BIG DATA HADOOP & SPARK TRAINING

Assignment on Oozie and Flume

APRIL 8, 2018

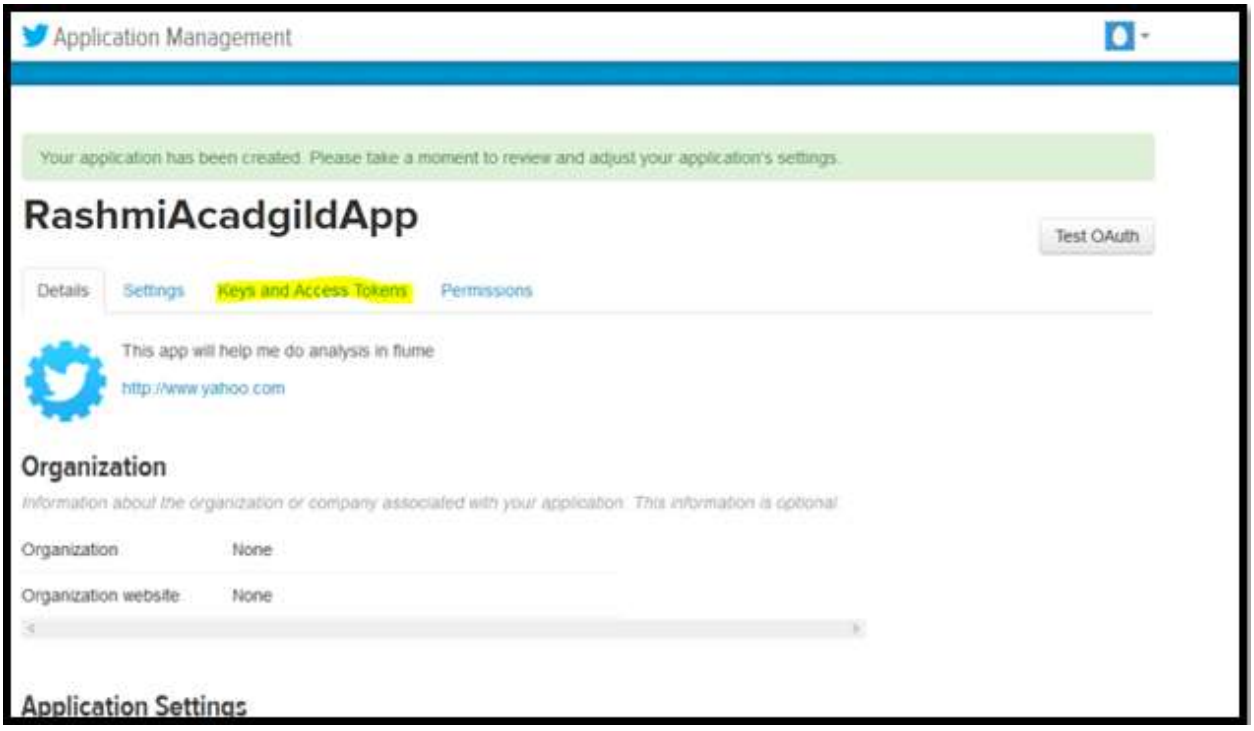
RASHMI KRISHNA

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

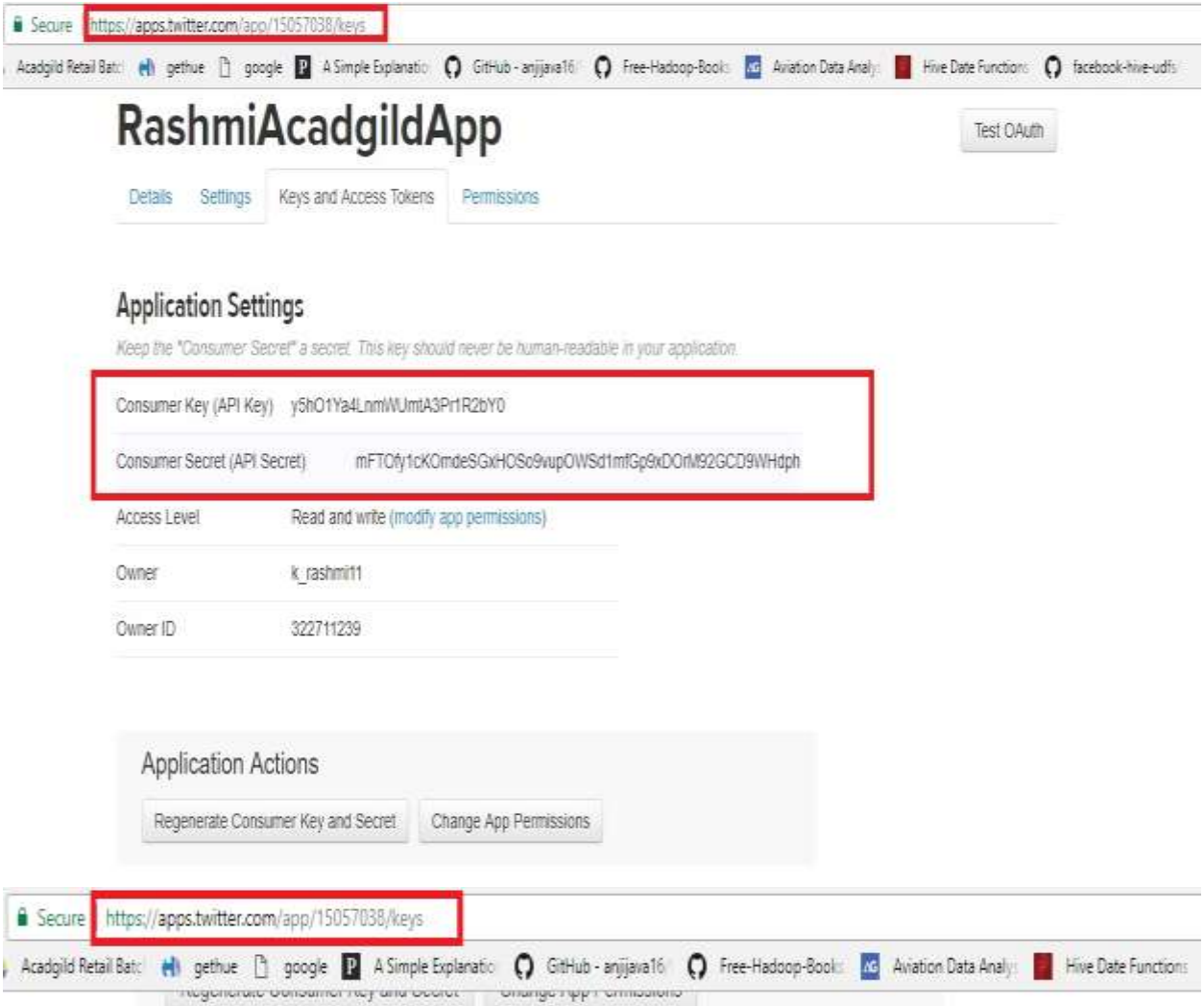
To create a flume agent, to stream the data from Twitter, first go to your twitter account and create an application from the below link:

https://apps.twitter.com/app

Provide all the required credentials to create a app.
On successful credential authorization, you will get your application screen, from which go to "Keys and Access Tokens" tab.



In this screen copy the required information, i.e. copy consumer key, consumer secret, token and token secret



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	322711239-h9wo8AnX9AKMzgcEWSMOaEJ91vXH6H0lJb6qyHhb
Access Token Secret	DyWtlf3fJTJ1UtFLBP0QVpK3dZMmTBjsRJdz2sAynhA9u
Access Level	Read and write
Owner	k_rashmi11
Owner ID	322711239

Use the information in the flume.conf_twitter file

flume.conf_twitter - Notepad

File Edit Format View Help

```
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=y5hO1Ya4LnmWUmtA3Pr1R2bY0
TwitterAgent.sources.Twitter.consumerSecret=mFTOfy1cKOmdeSGxHOSo9vupOWSd1mfGp9xDOrM92GCD9WHdph
TwitterAgent.sources.Twitter.accessToken=322711239-h9wo8AnX9AKMzgcEWSMOaEJ91vXH6H01Jb6qyHhb
TwitterAgent.sources.Twitter.accessTokenSecret=DyWtlf3fJTJ1UtFLBP0QVpK3dZMmTBjsRJdz2sAynhA9u
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:8020/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
```

Consumerkey,consumersecret, Token and Tokensecret access keys

HDFS output file path

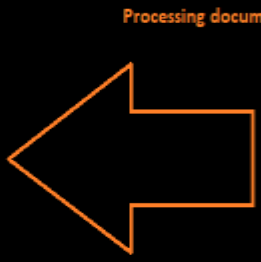
Stream the Twitter data using flume using the below command:

flume-ng agent --conf c -n TwitterAgent -f /home/acadgild/flume/flume.conf_twitter

```
[acadgild@localhost ~]$ flume-ng agent --conf c  -n TwitterAgent -f /home/acadgild/flume/flume.conf twitter
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 'c:/home/acadgild/install/flume/apache-flume-1.8.0-bin/lib/*:/home/acadgild/install/
hadoop/hadoop-2.6.5/contrib/capacity-scheduler/*.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/etc/hadoop/:/home/acadgild/install/hado
op/hadoop-2.6.5/share/hadoop/common/lib/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/*:/home/acadgild/install/hadoo
p/hadoop-2.6.5/share/hadoop/hdfs:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/hdfs/lib/*:/home/acadgild/install/hadoop/hadoo
p-2.6.5/share/hadoop/hdfs/*:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/yarn/lib/*:/home/acadgild/install/hadoop/hadoop-2.6
.5/share/hadoop/yarn/*:/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/hbase/hbase-1.2.6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/h
base/hbase-1.2.6:/home/acadgild/install/hbase/hbase-1.2.6/lib/activation-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopallian
ce-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apache
ds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.0.0-M20.jar:/home/acadgild/install/hbase/hb
ase-1.2.6/lib/api-util-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/
lib/avro-1.7.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/li
b/commons-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install/hbase/hbase-
1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-collections-3.2.2.jar:/home/acadgild/install/hbase
/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configuration-1.6.jar:/home/acadgild/i
nstall/hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-digester-1.8.jar:/home/acad
gild/install/hbase/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-httpclient-3.1.jar:/home/aca
dgild/install/hbase/hbase-1.2.6/lib/commons-io-2.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild
/install/hbase/hbase-1.2.6/lib/commons-logging-1.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild
/install/hbase/hbase-1.2.6/lib/commons-math3-3.1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/disruptor-3.3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/findbugs-annotations-1.3.9-1.jar:/home/a
cadgild/install/hbase/hbase-1.2.6/lib/guava-12.0.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/instal
l/hbase/hbase-1.2.6/lib/guice-servlet-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-annotations-2.5.1.jar:/home/acadgild/
install/hbase/hbase-1.2.6/lib/hadoop-auth-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-client-2.5.1.jar:/home/acadgild
/install/hbase/hbase-1.2.6/lib/hadoop-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-hdfs-2.5.1.jar:/home/acadgil
d/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-app-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-cli
ent-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-core-2.5.1.jar:/home/acadgild/install/hbase/h
base-1.2.6/lib/hadoop-mapreduce-client-jobclient-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-mapreduce-client-shuffle
-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-api-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-y
arn-client-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hadoop-yarn-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6
/lib/hadoop-yarn-server-common-2.5.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-annotations-1.2.6.jar:/home/acadgild/instal
l/hbase/hbase-1.2.6/lib/hbase-annotations-1.2.6-tests.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-client-1.2.6.jar:/home/aca
dgild/install/hbase/hbase-1.2.6/lib/hbase-common-1.2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/hbase-common-1.2.6-tests.jar:/h
```

```
18/04/08 16:35:28 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
18/04/08 16:35:28 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/acadgild/flume/flume.conf_tw
itter
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: TwitterAgent
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Processing:HDFS
18/04/08 16:35:28 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAgent]
18/04/08 16:35:28 INFO node.AbstractConfigurationProvider: Creating channels
18/04/08 16:35:28 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
18/04/08 16:35:28 INFO node.AbstractConfigurationProvider: Created channel MemChannel
18/04/08 16:35:28 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.TwitterSo
urce
18/04/08 16:35:29 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
18/04/08 16:35:29 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
18/04/08 16:35:29 INFO node.Application: Starting new configuration:{ sourceRunners:{Twitter=EventDrivenSourceRunner: { source:org.apac
he.flume.source.twitter.TwitterSource{name:Twitter,state:IDLE} }} sinkRunners:{HDFS=SinkRunner: { policy:org.apache.flume.sink.DefaultS
inkProcessor@1e62b4a counterGroup:{ name:null counters:{} } }} channels:{MemChannel=org.apache.flume.channel.MemoryChannel{name: MemCha
nnel}} }
18/04/08 16:35:29 INFO node.Application: Starting Channel MemChannel
18/04/08 16:35:29 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully
 registered new MBean.
18/04/08 16:35:29 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
18/04/08 16:35:29 INFO node.Application: Starting Sink HDFS
18/04/08 16:35:29 INFO node.Application: Starting Source Twitter
18/04/08 16:35:29 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource{name:Twitter,state:
IDLE} ...
18/04/08 16:35:29 INFO twitter.TwitterSource: Twitter source Twitter started.
18/04/08 16:35:29 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/04/08 16:35:29 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully register
ed new MBean.
18/04/08 16:35:29 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
```

```
18/04/08 16:35:39 INFO twitter.TwitterSource: Processed 200 docs
18/04/08 16:35:41 INFO twitter.TwitterSource: Processed 300 docs
18/04/08 16:35:44 INFO twitter.TwitterSource: Processed 400 docs
18/04/08 16:35:47 INFO twitter.TwitterSource: Processed 500 docs
18/04/08 16:35:50 INFO twitter.TwitterSource: Processed 600 docs
18/04/08 16:35:52 INFO twitter.TwitterSource: Processed 700 docs
18/04/08 16:35:55 INFO twitter.TwitterSource: Processed 800 docs
18/04/08 16:35:58 INFO twitter.TwitterSource: Processed 900 docs
18/04/08 16:36:01 INFO twitter.TwitterSource: Processed 1,000 docs
18/04/08 16:36:01 INFO twitter.TwitterSource: Total docs indexed: 1,000, total skipped docs: 0
18/04/08 16:36:01 INFO twitter.TwitterSource:     32 docs/second
18/04/08 16:36:01 INFO twitter.TwitterSource: Run took 31 seconds and processed:
18/04/08 16:36:01 INFO twitter.TwitterSource:     0.009 MB/sec sent to index
18/04/08 16:36:01 INFO twitter.TwitterSource:     0.27 MB text sent to index
18/04/08 16:36:01 INFO twitter.TwitterSource: There were 0 exceptions ignored:
18/04/08 16:36:03 INFO twitter.TwitterSource: Processed 1,100 docs
18/04/08 16:36:06 INFO twitter.TwitterSource: Processed 1,200 docs
18/04/08 16:36:08 INFO twitter.TwitterSource: Processed 1,300 docs
18/04/08 16:36:11 INFO twitter.TwitterSource: Processed 1,400 docs
18/04/08 16:36:14 INFO twitter.TwitterSource: Processed 1,500 docs
18/04/08 16:36:16 INFO twitter.TwitterSource: Processed 1,600 docs
18/04/08 16:36:18 INFO twitter.TwitterSource: Processed 1,700 docs
18/04/08 16:36:21 INFO twitter.TwitterSource: Processed 1,800 docs
18/04/08 16:36:23 INFO twitter.TwitterSource: Processed 1,900 docs
18/04/08 16:36:26 INFO twitter.TwitterSource: Processed 2,000 docs
18/04/08 16:36:26 INFO twitter.TwitterSource: Total docs indexed: 2,000, total skipped docs: 0
18/04/08 16:36:26 INFO twitter.TwitterSource:     35 docs/second
18/04/08 16:36:26 INFO twitter.TwitterSource: Run took 56 seconds and processed:
18/04/08 16:36:26 INFO twitter.TwitterSource:     0.01 MB/sec sent to index
18/04/08 16:36:26 INFO twitter.TwitterSource:     0.539 MB text sent to index
18/04/08 16:36:26 INFO twitter.TwitterSource: There were 0 exceptions ignored:
18/04/08 16:36:28 INFO twitter.TwitterSource: Processed 2,100 docs
18/04/08 16:36:31 INFO twitter.TwitterSource: Processed 2,200 docs
18/04/08 16:36:33 INFO twitter.TwitterSource: Processed 2,300 docs
18/04/08 16:36:36 INFO twitter.TwitterSource: Processed 2,400 docs
18/04/08 16:36:38 INFO twitter.TwitterSource: Processed 2,500 docs
18/04/08 16:36:41 INFO twitter.TwitterSource: Processed 2,600 docs
18/04/08 16:36:44 INFO twitter.TwitterSource: Processed 2,700 docs
```




```
663
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.stop.time == 15231866470
27
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.batch.complete == 0
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.batch.empty == 60
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.batch.underflow == 1
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.connection.closed.count
== 2
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.connection.creation.coun
t == 2
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.connection.failed.count
== 0
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.event.drain.attempt == 6
21
18/04/08 16:54:07 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: HDFS. sink.event.drain.sucess == 62
1
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /user/flume/tweets
18/04/08 16:54:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
e applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup   11430392 2018-04-08 16:45 /user/flume/tweets/FlumeData.1523185534093
-rw-r--r--   1 acadgild supergroup     881617 2018-04-08 16:54 /user/flume/tweets/FlumeData.1523186140569
[acadgild@localhost ~]$ hadoop fs -cat /user/flume/tweets/FlumeData.1523185534093
```

Tweets retrieved from twitter, saved in HDFS is read using the cat command:

— 지능은 어린 아이 or 실제 동물 마냥 짧은 생각
— 단어 의미 잘 모르는
— 말 오지 않는일 시간의 종결 서 함 = Ekuangkaooo(2018-04-08T16:42:27Z6@immure__ 찰칵 ,
찰칵 ... Twitter for Android82939223492771841 南→北 三島北
→福井県立大学(経済-経済)
LINE担当(自称)
いきものがかり/ポルノグラフィティ
騒いでるやついたらそいつです はらーじゅ【 feat Daiki Hara】 from_haraju(2018-04-08T16:42:27Z寒いんだよ～ 静岡県民にはきつい
(半袖 裸足 暖房なし) Twitter Lite$982939223459287042 ~振り向くな 後ろには
明日はないから 前を向け wpNobutaka SudoNbtk_Sd(2018-04-08T16:42:27ZPRT @muramura005: https://t.co/hNx7i10335 <a href="http://twitter.co
m/download/android" rel="nofollow">Twitter for Android^https://pbs.twimg.com/media/C2NNvguUQAANLEI.jpg https://twitter.com/SHINKAN3
4721410/status/820585041122050052/photo/1$982939223492771840 101 =♡ #워너원 #옹성우 #윤지성 #강다니엘 #박우진 #유연정
OR●or_auraya(2018-04-08T16:42:27Z RT @focuson_osw: 180408 팬사인회 #옹성우 #워너원 https://t.co/lKKlrPGjIg <a href="http://twitter.com/
download/android" rel="nofollow">Twitter for Android^https://pbs.twimg.com/media/DaQWgcTUQAAhOIi.jpg https://twitter.com/focuson_os
w/status/982935377521475584/photo/1$982939223484334085z6Republic of the Philippines,I AM PROUD FANBOY ARMY Rodgen AbacaroAbacaroRodge
n(2018-04-08T16:42:27Z RT @ARMYCLOCK_8AM: []Daily Reminder[]

Have you visited the BTS wiki page yet today?

https://t.co/2ANkIrqIk8

Browse it on the DESKTOP… Twitter for Android$982939223467745280 Islas
CanariasTToda la información de Canarias a diario.
Canarias7canarias7(2018-04-08T16:42:27Z Ha sido fundamental neutralizar el 1-0. No hay tanta ansiedad y se nota. Aunque, recordemos, el
empate vale de bien… https://t.co/k29hoSGCSt Twitter Web Client$982939223459287041 4.J
im Mitchellopseu616(2018-04-08T16:42:27Z RT @cmjackel: All 5 Acting Managers from #SarniaJail @OPSEU128 have confirmed that they are st
epping down. Thanks to all 5! #Solid #CrisisI… Twitter for Android$982
939223484522497 }
XVIIIN سالب جمعهم ونتائج صبري مل الأ ع جامع و دايب خيط ق متعلق و .. غايب لكن الحاضر أنا (: ! ! Salma "salma_seleman(2018-04-08T16:42:
72Z RT @Ayman_El_Teer: إذا فكرت بالمتاعب أسرعت إليك وإذا فكرت باللذات أسرعت إليك تقول أنا نمت فتنام أنا مرضت فتمرض أنا شفيت فتشفى أنا
a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android$982939223484452864 *Tepito, México, D.F.> انتهيت فـ
Arquitecto, Urbanista y vago de nacimiento, Creador de la filosofía de vida: Provocar el intercambio del conocimiento entre l@s que sa
ben l@s que saben Gabriel SánchezElarkidelbarrio(2018-04-08T16:42:27Z RT @hipmaricopa: 5 signs that a public space is welcoming to al
l https://t.co/xGcJnWXwoc

#placemaking #publicspaces #community #inclusion Twitter for iPhone$982
939223463313409 #SHINee ♥♥♥ TaeChing
♥ching_taemin(2018-04-08T16:42:27Z RT @violita_tiara: