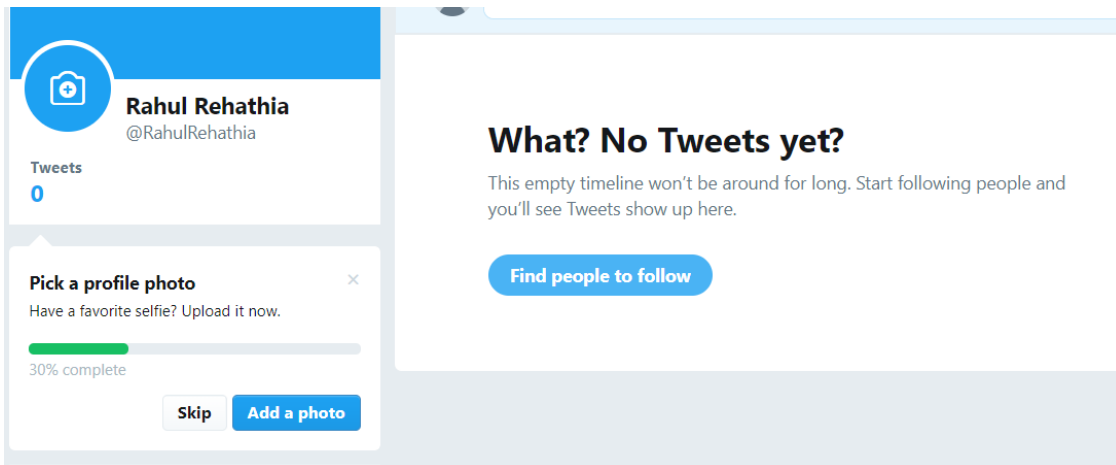


## Assignment 12.1

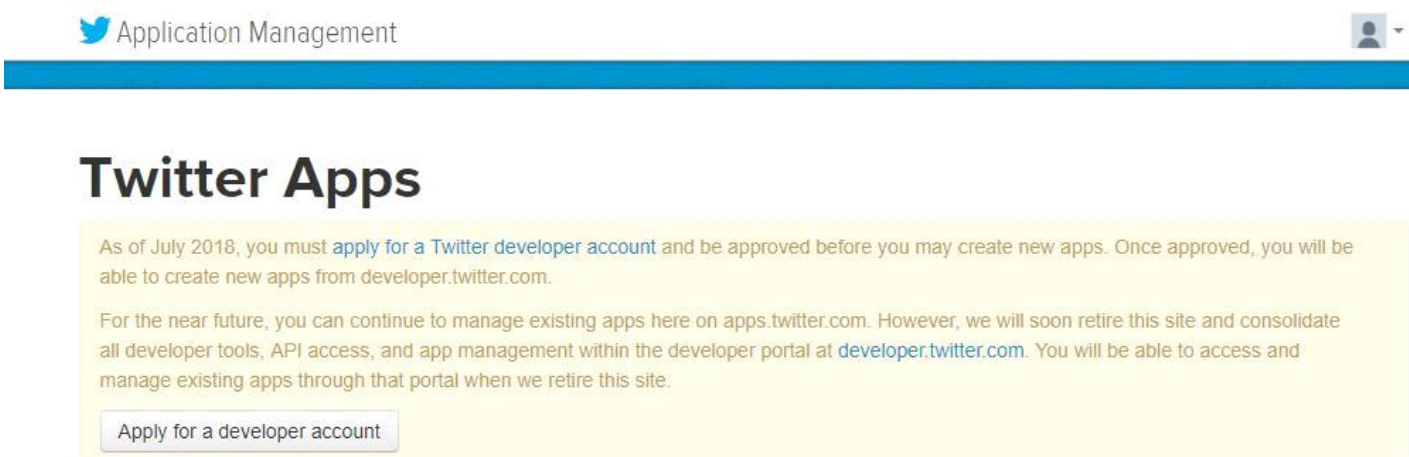
### Oozie and Flume

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

- 1) We have created new account on Twitter.



1. Then we go to the link: <https://apps.twitter.com/app> and click the 'create new app' button.



Then we have applied for Developer Account. Now we are waiting for approval.

As approval is still pending, we could not create a new application with required details like Name, Description, key, etc.

1. We have downloaded flume tar file from link :  
<https://drive.google.com/drive/u/0/folders/0B1QaXx7tpw3SWkMwVFBkc3djNFk> and extracted it.
1. Then we have edited .bashrc file and set the path of flume directory. Then closed the .bashrc file after saving it. And then in the terminal, we have used command '**source .bashrc**' to update the .bashrc file.

```
# Below 2 lines we have to add for kafka Installation

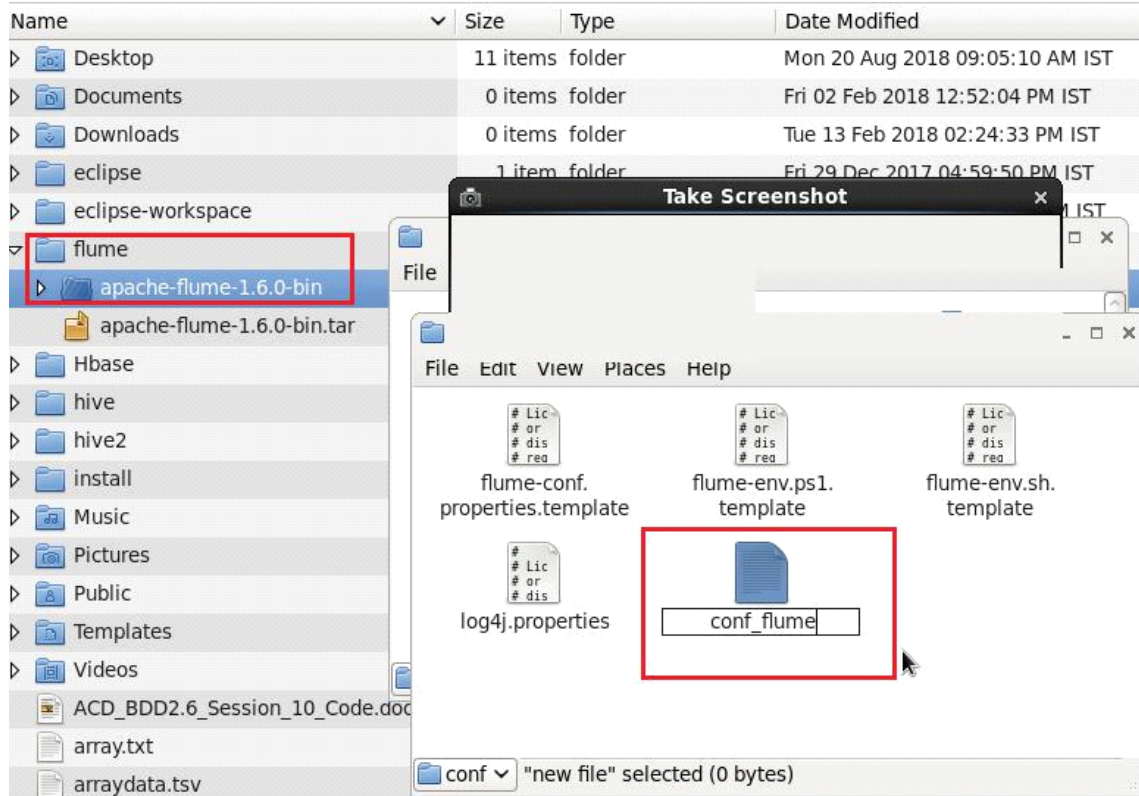
export KAFKA_HOME=/home/acadgild/install/kafka/kafka_2.12-0.10.1.1
export PATH=$PATH:$KAFKA_HOME/bin

# Below 2 lines we have to add for FLUME Installation

export FLUME_HOME=/home/acadgild/flume|
export PATH=$PATH:$FLUME_HOME/bin

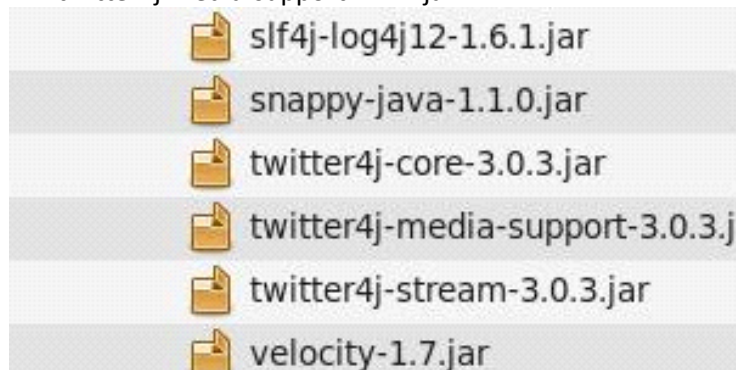
# Below 2 lines we have to add for zookeeper Installation
```

1. We have created a new file **conf\_flume** inside the **conf** directory of **apache-flume-1.6.0-bin** folder.



- We have verified that below jars placed in \$FLUME\_HOME/lib directory i.e. apache-flume-1.6.0-bin/lib folder :

- twitter4j-core-X.XX.jar
- twitter4j-stream-X.X.X.jar
- twitter4j-media-support-X.X.X.jar



1. We have copied the Flume configuration code from the link <https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWINidkk> and pasted it in the newly created file inside the conf directory of **apache-flume-1.6.0-bin** folder. Then we have saved this file as **flume.conf**

## Assignment 12.1

### Oozie and Flume

```
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-yDA9xq0ljo4CV32wSnqxs2BXh4RBIRKfxZG5ZrPC
TwitterAgent.sources.Twitter.accessTokenSecret=zDxePILZitS5tIWBhre0GWqps0FIj90adX8RZb6w8ZCwz
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
```

- 8) We run jps command to verify all hadoop daemons are running fine.

```
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$ jps  
4769 DataNode  
5129 ResourceManager  
6970 Jps  
4972 SecondaryNameNode  
4670 NameNode  
5230 NodeManager  
You have new mail in /var/spool/mail/acadgild
```

1. We have created a new directory inside HDFS path, where the Twitter tweet data should be stored.

**hadoop dfs -mkdir -p /user/flume/tweets**

```
mkdir: Use of this script to execute mkdir command is deprecated.  
[acadgild@localhost ~]$ hadoop dfs -mkdir -p /user/flume/tweets  
DEPRECATED: Use of this script to execute hdfs command is deprecated.  
Instead use the hdfs command for it.  
18/08/22 09:34:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

## Assignment 12.1

### Oozie and Flume

1. To fetch data from Twitter, we are using below command to fetch the twitter tweet data into the HDFS cluster path.

**flume-ng agent -n TwitterAgent -f /home/acadgild/flume/apache-flume-1.6.0-bin/conf/flume.conf**

```
applicator
[acadgild@localhost ~]$ flume-ng agent -n TwitterAgent -f /home/acadgild/flume/apache-flume-1.6.0-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-1.6.0-bin/lib/*:/home/acadgild/install/hadoop/hadoop-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/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/*:/home/acadgild/install/hadoop/hadoop-2.6.5/contrib/capacity-scheduler/*:/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
```

1. To check the contents of the tweet data we can use the following command:  
**hadoop dfs -ls /user/flume/tweets**

12) Then to display the tweet data inside the /user/flume/tweets folder we are using below command :

**hadoop dfs -cat /user/flume/tweets/<flumeData file name>**