

ASSIGNMENT FLUME

TASK: STREAMING TWITTER DATA INTO HDFS USING FLUME

STEP 1: LOGIN INTO TWITTER ACCOUNT

STEP 2: Go to the following link and click the create new app button.
<https://apps.twitter.com/app>

STEP 3: CREATE APPLICATION BY FILLING THE DETAILS AS FOLLOWS AND SCROLL DOWN
N CLICK CREATE APPLICATION

Create an application

Application Details

Name *

acadgildApp

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

This app will help me do analysis in flume

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

<http://www.yahoo.com>

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL, yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successful authentication? [Learn more](#) Applications should explicitly specify their callback URL on the request token screen.

STEP 4: AFTER THE SUCCESSFUL CREATION OF APP , CLICK ON KEYS AND ACCESS TOKENS
OPTION ON TOP.

STEP 5: COPY THE CONSUMER ACCESS KEY AND CONSUMER ACCESS SECRET KEY

STEP 6: THEN SCROLL DOWN AND CLICK ON GENERATE ACCESS TOKEN

STEP 7: COPY THE ACCESS TOKEN AND ACCESS SECRET TOKEN

STEP 8: Download flume tar file from below link and extract it.
<https://drive.google.com/drive/u/0/folders/0B1QaXx7tpw3SWkMwVFBkc3djenFk>

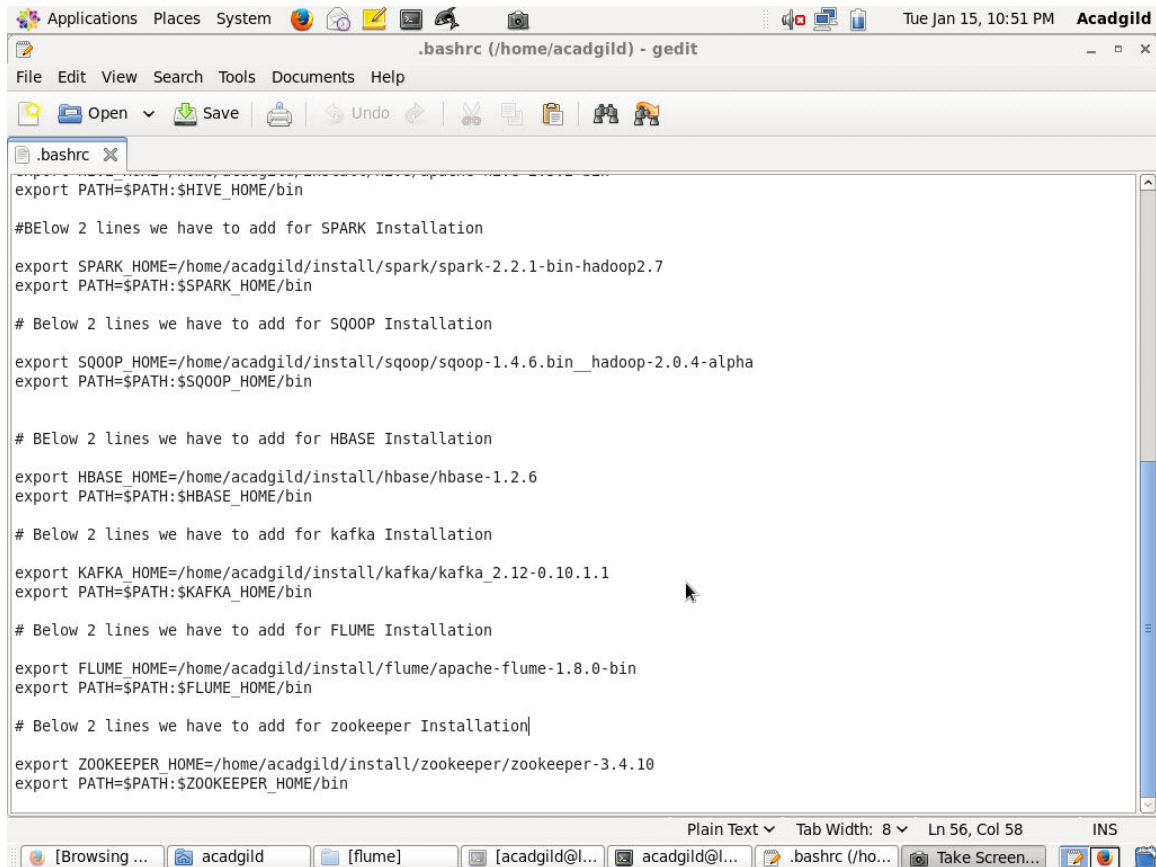
STEP 9: NOW GO TO HDFS SHELL AND TYPE BELOW COMMAND

COMMAND: `sudo gedit .bashrc`

EXPLANATION: A FILE WILL OPEN IN WHICH WE HAVE TO GIVE THE PATH ADDRESS SAME
AS WHERE OUR FLUME EXTRACTED FILE EXISTS (THE SAME AS BELOW)

`export FLUME_HOME=/home/acadgild/install/flume/apache-flume-1.8.0-bin`

```
export PATH=$PATH:$FLUME_HOME/bin
```



```
.bashrc (/home/acadgild) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
.bashrc
export PATH=$PATH:$HIVE_HOME/bin

#Below 2 lines we have to add for SPARK Installation
export SPARK_HOME=/home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin

# Below 2 lines we have to add for SQOOP Installation
export SQOOP_HOME=/home/acadgild/install/sqoop/sqoop-1.4.6.bin_hadoop-2.0.4-alpha
export PATH=$PATH:$SQOOP_HOME/bin

# Below 2 lines we have to add for HBASE Installation
export HBASE_HOME=/home/acadgild/install/hbase/hbase-1.2.6
export PATH=$PATH:$HBASE_HOME/bin

# 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/install/flume/apache-flume-1.8.0-bin
export PATH=$PATH:$FLUME_HOME/bin

# Below 2 lines we have to add for zookeeper Installation
export ZOOKEEPER_HOME=/home/acadgild/install/zookeeper/zookeeper-3.4.10
export PATH=$PATH:$ZOOKEEPER_HOME/bin
```

NOW AFTER ENTERING THE CORRECT PATH AND EXPORT ADDRESS SAVE IT AND CLOSE IT.

STEP 10: NOW TYPE THE BELOW COMMAND

COMMAND: `source .bashrc`

```
[acadgild@localhost ~]$ source .bashrc
[acadgild@localhost ~]$
```

STEP 11: GO TO FLUME EXTRACTED FOLDER AND THEN OPEN THE "conf" FILE. THEN CREATE THE A NEW FILE "flume.conf". NOW DOWNLOAD THE FLUME CONFIGURATION CODE FROM BELOW LINK AND THEN PASTE THE SAME IN THE NEW CREATED FILE.

<https://drive.google.com/open?id=0B1QaXx7tpw3Sb3U4LW9SWlNidkk>

```
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= JBHLOPoQCx9BC0S4dX4KR90v0
TwitterAgent.sources.Twitter.consumerSecret=dljRzS8ujMyZ00vNWTy85eDovk0F0CNC6BzK9yGLaJdkG1z0MW
TwitterAgent.sources.Twitter.accessToken=1085075367189196800-q0gSnMPpCsDEexDgP0x5ydy2GICaCh
TwitterAgent.sources.Twitter.accessTokenSecret=XIbG5yYeXZBd14W3cMdBM6BiRZdZkbPuYGs1TmzCAnB6k |
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
```

STEP 12: NOW OPEN THE FILE AND THEN COPY THE CONSUMER KEYS AND ACCESS TOKENS IN REQUIRED PLACE. REFER SCREENSHOT.

```
TwitterAgent.sources.Twitter.consumerKey=JBHLOPoQCx9BC0S4dX4KR90v0
TwitterAgent.sources.Twitter.consumerSecret=dljRzS8ujMyZ00vNWTy85eDovk0F0CNC6BzK9yGLaJdkG1z0MW
TwitterAgent.sources.Twitter.accessToken=1085075367189196800-q0gSnMPpCsDEexDgP0x5ydy2GICaCh
TwitterAgent.sources.Twitter.accessTokenSecret=XIbG5yYeXZBd14W3cMdBM6BiRZdZkbPuYGs1TmzCAnB6k
```

STEP 13: NOW GO TO HDFS SHELL AND START THE HADOOP DEMONS USING BELOW COMMANDS.

COMMAND: start-all.sh
COMMAND:jps

SOLUTION REPORT:

```
[acadgild@localhost ~]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
19/01/15 22:59:58 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 3058. Stop it first.
localhost: datanode running as process 3155. Stop it first.
Starting secondary namenodes [0.0.0.0]
```

```
0.0.0.0: secondarynamenode running as process 3408. Stop it first.
19/01/15 23:00:10 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
starting yarn daemons
resourcemanager running as process 3612. Stop it first.
localhost: nodemanager running as process 3712. Stop it first.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ jps
3408 SecondaryNameNode
3712 NodeManager
3058 NameNode
3155 DataNode
5716 Application
3612 ResourceManager
7694 Jps
```

EXPLANATION : NOW CREATE A FILE IN WHICH THE TWEET DATA SHOULD BE STREAMING IN. "NOTE: THE PATH SHOULD BE SAME AS THE SINK PATH GIVEN IN THE CONFIGURATION FILE"

COMMAND: `hadoop fs -mkdir -p /user/flume/tweets/`

SOLUTION REPORT:

```
[acadgild@localhost ~]$ hadoop fs -mkdir /user/flume/tweets
19/01/15 23:03:16 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
You have new mail in /var/spool/mail/acadgild
```

STEP 14: FOR FETCHING DATA USE BELOW COMMAND

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

EXPLANATION: THE FLUME AGENT WILL COMPILE THE CONFIGURATION FILE WHICH IS GIVEN IN THE PATH ADDRESS AND FETCH THE TWEET DATA AND STORE IT IN HDFS.

SOLUTION REPORT:

```
at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingCon
structorAccessorImpl.java:45)
  at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
  at
org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:791)
  at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:731)
  at org.apache.hadoop.ipc.Client.call(Client.java:1474)
  at org.apache.hadoop.ipc.Client.call(Client.java:1401)
  at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngin
e.java:232)
  at com.sun.proxy.$Proxy13.create(Unknown Source)
  at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.cr
eate(ClientNamenodeProtocolTranslatorPB.java:295)
```

```
        at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
        at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccesso
rImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at
org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInv
ocationHandler.java:187)
        at
org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocatio
nHandler.java:102)
        at com.sun.proxy.$Proxy14.create(Unknown Source)
        at
org.apache.hadoop.hdfs.DFSOutputStream.newStreamForCreate(DFSOutputStre
am.java:1721)
        at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1657)
        at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1582)
        at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:397)
        at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:393)
        at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResol
ver.java:81)
        at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:393)
        at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:337)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:908)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:889)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:786)
        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:775)
        at
org.apache.flume.sink.hdfs.HDFSDataStream.doOpen(HDFSDataStream.java:81
)
        at
org.apache.flume.sink.hdfs.HDFSDataStream.open(HDFSDataStream.java:108)
        at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:262)
        at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:252)
        at
org.apache.flume.sink.hdfs.BucketWriter$9$1.run(BucketWriter.java:701)
        at
org.apache.flume.auth.SimpleAuthenticator.execute(SimpleAuthenticator.j
ava:50)
        at
org.apache.flume.sink.hdfs.BucketWriter$9.call(BucketWriter.java:698)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
```

```

        at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
Caused by: java.net.ConnectException: Connection refused
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
        at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
        at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
        at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
        at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:609)
    )
        at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:707)
        at
org.apache.hadoop.ipc.Client$Connection.access$2800(Client.java:370)
        at org.apache.hadoop.ipc.Client.getConnection(Client.java:1523)
        at org.apache.hadoop.ipc.Client.call(Client.java:1440)
        ... 33 more
19/01/15 22:29:45 INFO twitter.TwitterSource: Processed 25,000 docs
19/01/15 22:29:45 INFO twitter.TwitterSource: Total docs indexed: 25,000,
total skipped docs: 0
19/01/15 22:29:45 INFO twitter.TwitterSource:      45 docs/second
19/01/15 22:29:45 INFO twitter.TwitterSource: Run took 551 seconds and
processed:
19/01/15 22:29:45 INFO twitter.TwitterSource:      0.012 MB/sec sent to
index
19/01/15 22:29:45 INFO twitter.TwitterSource:      6.806 MB text sent to
index
19/01/15 22:29:45 INFO twitter.TwitterSource: There were 0 exceptions
ignored:
19/01/15 22:29:47 INFO twitter.TwitterSource: Processed 25,100 docs
19/01/15 22:29:49 INFO hdfs.BucketWriter: Creating
hdfs://localhost:9000/user/flume/tweets/FlumeData.1547571037727.tmp
19/01/15 22:29:50 INFO twitter.TwitterSource: Processed 25,200 docs
19/01/15 22:29:50 WARN hdfs.HDFSEventSink: HDFS IO error
java.net.ConnectException: Call From localhost.localdomain/127.0.0.1 to
localhost:9000 failed on connection exception: java.net.ConnectException:
Connection refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native
Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructor
AccessorImpl.java:62)
        at

```

```
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
    at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
    at
org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:791)
    at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:731)
    at org.apache.hadoop.ipc.Client.call(Client.java:1474)
    at org.apache.hadoop.ipc.Client.call(Client.java:1401)
    at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:232)
    at com.sun.proxy.$Proxy13.create(Unknown Source)
    at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.create(ClientNamenodeProtocolTranslatorPB.java:295)
    at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
    at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:187)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
    at com.sun.proxy.$Proxy14.create(Unknown Source)
    at
org.apache.hadoop.hdfs.DFSOutputStream.newStreamForCreate(DFSOutputStream.java:1721)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1657)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1582)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSystem.java:397)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSystem.java:393)
    at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSystem.java:393)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSystem.java:337)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:908)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:889)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:786)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:775)
    at
org.apache.flume.sink.hdfs.HDFSDataStream.doOpen(HDFSDataStream.java:81
```

```

)
    at
org.apache.flume.sink.hdfs.HDFSDataStream.open(HDFSDataStream.java:108)
    at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:262)
    at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:252)
    at
org.apache.flume.sink.hdfs.BucketWriter$9$1.run(BucketWriter.java:701)
    at
org.apache.flume.auth.SimpleAuthenticator.execute(SimpleAuthenticator.java:50)
    at
org.apache.flume.sink.hdfs.BucketWriter$9.call(BucketWriter.java:698)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: java.net.ConnectException: Connection refused
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
    at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
    at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:609)
)
    at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:707)
    at
org.apache.hadoop.ipc.Client$Connection.access$2800(Client.java:370)
    at org.apache.hadoop.ipc.Client.getConnection(Client.java:1523)
    at org.apache.hadoop.ipc.Client.call(Client.java:1440)
    ... 33 more
19/01/15 22:29:51 INFO twitter.TwitterSource: Processed 25,300 docs
19/01/15 22:29:54 INFO twitter.TwitterSource: Processed 25,400 docs
19/01/15 22:29:55 INFO hdfs.BucketWriter: Creating
hdfs://localhost:9000/user/flume/tweets/FlumeData.1547571037728.tmp
19/01/15 22:29:56 WARN hdfs.HDFSEventSink: HDFS IO error
java.net.ConnectException: Call From localhost.localdomain/127.0.0.1 to
localhost:9000 failed on connection exception: java.net.ConnectException:
Connection refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused
    at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native
Method)
    at

```



```
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructor
AccessorImpl.java:62)
    at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingCon
structorAccessorImpl.java:45)
    at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
    at
org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:791)
    at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:731)
    at org.apache.hadoop.ipc.Client.call(Client.java:1474)
    at org.apache.hadoop.ipc.Client.call(Client.java:1401)
    at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngin
e.java:232)
    at com.sun.proxy.$Proxy13.create(Unknown Source)
    at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.cr
eate(ClientNamenodeProtocolTranslatorPB.java:295)
    at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
    at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccesso
rImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInv
ocationHandler.java:187)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocatio
nHandler.java:102)
    at com.sun.proxy.$Proxy14.create(Unknown Source)
    at
org.apache.hadoop.hdfs.DFSOutputStream.newStreamForCreate(DFSOutputStre
am.java:1721)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1657)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1582)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:397)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:393)
    at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResol
ver.java:81)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:393)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:337)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:908)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:889)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:786)
```

```

        at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:775)
        at
org.apache.flume.sink.hdfs.HDFSDataStream.doOpen(HDFSDataStream.java:81
)
        at
org.apache.flume.sink.hdfs.HDFSDataStream.open(HDFSDataStream.java:108)
        at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:262)
        at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:252)
        at
org.apache.flume.sink.hdfs.BucketWriter$9$1.run(BucketWriter.java:701)
        at
org.apache.flume.auth.SimpleAuthenticator.execute(SimpleAuthenticator.j
ava:50)
        at
org.apache.flume.sink.hdfs.BucketWriter$9.call(BucketWriter.java:698)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.ja
va:1149)
        at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.j
ava:624)
        at java.lang.Thread.run(Thread.java:748)
Caused by: java.net.ConnectException: Connection refused
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
        at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.j
ava:206)
        at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
        at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
        at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:609
)
        at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:707)
        at
org.apache.hadoop.ipc.Client$Connection.access$2800(Client.java:370)
        at org.apache.hadoop.ipc.Client.getConnection(Client.java:1523)
        at org.apache.hadoop.ipc.Client.call(Client.java:1440)
        ... 33 more
19/01/15 22:29:56 INFO twitter.TwitterSource: Processed 25,500 docs
19/01/15 22:29:59 INFO twitter.TwitterSource: Processed 25,600 docs
19/01/15 22:30:01 INFO twitter.TwitterSource: Processed 25,700 docs
19/01/15 22:30:01 INFO hdfs.BucketWriter: Creating
hdfs://localhost:9000/user/flume/tweets/FlumeData.1547571037729.tmp
19/01/15 22:30:02 WARN hdfs.HDFSEventSink: HDFS IO error
java.net.ConnectException: Call From localhost.localdomain/127.0.0.1 to
localhost:9000 failed on connection exception: java.net.ConnectException:
Connection refused; For more details see:

```

```
http://wiki.apache.org/hadoop/ConnectionRefused
    at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native
Method)
    at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructor
AccessorImpl.java:62)
    at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingCon
structorAccessorImpl.java:45)
    at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
    at
org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:791)
    at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:731)
    at org.apache.hadoop.ipc.Client.call(Client.java:1474)
    at org.apache.hadoop.ipc.Client.call(Client.java:1401)
    at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngin
e.java:232)
    at com.sun.proxy.$Proxy13.create(Unknown Source)
    at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.cr
eate(ClientNamenodeProtocolTranslatorPB.java:295)
    at sun.reflect.GeneratedMethodAccessor2.invoke(Unknown Source)
    at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccesso
rImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInv
ocationHandler.java:187)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocatio
nHandler.java:102)
    at com.sun.proxy.$Proxy14.create(Unknown Source)
    at
org.apache.hadoop.hdfs.DFSOutputStream.newStreamForCreate(DFSOutputStre
am.java:1721)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1657)
    at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1582)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:397)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:393)
    at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResol
ver.java:81)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:393)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
```

```

em.java:337)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:908)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:889)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:786)
    at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:775)
    at
org.apache.flume.sink.hdfs.HDFSDataStream.doOpen(HDFSDataStream.java:81
)
    at
org.apache.flume.sink.hdfs.HDFSDataStream.open(HDFSDataStream.java:108)
    at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:262)
    at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:252)
    at
org.apache.flume.sink.hdfs.BucketWriter$9$1.run(BucketWriter.java:701)
    at
org.apache.flume.auth.SimpleAuthenticator.execute(SimpleAuthenticator.j
ava:50)
    at
org.apache.flume.sink.hdfs.BucketWriter$9.call(BucketWriter.java:698)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.ja
va:1149)
    at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.j
ava:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: java.net.ConnectException: Connection refused
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
    at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.j
ava:206)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
    at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:609
)
    at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:707)
    at
org.apache.hadoop.ipc.Client$Connection.access$2800(Client.java:370)
    at org.apache.hadoop.ipc.Client.getConnection(Client.java:1523)
    at org.apache.hadoop.ipc.Client.call(Client.java:1440)
    ... 33 more
19/01/15 22:30:03 INFO twitter.TwitterSource: Processed 25,800 docs
19/01/15 22:30:05 INFO twitter.TwitterSource: Processed 25,900 docs
19/01/15 22:30:07 INFO twitter.TwitterSource: Processed 26,000 docs
19/01/15 22:30:07 INFO twitter.TwitterSource: Total docs indexed: 26,000,
total skipped docs: 0

```

```
19/01/15 22:30:07 INFO twitter.TwitterSource:      45 docs/second
19/01/15 22:30:07 INFO twitter.TwitterSource: Run took 573 seconds and
processed:
19/01/15 22:30:07 INFO twitter.TwitterSource:      0.012 MB/sec sent to
index
19/01/15 22:30:07 INFO twitter.TwitterSource:      7.083 MB text sent to
index
19/01/15 22:30:07 INFO twitter.TwitterSource: There were 0 exceptions
ignored:
19/01/15 22:30:08 INFO hdfs.BucketWriter: Creating
hdfs://localhost:9000/user/flume/tweets/FlumeData.1547571037730.tmp
19/01/15 22:30:09 WARN hdfs.HDFSEventSink: HDFS IO error
java.net.ConnectException: Call From localhost.localdomain/127.0.0.1 to
localhost:9000 failed on connection exception: java.net.ConnectException:
Connection refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused
    at sun.reflect.NativeConstructorAccessorImpl.newInstance0 (Native
Method)
    at
    sun.reflect.NativeConstructorAccessorImpl.newInstance (NativeConstructor
AccessorImpl.java:62)
    at
    sun.reflect.DelegatingConstructorAccessorImpl.newInstance (DelegatingCon
structorAccessorImpl.java:45)
    at java.lang.reflect.Constructor.newInstance (Constructor.java:423)
    at
    org.apache.hadoop.net.NetUtils.wrapWithMessage (NetUtils.java:791)
    at org.apache.hadoop.net.NetUtils.wrapException (NetUtils.java:731)
    at org.apache.hadoop.ipc.Client.call (Client.java:1474)
    at org.apache.hadoop.ipc.Client.call (Client.java:1401)
    at
    org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke (ProtobufRpcEngin
e.java:232)
    at com.sun.proxy.$Proxy13.create (Unknown Source)
    at
    org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.cr
eate (ClientNamenodeProtocolTranslatorPB.java:295)
    at sun.reflect.GeneratedMethodAccessor2.invoke (Unknown Source)
    at
    sun.reflect.DelegatingMethodAccessorImpl.invoke (DelegatingMethodAccesso
rImpl.java:43)
    at java.lang.reflect.Method.invoke (Method.java:498)
    at
    org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod (RetryInv
ocationHandler.java:187)
    at
    org.apache.hadoop.io.retry.RetryInvocationHandler.invoke (RetryInvocatio
nHandler.java:102)
    at com.sun.proxy.$Proxy14.create (Unknown Source)
    at
    org.apache.hadoop.hdfs.DFSOutputStream.newStreamForCreate (DFSOutputStre
am.java:1721)
    at org.apache.hadoop.hdfs.DFSClient.create (DFSClient.java:1657)
```

```
        at org.apache.hadoop.hdfs.DFSClient.create(DFSClient.java:1582)
        at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:397)
        at
org.apache.hadoop.hdfs.DistributedFileSystem$6.doCall(DistributedFileSy
stem.java:393)
        at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResol
ver.java:81)
        at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:393)
        at
org.apache.hadoop.hdfs.DistributedFileSystem.create(DistributedFileSyst
em.java:337)
            at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:908)
            at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:889)
            at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:786)
            at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:775)
            at
org.apache.flume.sink.hdfs.HDFSDataStream.doOpen(HDFSDataStream.java:81
)
            at
org.apache.flume.sink.hdfs.HDFSDataStream.open(HDFSDataStream.java:108)
            at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:262)
            at
org.apache.flume.sink.hdfs.BucketWriter$1.call(BucketWriter.java:252)
            at
org.apache.flume.sink.hdfs.BucketWriter$9$1.run(BucketWriter.java:701)
            at
org.apache.flume.auth.SimpleAuthenticator.execute(SimpleAuthenticator.j
ava:50)
            at
org.apache.flume.sink.hdfs.BucketWriter$9.call(BucketWriter.java:698)
            at java.util.concurrent.FutureTask.run(FutureTask.java:266)
            at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.ja
va:1149)
            at
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.j
ava:624)
            at java.lang.Thread.run(Thread.java:748)
Caused by: java.net.ConnectException: Connection refused
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:717)
        at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.j
ava:206)
            at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
            at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
```

```
        at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:609
)
        at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:707)
        at
org.apache.hadoop.ipc.Client$Connection.access$2800(Client.java:370)
        at org.apache.hadoop.ipc.Client.getConnection(Client.java:1523)
        at org.apache.hadoop.ipc.Client.call(Client.java:1440)
        ... 33 more
19/01/15 22:30:10 INFO twitter.TwitterSource: Processed 26,100 docs
Exception in thread "Twitter Stream consumer-1[Receiving stream]"
java.lang.OutOfMemoryError: Java heap space
        at java.util.Arrays.copyOf(Arrays.java:3332)
        at
java.lang.AbstractStringBuilder.ensureCapacityInternal(AbstractStringBu
ilder.java:124)
        at
java.lang.AbstractStringBuilder.append(AbstractStringBuilder.java:596)
        at java.lang.StringBuffer.append(StringBuffer.java:367)
        at java.io.BufferedReader.readLine(BufferedReader.java:358)
        at java.io.BufferedReader.readLine(BufferedReader.java:389)
        at
twitter4j.StatusStreamBase.handleNextElement(StatusStreamBase.java:85)
        at twitter4j.StatusStreamImpl.next(StatusStreamImpl.java:57)
        at
twitter4j.TwitterStreamImpl$TwitterStreamConsumer.run(TwitterStreamImpl
.java:478)
Exception in thread "Twitter4J Async Dispatcher[0]"
java.lang.OutOfMemoryError: Java heap space
        at java.util.HashMap.newNode(HashMap.java:1747)
        at java.util.HashMap.putVal(HashMap.java:631)
        at java.util.HashMap.put(HashMap.java:612)
        at twitter4j.internal.org.json.JSONObject.put(JSONObject.java:765)
        at
twitter4j.internal.org.json.JSONObject.putOnce(JSONObject.java:788)
        at
twitter4j.internal.org.json.JSONObject.<init>(JSONObject.java:207)
        at
twitter4j.internal.org.json.JSONTokener.nextValue(JSONTokener.java:297)
        at
twitter4j.internal.org.json.JSONObject.<init>(JSONObject.java:207)
        at
twitter4j.internal.org.json.JSONObject.<init>(JSONObject.java:311)
        at twitter4j.StatusStreamBase$1.run(StatusStreamBase.java:104)
        at
twitter4j.internal.async.ExecuteThread.run(DispatcherImpl.java:116)
19/01/15 22:30:15 ERROR hdfs.HDFSEventSink: process failed
java.lang.OutOfMemoryError: Java heap space
        at
java.util.zip.InflaterInputStream.<init>(InflaterInputStream.java:88)
        at
java.util.zip.ZipFile$ZipFileInflaterInputStream.<init>(ZipFile.java:40
```

```

8)
    at java.util.zip.ZipFile.getInputStream(ZipFile.java:389)
    at java.util.jar.JarFile.getInputStream(JarFile.java:447)
    at
sun.net.www.protocol.jar.JarURLConnection.getInputStream(JarURLConnection.java:162)
    at java.net.URL.openStream(URL.java:1045)
    at
org.apache.hadoop.conf.Configuration.parse(Configuration.java:2420)
    at
org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2491)
    at
org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2444)
    at
org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2361)
    at
org.apache.hadoop.conf.Configuration.set(Configuration.java:1099)
    at
org.apache.hadoop.conf.Configuration.set(Configuration.java:1071)
    at
org.apache.hadoop.conf.Configuration.setBoolean(Configuration.java:1409)
)
    at
org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:226)
    at
org.apache.flume.sink.hdfs.BucketWriter.append(BucketWriter.java:541)
    at
org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:401)
)
    at
org.apache.flume.sink.DefaultSinkProcessor.process(DefaultSinkProcessor.java:67)
    at
org.apache.flume.SinkRunner$PollingRunner.run(SinkRunner.java:145)
    at java.lang.Thread.run(Thread.java:748)
Exception in thread "SinkRunner-PollingRunner-DefaultSinkProcessor"
java.lang.OutOfMemoryError: Java heap space
    at
java.util.zip.InflaterInputStream.<init>(InflaterInputStream.java:88)
    at
java.util.zip.ZipFile$ZipFileInflaterInputStream.<init>(ZipFile.java:408)
8)
    at java.util.zip.ZipFile.getInputStream(ZipFile.java:389)
    at java.util.jar.JarFile.getInputStream(JarFile.java:447)
    at
sun.net.www.protocol.jar.JarURLConnection.getInputStream(JarURLConnection.java:162)
    at java.net.URL.openStream(URL.java:1045)
    at
org.apache.hadoop.conf.Configuration.parse(Configuration.java:2420)
    at

```



```
org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:24
91)
    at
org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2
444)
    at
org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2361)
    at
org.apache.hadoop.conf.Configuration.set(Configuration.java:1099)
    at
org.apache.hadoop.conf.Configuration.set(Configuration.java:1071)
    at
org.apache.hadoop.conf.Configuration.setBoolean(Configuration.java:1409
)
    at
org.apache.flume.sink.hdfs.BucketWriter.open(BucketWriter.java:226)
    at
org.apache.flume.sink.hdfs.BucketWriter.append(BucketWriter.java:541)
    at
org.apache.flume.sink.hdfs.HDFSEventSink.process(HDFSEventSink.java:401
)
    at
org.apache.flume.sink.DefaultSinkProcessor.process(DefaultSinkProcessor
.java:67)
    at
org.apache.flume.SinkRunner$PollingRunner.run(SinkRunner.java:145)
    at java.lang.Thread.run(Thread.java:748)
```

STEP 17: WE CAN CHECK THE TWEET DATA BY TYPING THE BELOW COMMAND IN HDFS.
REFER SCREENSHOT

COMMAND: `hadoop fs -cat /user/flume/tweets/FlumeData.1450940925854`

EXPLANATION: YOU CAN SEE THE TWITTER DATA HAS BEEN STORED INTO HDFS. REFER
SCREENSHOT

```
... https://t.co/rKcFWoUcc #education", "contributors": null, "geo": null, "entitie
": {"symbols": [], "urls": [{"expanded_url": "http://bit.ly/1QJreHR", "indices": [103,
26], "display_url": "bit.ly/1QJreHR", "url": "https://t.co/rKcFWoUcc"}], "hashtags"
[{"text": "education", "indices": [127, 137]}], "user_mentions": [], "is_quote_status
": false, "source": "<a href='\"http://twitterfeed.com\"' rel='\"nofollow\">twitterfeed
</a>", "favorited": false, "in_reply_to_user_id": null, "retweet_count": 0, "id_str":
679921891398660096", "user": {"location": "Coventry", "default_profile": false, "prof
le_background_tile": false, "statuses_count": 33000, "lang": "en", "profile_link colo
": "2FC2EF", "profile_banner_url": "https://pbs.twimg.com/profile_banners/25467890
4/1441620563", "id": 2546789014, "following": null, "protected": false, "favourites_co
nt": 508, "profile_text_color": "666666", "verified": false, "description": "Amateur t
nerd . Working", "contributors_enabled": false, "profile_sidebar_border_color": "1
1A1E", "name": "Angela Dubravski", "profile_background_color": "1A1B1F", "created_at
": "Wed May 14 06:28:35 +0000 2014", "default_profile_image": false, "followers_coun
": 729, "profile_image_url_https": "https://pbs.twimg.com/profile_images/659502566
54627328/sfQgy6LM_normal.jpg", "geo_enabled": false, "profile_background_image_url
": "http://abs.twimg.com/images/themes/theme9/bg.gif", "profile_background_image_u
l_https": "https://abs.twimg.com/images/themes/theme9/bg.gif", "follow_request se
t": null, "url": null, "utc_offset": null, "time_zone": null, "notifications": null, "pro
file_use_background_image": true, "friends_count": 1192, "profile_sidebar_fill_color
": "252429", "screen_name": "citaCrowl", "id_str": "2546789014", "profile_image_url": "
http://pbs.twimg.com/profile_images/659502566854627328/sfQgy6LM_normal.jpg", "lis
ed_count": 91, "is_translator": false}}

... https://t.co/rKcFWoUcc #education", "contributors": null, "geo": null, "entitie
": {"symbols": [], "urls": [{"expanded_url": "http://bit.ly/1QJreHR", "indices": [103,
26], "display_url": "bit.ly/1QJreHR", "url": "https://t.co/rKcFWoUcc"}], "hashtags"
[{"text": "education", "indices": [127, 137]}], "user_mentions": [], "is_quote_status
": false, "source": "<a href='\"http://twitterfeed.com\"' rel='\"nofollow\">twitterfeed
</a>", "favorited": false, "in_reply_to_user_id": null, "retweet_count": 0, "id_str":
679921891398660096", "user": {"location": "Coventry", "default_profile": false, "prof
le_background_tile": false, "statuses_count": 33000, "lang": "en", "profile_link colo
": "2FC2EF", "profile_banner_url": "https://pbs.twimg.com/profile_banners/25467890
4/1441620563", "id": 2546789014, "following": null, "protected": false, "favourites_co
nt": 508, "profile_text_color": "666666", "verified": false, "description": "Amateur t
nerd . Working", "contributors_enabled": false, "profile_sidebar_border_color": "1
1A1E", "name": "Angela Dubravski", "profile_background_color": "1A1B1F", "created_at
": "Wed May 14 06:28:35 +0000 2014", "default_profile_image": false, "followers_coun
": 729, "profile_image_url_https": "https://pbs.twimg.com/profile_images/659502566
54627328/sfQgy6LM_normal.jpg", "geo_enabled": false, "profile_background_image_url
": "http://abs.twimg.com/images/themes/theme9/bg.gif", "profile_background_image_u
l_https": "https://abs.twimg.com/images/themes/theme9/bg.gif", "follow_request se
t": null, "url": null, "utc_offset": null, "time_zone": null, "notifications": null, "pro
file_use_background_image": true, "friends_count": 1192, "profile_sidebar_fill_color
": "252429", "screen_name": "citaCrowl", "id_str": "2546789014", "profile_image_url": "
http://pbs.twimg.com/profile_images/659502566854627328/sfQgy6LM_normal.jpg", "lis
ed_count": 91, "is_translator": false}}
```