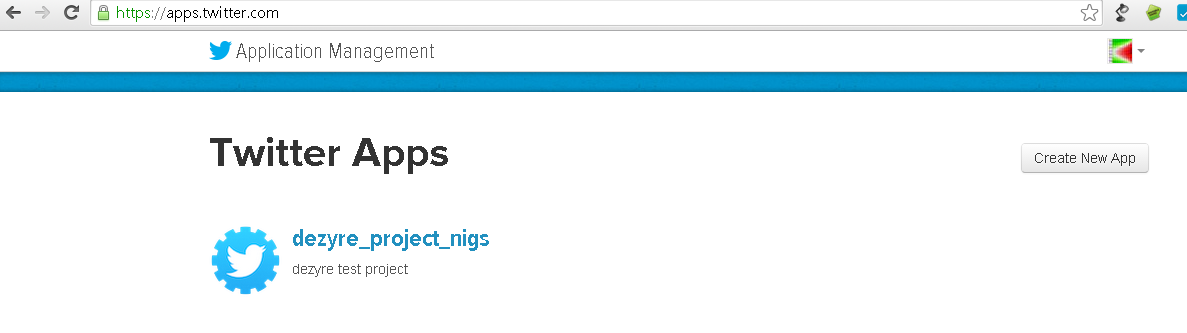
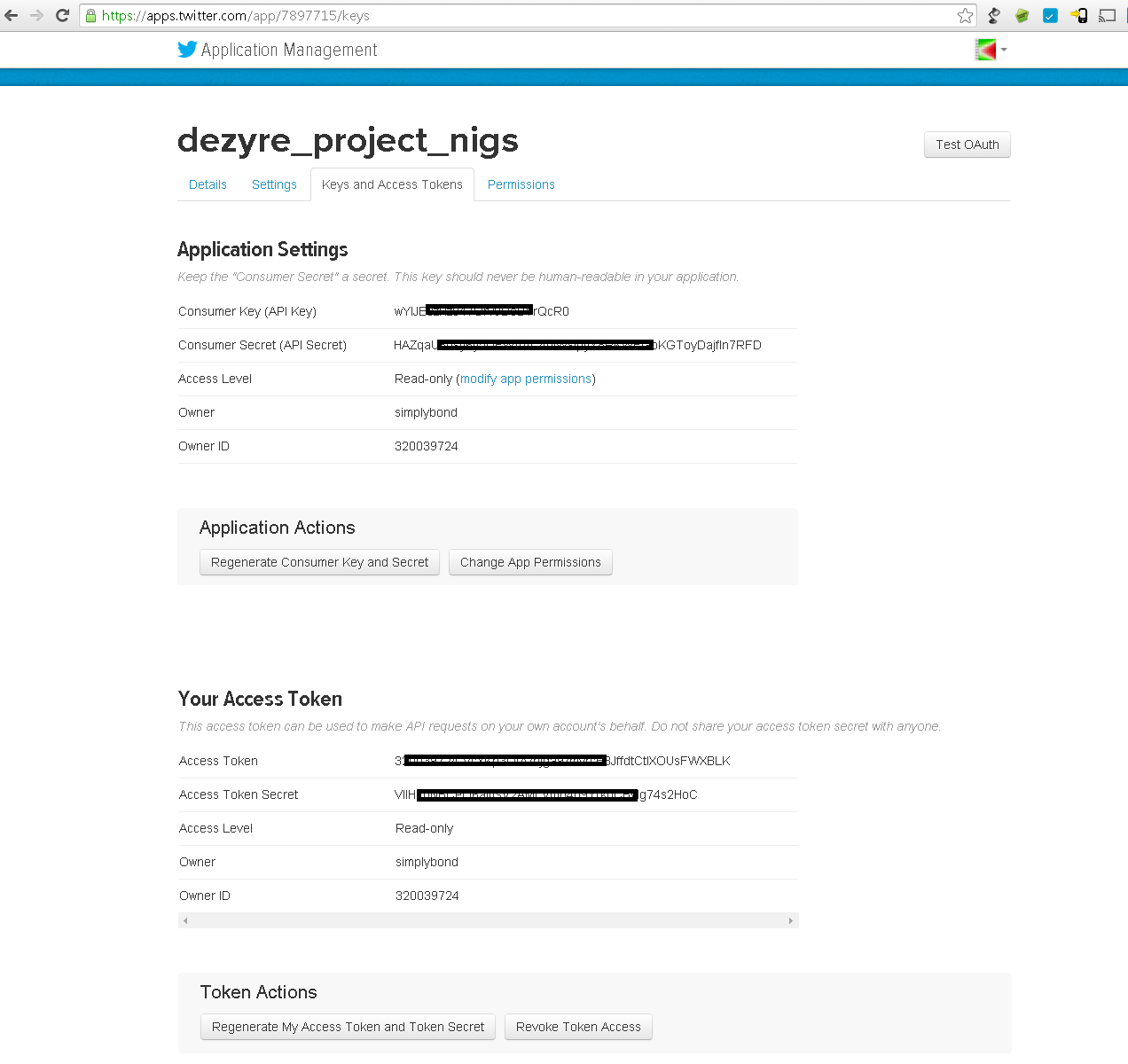
**StockTweetsAnalysis Project**

**Problem Statement: To analyze tweets for fortune 100 companies from S&P 500 index. Below questions are asked to the data**

1. **What are the most discussed stocks on twitter?**
2. **Who are the twitter users who tweet about fortune 100 companies and have most number of followers on twitter?**
3. **What are the trending words or Hashtags investment/trader community is talking about?**
4. **What are 10 high frequency two word combinations among tweets?**

# Created twitter application



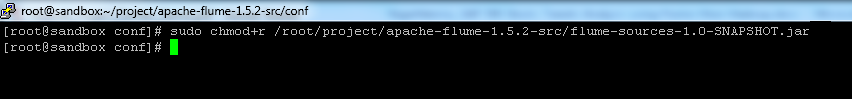


# Flume Agent Setup

**Downloaded flume agent from apache-**



**Building flume source and copying flume library-**

****

**Configuring flume.conf –**

Updated below lines with keys from twitter application

TwitterAgent.sources.Twitter.consumerKey

TwitterAgent.sources.Twitter.consumerSecret

TwitterAgent.sources.Twitter.accessToken

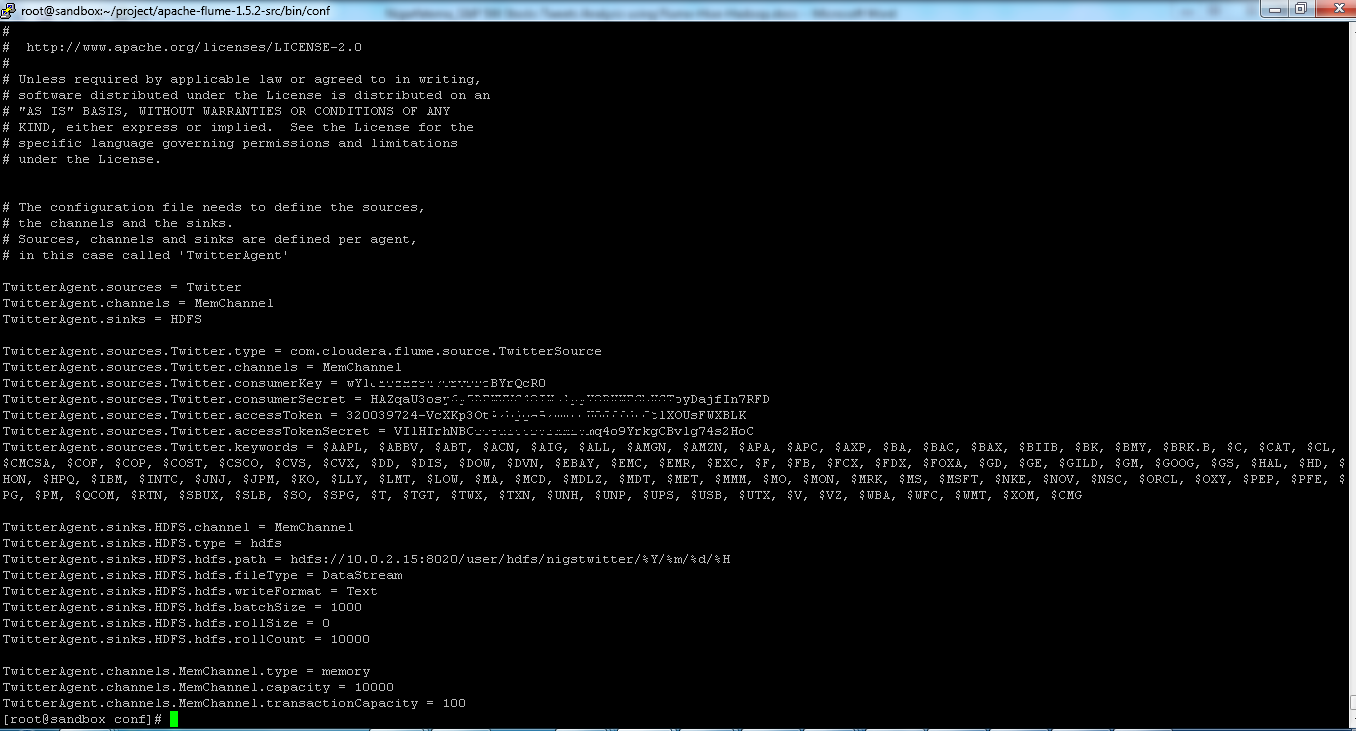
TwitterAgent.sources.Twitter.accessTokenSecret

Updated below line with path of Hortonworks Sandbox HDFS directory where tweets would go-

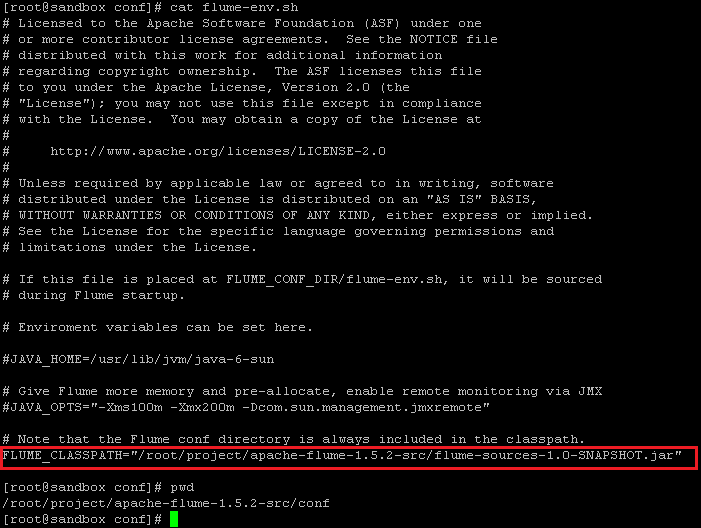
TwitterAgent.sources.Twitter.keywords

**Keywords used**: Keywords used are the ticker symbols of fortune 100 companies from S&P 500 index. List of the ticker symbols is as below-

**$AAPL, $ABBV, $ABT, $ACN, $AIG, $ALL, $AMGN, $AMZN, $APA, $APC, $AXP, $BA, $BAC, $BAX, $BIIB, $BK, $BMY, $BRK.B, $C, $CAT, $CL, $CMCSA, $COF, $COP, $COST, $CSCO, $CVS, $CVX, $DD, $DIS, $DOW, $DVN, $EBAY, $EMC, $EMR, $EXC, $F, $FB, $FCX, $FDX, $FOXA, $GD, $GE, $GILD, $GM, $GOOG, $GS, $HAL, $HD, $HON, $HPQ, $IBM, $INTC, $JNJ, $JPM, $KO, $LLY, $LMT, $LOW, $MA, $MCD, $MDLZ, $MDT, $MET, $MMM, $MO, $MON, $MRK, $MS, $MSFT, $NKE, $NOV, $NSC, $ORCL, $OXY, $PEP, $PFE, $PG, $PM, $QCOM, $RTN, $SBUX, $SLB, $SO, $SPG, $T, $TGT, $TWX, $TXN, $UNH, $UNP, $UPS, $USB, $UTX, $V, $VZ, $WBA, $WFC, $WMT, $XOM, $CMG**

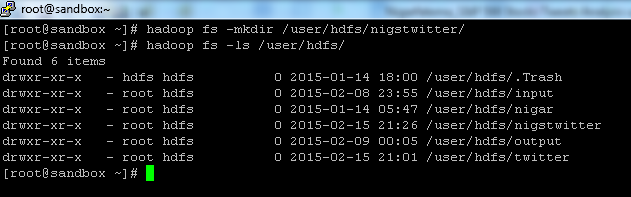
****

**Configuring flume-env.sh-**

****

# HDFS Directory

Created following HDFS directory to receive tweet files “/user/hdfs/nigstwitter/”

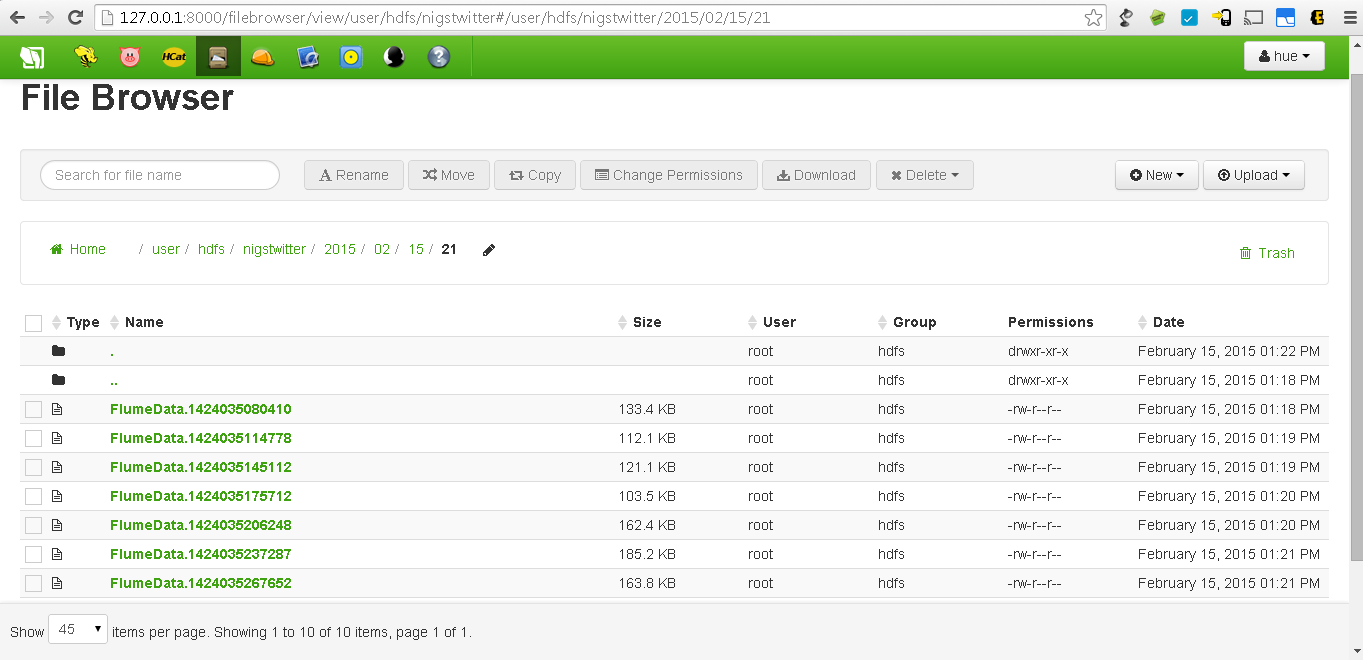


# Flume execution

**Below command is used to execute flume**

flume-ng agent -n TwitterAgent -c conf -f /root/project/apache-flume-1.5.2-src/bin/conf/flume.conf

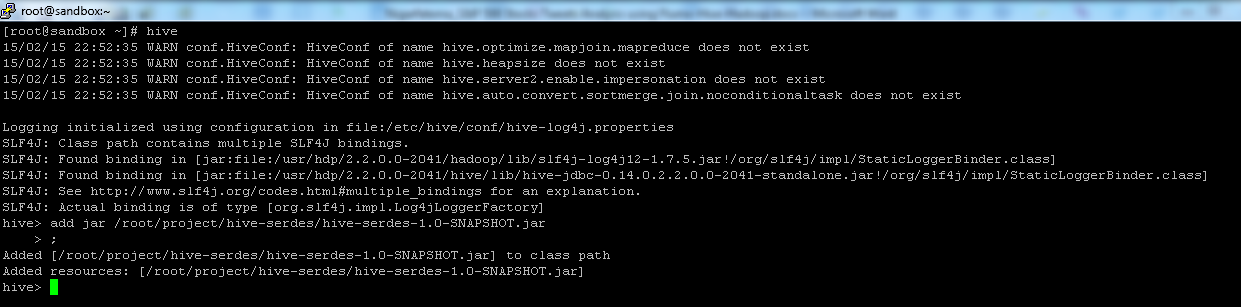
**After flume execution twitter files arrived in HDFS** “/user/hdfs/nigstwitter/



# Hive Setup

1. Log in to Hive
2. Add Hive SerDe jar file

add jar /root/project/hive-serdes/hive-serdes-1.0-SNAPSHOT.jar



1. **Create Hive table to load tweets**

**Use below create table script to create table**

CREATE EXTERNAL TABLE stocktweets (

id BIGINT,

created\_at STRING,

source STRING,

favorited BOOLEAN,

retweet\_count INT,

retweeted\_status STRUCT<

text:STRING,

user:STRUCT<screen\_name:STRING,name:STRING>>,

entities STRUCT<

urls:ARRAY<STRUCT<expanded\_url:STRING>>,

user\_mentions:ARRAY<STRUCT<screen\_name:STRING,name:STRING>>,

hashtags:ARRAY<STRUCT<text:STRING>>>,

text STRING,

user STRUCT<

screen\_name:STRING,

name:STRING,

friends\_count:INT,

followers\_count:INT,

statuses\_count:INT,

verified:BOOLEAN,

utc\_offset:INT,

time\_zone:STRING>,

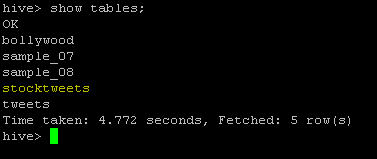
in\_reply\_to\_screen\_name STRING

)

PARTITIONED BY (datehour INT)

ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'

LOCATION '/user/hdfs/nigstwitter';



**Load stocktweets table with tweets**

LOAD DATA INPATH '/user/hdfs/nigstwitter/2015/02/15/21'

INTO TABLE `default.stocktweets`

PARTITION (datehour='2015021521');

1. **Analyzing data using HiveQL**

**Analysis Query 1:**

|  |  |
| --- | --- |
| **Objective** | **To know top 50 users those have the most number of followers** |
| **Query** | select distinct user.screen\_name, user.followers\_count c from stocktweets order by c desc limit 10; |
| **Result** |  |
| **Execution**  **Screenshot** |  |

**Analysis Query 2:**

|  |  |
| --- | --- |
| **Objective** | **To know list of trending words or Hashtags investment/trader community is talking about** |
| **Query** | select distinct t.c from (select entities.hashtags c from stocktweets) t; |
| **Result** |  |
| **Execution**  **Screenshot** |  |

**Analysis Query 3:**

|  |  |
| --- | --- |
| **Objective** | **To get a list of most frequent two word combinations in tweets. Need top 10 high frequency two word combinations among tweets** |
| **Query** | SELECT explode(ngrams(sentences(lower(text)), 2, 10)) FROM stocktweets; |
| **Result** |  |
| **Execution**  **Screenshot** |  |

**Analysis Query 4:**

|  |  |
| --- | --- |
| **Objective** | **To get the most mentioned/discussed top 15 stocks on twitter.**  Below query lists most discussed stocks/ticker symbols on twitter. |
| **Query** | SELECT text, COUNT(\*) c FROM stocktweets LATERAL VIEW explode(split(text, ' ')) stocktweets as text where text like '%$%' GROUP BY text order by c desc limit 15; |
| **Result** | $AAPL (Apple Computers) is the most mentioned stock on twitter. Followed by $HYG (High Yid Corp Bond ETF), $JPM (JP Morgan Chase) and $FB (Facebook) |
| **Execution**  **Screenshot** |  |