ASSIGNMENT 11.3

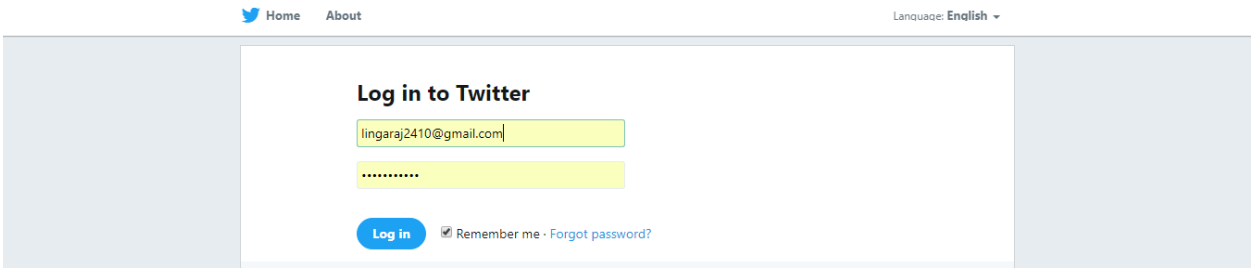
Problem Statement:

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

Solution:

Twitter data streaming into HDFS using Flume:

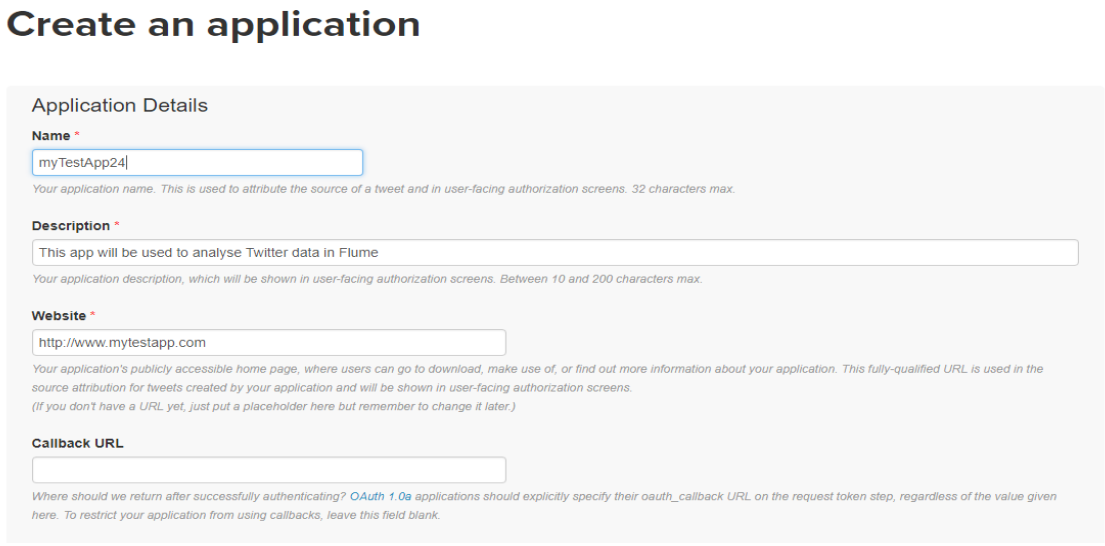
Step 1: Login to the twitter account



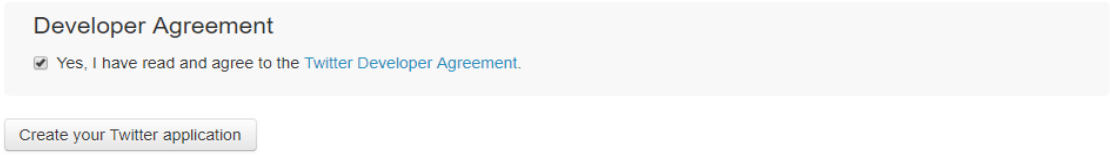
Step 2: Go to the following link and click the 'create new app' button.

https://apps.twitter.com/app

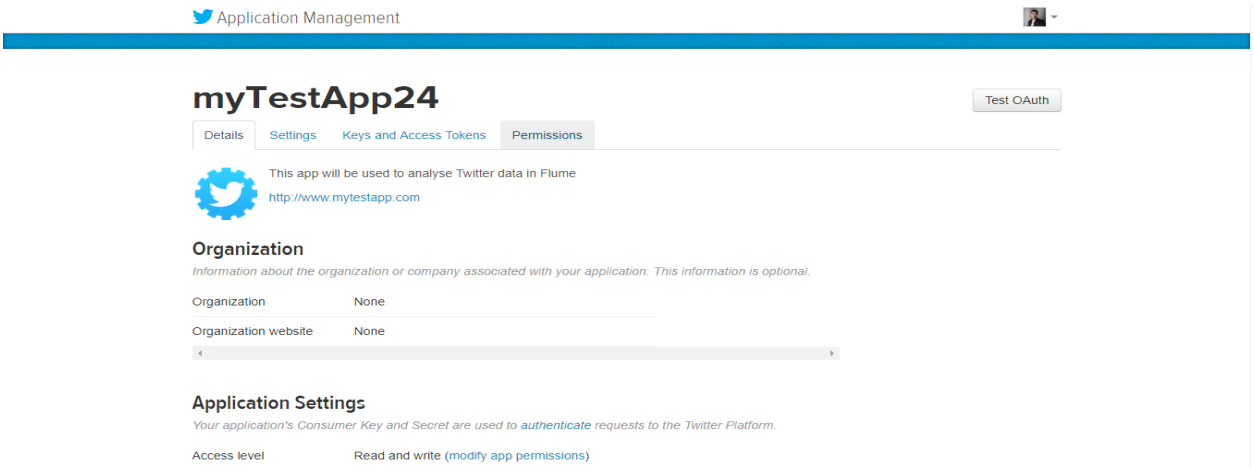
Step 3: Enter the necessary details.



Step 4: Accept the developer agreement and select the ‘create your Twitter application’ button.

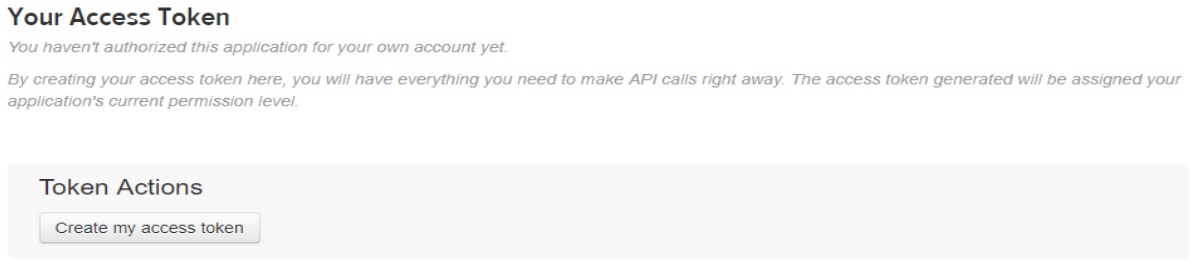


Step 5: Select the ‘Keys and Access Token’ tab.

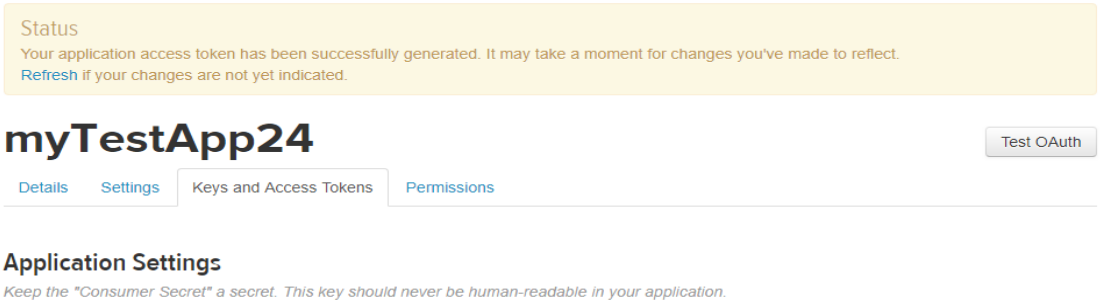


Step 6: Copy the consumer key and the consumer secret code.

Step 7: Scroll down further and select the ‘create my access token’ button.



Now, you will receive a message stating “that you have successfully generated your application access token”.



Step 8: Copy the Access Token and Access token Secret code.

Step 9: Create a configuration file that is used to specify Flume properties in compliance with Twitter streaming. Save this file in the ‘conf’ deirectory of Flume home path.

Here is the properties file I have created that includes Twitter app credentials as well as the HDFS path into which Tweets will 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=bTeT3P3RhV3XfyZ2XKKk49dac
TwitterAgent.sources.Twitter.consumerSecret=L2aNWtNxxeKUGHME71RDIPg6UjCB44il0fGFmmQqHV68my7KFd
TwitterAgent.sources.Twitter.accessToken=2221479913-T6KiphqccmfYjzIL6k6AZDG0RS6R8iINMD6SDW4
TwitterAgent.sources.Twitter.accessTokenSecret=IW6Rm3eDnaKeu4leiJT4tzmr7bxOheJmJAvpwIQsNQHQL
TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, spark, 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
```

We can mention which keywords tweet data to be collected from the Twitter application. We can change the keywords in the TwitterAgent.sources.Twitter.keywords property. In our example, we are fetching tweet data related to Hadoop, election, sports, cricket and Big data.

Step 10: Open a new terminal and start all the Hadoop daemons, before running the flume command to fetch the twitter data. Use the ‘jps’ command to see the running Hadoop daemons.

```
[acadgild@localhost sbin]$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/01/23 16:48:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: datanode running as process 2810. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-secondarynamenode-localhost.loca
ldomain.out
18/01/23 16:48:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
starting yarn daemons
resourcemanager running as process 3101. Stop it first.
localhost: nodemanager running as process 3204. Stop it first.
[acadgild@localhost sbin]$ jps
4082 Jps
3204 NodeManager
3881 SecondaryNameNode
2810 DataNode
3101 ResourceManager
3663 NameNode
[acadgild@localhost sbin]$
```

Step 11: Create a new directory inside HDFS path, where the Twitter tweet data should be stored.

$ hadoop fs –mkdir –p /user/acadgild/twitter_data

```
[acadgild@localhost sbin]$ hadoop fs -mkdir /user/acadgild/twitter_data
18/01/23 16:53:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
[acadgild@localhost sbin]$ hadoop fs -ls /user/acadgild/
18/01/23 16:53:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 6 items
drwxr-xr-x   - acadgild supergroup          0 2015-11-20 11:46 /user/acadgild/Pictures
drwxr-xr-x   - acadgild supergroup          0 2018-01-23 12:32 /user/acadgild/_sqoop
drwxr-xr-x   - acadgild supergroup          0 2018-01-23 10:13 /user/acadgild/emp_info
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 02:03 /user/acadgild/oozie-acad
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 02:00 /user/acadgild/share
drwxr-xr-x   - acadgild supergroup          0 2018-01-23 16:53 /user/acadgild/twitter_data
[acadgild@localhost sbin]$
```

Step 12: 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 /usr/local/flume/conf/acadgild.conf

```
[acadgild@localhost sbin]$ flume-ng agent -n TwitterAgent -f /usr/local/flume/conf/acadgild.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop) for HDFS access
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including HBASE libraries found via (/usr/local/hbase/bin/hbase) for HBASE access
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
+ exec /usr/local/java/bin/java -Xmx20m -cp '/usr/local/flume/lib/*:/usr/local/hadoop-2.6.0/contrib/capacity-scheduler/*.jar:
/usr/local/hadoop-2.6.0/etc/hadoop:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/activation-1.1.jar:/usr/local/hadoop-2.6.0
/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/apacheds-kerberos-codec-
2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop
/common/lib/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/asm-3.2.jar:/usr/local/hadoop-2.6.0/share/
hadoop/common/lib/avro-1.7.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoo
p-2.6.0/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-cli-
1.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-codec-1.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib
/commons-collections-3.2.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-compress-1.4.1.jar:/usr/local/hadoop-2
.6.0/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-digester-1

18/01/23 17:38:02 INFO node.AbstractConfigurationProvider: Creating channels
18/01/23 17:38:02 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
18/01/23 17:38:02 INFO node.AbstractConfigurationProvider: Created channel MemChannel
18/01/23 17:38:02 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter
.TwitterSource
18/01/23 17:38:02 INFO twitter.TwitterSource: Consumer Key:        'EZcZZXOhCY35QZHCL78LCvKBm'
18/01/23 17:38:02 INFO twitter.TwitterSource: Consumer Secret:     'et9R8KBOtawc9qvD5maT1SpfCkAzLecGQxTWsy5sOExoIGtMxm'
18/01/23 17:38:02 INFO twitter.TwitterSource: Access Token:        '2221479913-sLoCSKtd5u7LfuwWTlXz8eadv91fWT6qYidRNlz'
18/01/23 17:38:02 INFO twitter.TwitterSource: Access Token Secret: 'rPsEerMTxsfE1fEvy43nhVVn33Jff1Gzb8DbLo6YDO06a'
18/01/23 17:38:02 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
18/01/23 17:38:03 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
18/01/23 17:38:03 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { sourc
e:org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache
.flume.sink.DefaultSinkProcessor@330e14bc counterGroup:{ name:null counters:{} } }} channels:{MemChannel=org.apache.flume.cha
nnel.MemoryChannel{name: MemChannel}} }
18/01/23 17:38:03 INFO node.Application: Starting Channel MemChannel
18/01/23 17:38:03 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Su
ccessfully registered new MBean.
18/01/23 17:38:03 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
18/01/23 17:38:03 INFO node.Application: Starting Sink HDFS
18/01/23 17:38:03 INFO node.Application: Starting Source Twitter
18/01/23 17:38:03 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twit
ter,state:IDLE} ...
18/01/23 17:38:03 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfull
y registered new MBean.
18/01/23 17:38:03 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/01/23 17:38:03 INFO twitter.TwitterSource: Twitter source Twitter started.
18/01/23 17:38:03 INFO twitter4j.TwitterStreamImpl: Establishing connection.
```

```
18/01/25 18:16:40 INFO twitter.TwitterSource: Processed 400 docs
18/01/25 18:16:42 INFO twitter.TwitterSource: Processed 500 docs
18/01/25 18:16:44 INFO twitter.TwitterSource: Processed 600 docs
18/01/25 18:16:47 INFO twitter.TwitterSource: Processed 700 docs
18/01/25 18:16:50 INFO twitter.TwitterSource: Processed 800 docs
18/01/25 18:16:52 INFO twitter.TwitterSource: Processed 900 docs
18/01/25 18:16:55 INFO twitter.TwitterSource: Processed 1,000 docs
18/01/25 18:16:55 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
18/01/25 18:16:55 INFO twitter.TwitterSource:     38 docs/second
18/01/25 18:16:55 INFO twitter.TwitterSource: Run took 26 seconds and processed:
18/01/25 18:16:55 INFO twitter.TwitterSource:     0.01 MB/sec sent to index
18/01/25 18:16:55 INFO twitter.TwitterSource:     0.271 MB text sent to index
18/01/25 18:16:55 INFO twitter.TwitterSource: There were 0 exceptions ignored:
18/01/25 18:16:58 INFO twitter.TwitterSource: Processed 1,100 docs
18/01/25 18:17:00 INFO twitter.TwitterSource: Processed 1,200 docs
18/01/25 18:17:02 INFO twitter.TwitterSource: Processed 1,300 docs
18/01/25 18:17:05 INFO twitter.TwitterSource: Processed 1,400 docs
18/01/25 18:17:08 INFO twitter.TwitterSource: Processed 1,500 docs
18/01/25 18:17:12 INFO twitter.TwitterSource: Processed 1,600 docs
18/01/25 18:17:16 INFO twitter.TwitterSource: Processed 1,700 docs
18/01/25 18:17:18 INFO twitter.TwitterSource: Processed 1,800 docs
18/01/25 18:17:20 INFO twitter.TwitterSource: Processed 1,900 docs
18/01/25 18:17:22 INFO twitter.TwitterSource: Processed 2,000 docs
```

We can derive from the logs above that the Flume agent is able to fetch Tweets from Twitter and storing the related data into HDFS directory which we have mentioned in the configuration file.

Let’s stop this process and check the output in output directory.

```
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/twitter_data
18/01/25 18:28:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r--   1 acadgild supergroup    3917637 2018-01-25 18:19 /user/acadgild/twitter_data/FlumeData.1516884393277
[acadgild@localhost ~]$
```

$ hadoop fs –cat /user/acadgild/twitter_data/FlumeData.1516884393277

```
른 | 판원 넬원 옹원 짹원 Wanna One || ♡방탄소년단 || 엑소♡ || 세븐틴♡ (이종석•박보검) (คชา,ริท) ❤#chanbaek #kookmin #minso
(핑크보이okik_67(2018-01-25T18:19:12Z RT @BESTFRIEND__92: 180125 BAEKHYIN IG LIVE
แพคฮยอน : ส่วนสูงของผมหรอครับ? กี่ปีผ่านไปก็เหมือนเดิม 'ㅅ' a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone</a>956509257209229319 국♡ 국:반탄소년단:02:ｻﾝｻﾁｼﾞﾃﾞｽ:@BTS_twt
                                                                                                                  코코로:1tosWyEPoHePB4(2018-01-25T1
8:19:12Z jiminnnn_1013 めちゃくそ美人よね。がんばってみるる。 a href="http://twitter.com/download/iphone" rel="no
follow">Twitter for iPhone</a>956509257226072064 ービット紳士です。フォローRT等お気軽に。
2017年までゲーム業界人でしたがマジで楽がしたいので脱社畜が目標。
豆腐メンタル/Destiny大好き/代表作(雑用)ネットハイP/閃乱カグラほか高木P作品/幕末Rock
/ブログhttp://bit.ly/Gentleblog
/動画http://bit.ly/GentleYT L 今世紀最高の紳士ッピー1AmNo1Gentleman(2018-01-25T18:19:12Zz@DirtyMeganeMAN 配信しながらや
っていいですか！ a href="http://twitter.com" rel="nofollow">Twitter Web Client</a> 956509257226010625
응원하다☆정용화☆CNBLUE☆ 한사랑 타소리+드럼소리+피아노소리 WAKEUP Realize
❤언제나화이팅 그대 날래나래☆ aRaeJYH(2018-01-25T18:19:12Z RT @CNlove_binbin: #경희대_진실을_밝혀
We will stay young forever a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a> X1 R
      T
          /
{"type":"record","name":"Doc","doc":"adoc","fields":[{"name":"id","type":"string"},{"name":"user_friends_count","type":["int"
,"null"]},{"name":"user_location","type":["string","null"]},{"name":"user_description","type":["string","null"]},{"name":"use
r_statuses_count","type":["int","null"]},{"name":"user_followers_count","type":["int","null"]},{"name":"user_name","type":["s
tring","null"]},{"name":"user_screen_name","type":["string","null"]},{"name":"created_at","type":["string","null"]},{"name":"
text","type":["string","null"]},{"name":"retweet_count","type":["long","null"]},{"name":"retweeted","type":["boolean","null"]
},{"name":"in_reply_to_user_id","type":["long","null"]},{"name":"source","type":["string","null"]},{"name":"in_reply_to_statu
s_id","type":["long","null"]},{"name":"media_url_https","type":["string","null"]},{"name":"expanded_url","type":["string","nu
ll"]}]}'4V-g9D LD 956509257226055681 みあか$全裸で着衣泳
nwdxf(2018-01-25T18:19:12Z RT @yuzuki lestokyo: 歌ってるゆづきさん
```

Since we haven’t mentioned any specified language for Tweets, we are seeing tweets from various languages. This is how we can fetch Twitter data in the real time with the use of **Apache Flume**.