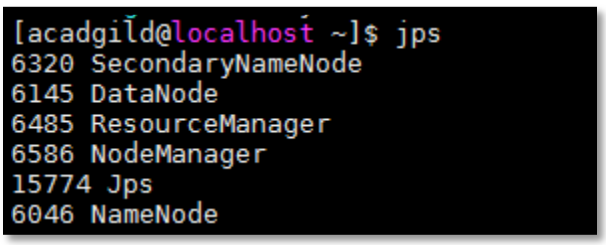

Application Management
Status
Your application access token has been successfully generated. It may take a moment for changes you've made to reflect.
Refresh if your changes are not yet indicated.
PrachiAcadGildApp
Test OAuth
Details
Settings
Keys and Access Tokens
Permissions
Application Settings
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.
Consumer Key (API Key) 0SfS80rjdZZIgjjrqapR1x6ul
Consumer Secret (API Secret) geiMmtEILCEPQKO3oRBhMxL6EPqlpgymTI4dMJedrw070yo3tL
Access Level Read and write (modify app permissions)
Owner ExpPrachi
Owner ID 995501284315705345


Your Access Token
This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.
Access Token 995501284315705345-J6WIANEEgf3Oj4LVFaA93bR4LmEvlg4
Access Token Secret mbBVYDRYXbhgG2PabAZpmXaT1rqo6Po5rqLi7d6J5j6uS
Access Level Read and write
Owner ExpPrachi
Owner ID 995501284315705345
Token Actions
Regenerate My Access Token and Token Secret
Revoke Token Access

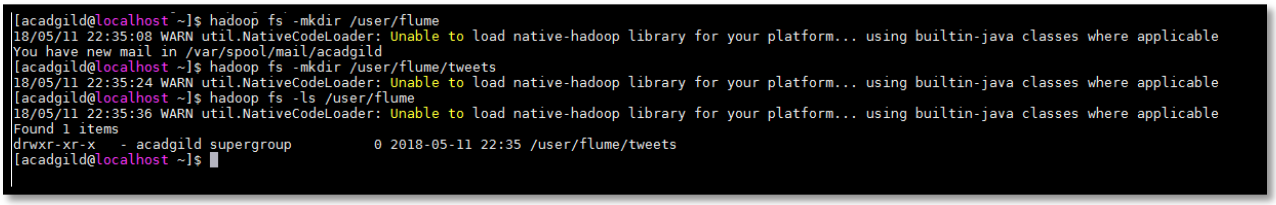
**Edit the above entries in the configuration Files**




**Make sure the HDFS file is working**

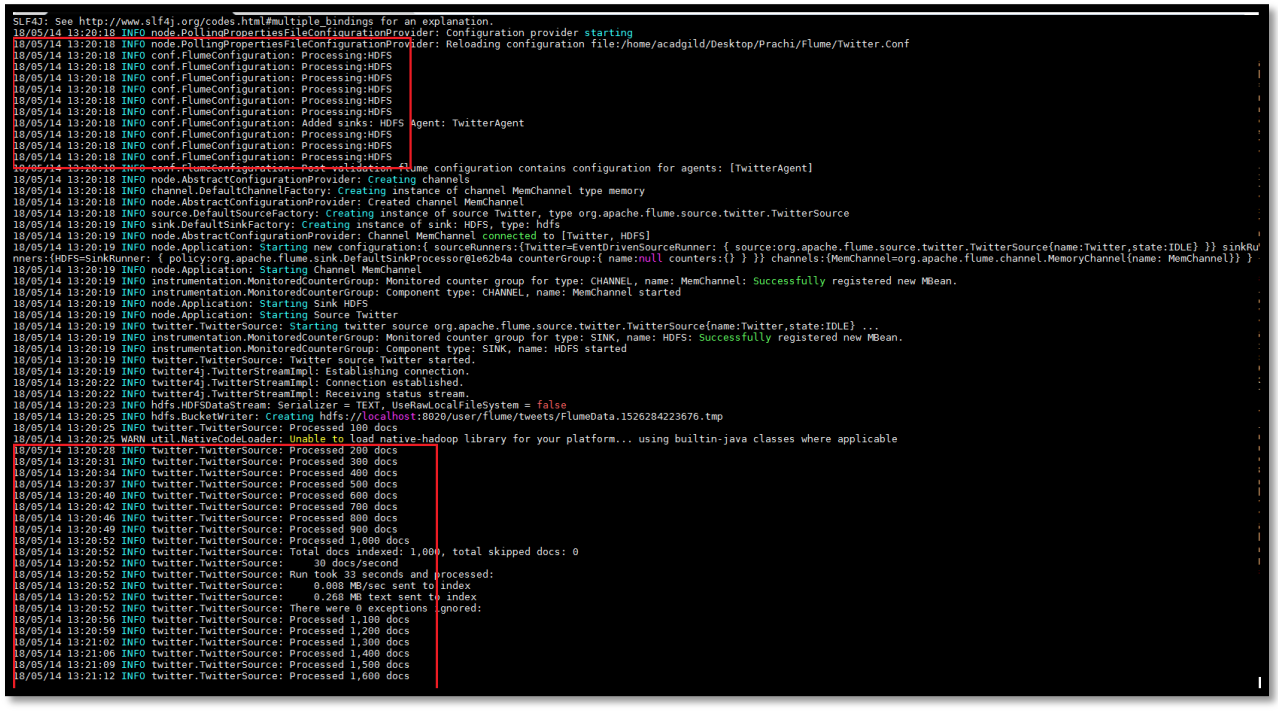



**Make sure the directory mentioned in configuration file is created n HDFS.**



For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path. **flume-ng agent –conf c -n TwitterAgent -f <location of created/edited conf file>**

**flume-ng agent –conf c -n TwitterAgent  -f /home/acadgild/Desktop/Prachi/Flume/ Twitter.Conf**

## The comments from Twitter are loaded into HDFS folder

```
[acadgild@localhost ~]$ hadoop fs  -ls /user/flume/tweets
18/05/14 13:59:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup    8885532 2018-05-14 13:30 /user/flume/tweets/FlumeData.1526284223676
-rw-r--r--   1 acadgild supergroup    3584352 2018-05-14 13:40 /user/flume/tweets/FlumeData.1526284829820
You have new mail in /var/spool/mail/acadgild
```

## Using the cat command to see the output