Exercise #1:

ANALYZING SOCIAL MEDIA AND CUSTOMER SENTIMENT WITH APACHE NIFI AND HDP SEARCH

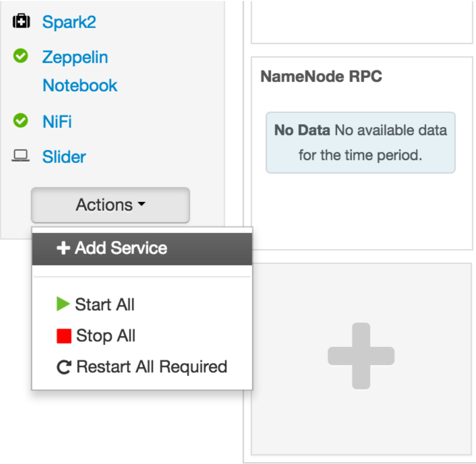
INTRODUCTION

We will use Solr and the LucidWorks HDP Search to view our streamed data in real time to gather insights as the data arrives in our Hadoop cluster. Next, we will use Hive to analyze the social sentiment after we have finished collecting our data from NiFi.

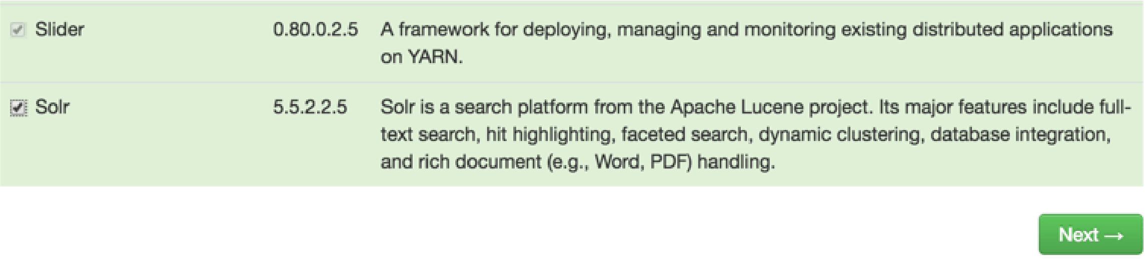
Finally, we will use Apache Zeppelin to create charts, so we can visualize our data directly inside of our Hadoop cluster.

CONFIGURE AND START SOLR

Make sure that Ambari Infra is stopped, we now need to install HDP Search. Login to Ambari user credentials: Username – **raj_ops** and Password – **raj_ops**. Click on Actions button at the bottom and then Add Service:

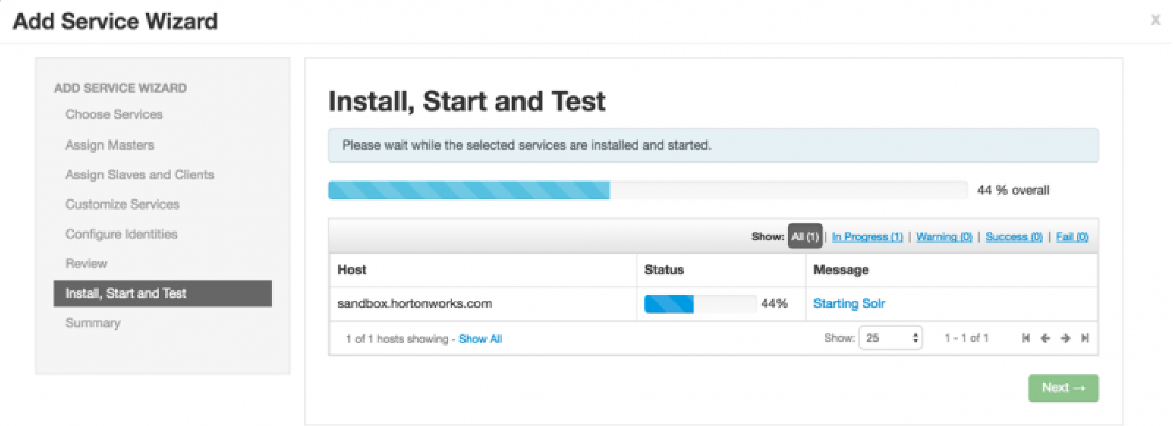


Next, you will view a list of services that you can add. Scroll to the bottom and select `Solr`, then press `Next`.



☑ Slider	0.80.0.2.5	A framework for deploying, managing and monitoring existing distributed applications on YARN.
☑ Solr	5.5.2.2.5	Solr is a search platform from the Apache Lucene project. Its major features include full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling.

Accept all default values in next few pages, and then you can see the progress of your installation:



Install, Start and Test

Please wait while the selected services are installed and started.

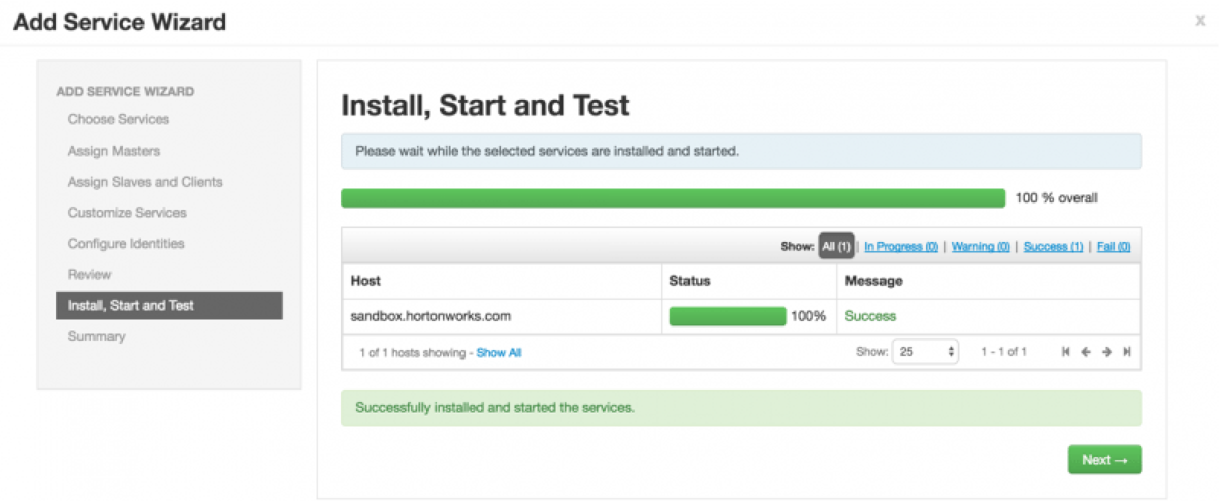
44 % overall

Show: All (1) | In Progress (1) | Warning (0) | Success (0) | Fail (0)

Host	Status	Message
sandbox.hortonworks.com	44%	Starting Solr

1 of 1 hosts showing - Show All   Show: 25   1 - 1 of 1

After a minute, you can see Solr successfully installed:



Install, Start and Test

Please wait while the selected services are installed and started.

100 % overall

Show: All (1) | In Progress (0) | Warning (0) | Success (1) | Fail (0)

Host	Status	Message
sandbox.hortonworks.com	100%	Success

1 of 1 hosts showing - Show All   Show: 25   1 - 1 of 1

Successfully installed and started the services.

Press Next, you will be asked to restart some services. Restart HDFS, YARN, Mapreduce2 and HBase.

We just need to make a few quick changes.

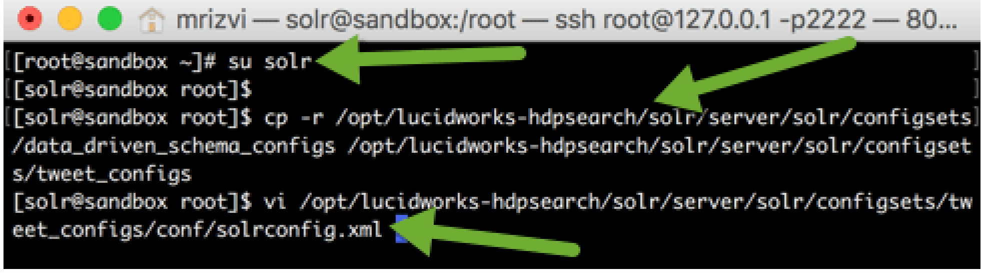
Open your terminal shell and SSH back into the sandbox. We're going to need to run the following commands as the **solr** user. Run

```
su solr
```

Then we need to edit the following file path to make sure that Solr can recognize a tweet's timestamp format. First we're going to copy the config set over to twitter's **tweet_configs** folder:

```
cp -r /opt/lucidworks-
hdpsearch/solr/server/solr/configsets/data_driven_schema_configs
/opt/lucidworks-hdpsearch/solr/server/solr/configsets/tweet_configs

vi /opt/lucidworks-
hdpsearch/solr/server/solr/configsets/tweet_configs/conf/solrconfig.xml
```


mrizvi — solr@sandbox:/root — ssh root@127.0.0.1 -p2222 — 80...

```
[root@sandbox ~]# su solr
[solr@sandbox root]$
[solr@sandbox root]$ cp -r /opt/lucidworks-hdpsearch/solr/server/solr/configsets
/data_driven_schema_configs /opt/lucidworks-hdpsearch/solr/server/solr/configset
s/tweet_configs
[solr@sandbox root]$ vi /opt/lucidworks-hdpsearch/solr/server/solr/configsets/tw
eet_configs/conf/solrconfig.xml
```

Once the file is opened in `vi` type

Note In **vi** the command below should not be run in **INSERT** mode. `/` will do a find for the text that you type after it.

```
/solr.ParseDateFieldUpdateProcessorFactory
```

This will bring you to the part of the config where we need to add the following:

```
<str>EEE MMM d HH:mm:ss Z yyyy</str>
```

Make sure this is inserted just above all of the other `<str>` tags.

Note In `vi`, to type or insert anything you must be in *insert mode*. Press `i` on your keyboard to enter insert mode in `vi`.

After inserting the above, the portion of the file should look something like this:

```
<processor class="solr.ParseLongFieldUpdateProcessorFactory"/>
  <processor class="solr.ParseDateFieldUpdateProcessorFactory">
    <arr name="format">
      <str>EEE MMM d HH:mm:ss Z yyyy</str>
      <str>yyyy-MM-dd'T'HH:mm:ss.SSSZ</str>
      <str>yyyy-MM-dd'T'HH:mm:ss,SSSZ</str>
      <str>yyyy-MM-dd'T'HH:mm:ss.SSS</str>
      <str>yyyy-MM-dd'T'HH:mm:ss,SSS</str>
      <str>yyyy-MM-dd'T'HH:mm:ssZ</str>
      </arr>
    </processor>
</processor>
```

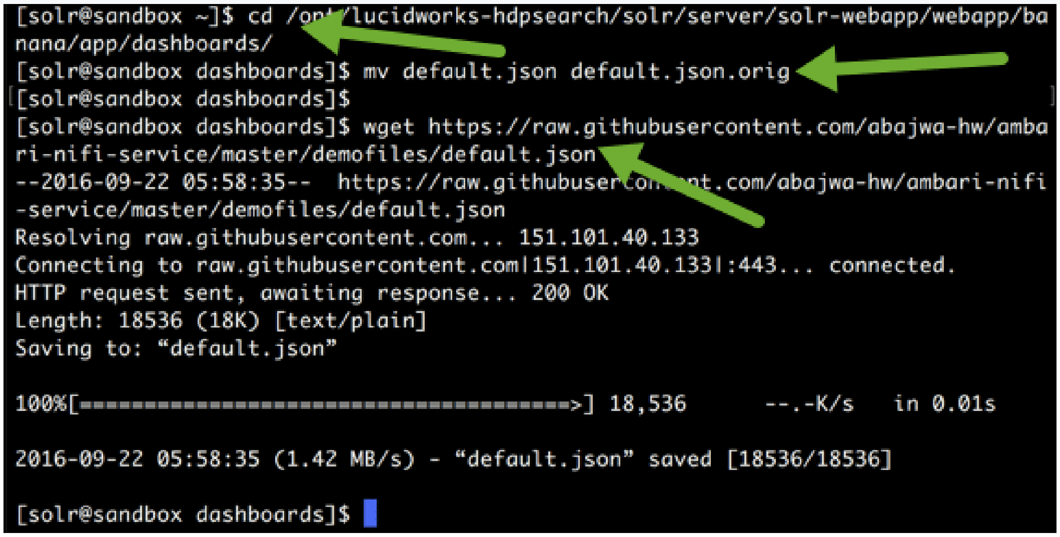
Finally press the **Escape key** on your keyboard and type `:wq` to save and close the `solrconfig.xml` file.

Next we need to replace a JSON file. Use the following commands to move the original and download the replacement file:

```
cd /opt/lucidworks-hdpsearch/solr/server/solr-
webapp/webapp/banana/app/dashboards/

mv default.json default.json.orig

wget https://raw.githubusercontent.com/hortonworks/data-
tutorials/master/tutorials/hdp/analyzing-social-media-and-customer-sentiment-
with-apache-nifi-and-hdp-search/assets/default.json
```



```
[solr@sandbox ~]$ cd /opt/lucidworks-hdpsearch/solr/server/solr-webapp/webapp/ba
nana/app/dashboards/
[solr@sandbox dashboards]$ mv default.json default.json.orig
[solr@sandbox dashboards]$
[solr@sandbox dashboards]$ wget https://raw.githubusercontent.com/abajwa-hw/amba
ri-nifi-service/master/demofiles/default.json
--2016-09-22 05:58:35--  https://raw.githubusercontent.com/abajwa-hw/ambari-nifi
-service/master/demofiles/default.json
Resolving raw.githubusercontent.com... 151.101.40.133
Connecting to raw.githubusercontent.com|151.101.40.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 18536 (18K) [text/plain]
Saving to: “default.json”

100%[======================================>] 18,536      --.-K/s   in 0.01s

2016-09-22 05:58:35 (1.42 MB/s) - “default.json” saved [18536/18536]

[solr@sandbox dashboards]$
```

Then we are going to add a collection called “tweets”

```
/opt/lucidworks-hdpsearch/solr/bin/solr create -c tweets -d tweet_configs -s
1 -rf 1 -p 8983
```

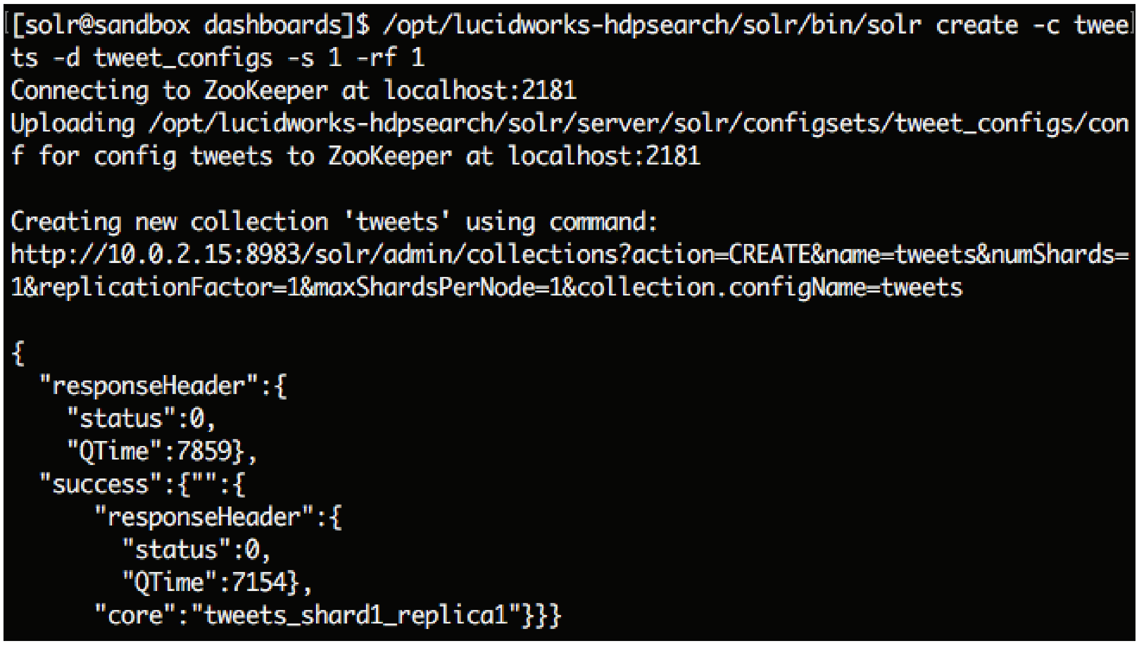
Note: Here -c indicates the name

-d is the config directory

-s is the number of shards

-rf is the replication factor

-p is the port at which Solr is running



```
[solr@sandbox dashboards]$ /opt/lucidworks-hdpsearch/solr/bin/solr create -c twee
ts -d tweet_configs -s 1 -rf 1
Connecting to ZooKeeper at localhost:2181
Uploading /opt/lucidworks-hdpsearch/solr/server/solr/configsets/tweet_configs/con
f for config tweets to ZooKeeper at localhost:2181

Creating new collection 'tweets' using command:
http://10.0.2.15:8983/solr/admin/collections?action=CREATE&name=tweets&numShards=
1&replicationFactor=1&maxShardsPerNode=1&collection.configName=tweets

{
  "responseHeader":{
    "status":0,
    "QTime":7859},
  "success":{"":{
      "responseHeader":{
        "status":0,
        "QTime":7154},
      "core":"tweets_shard1_replica1"}}}
```

We can now go back to running commands as the **root** user. Run

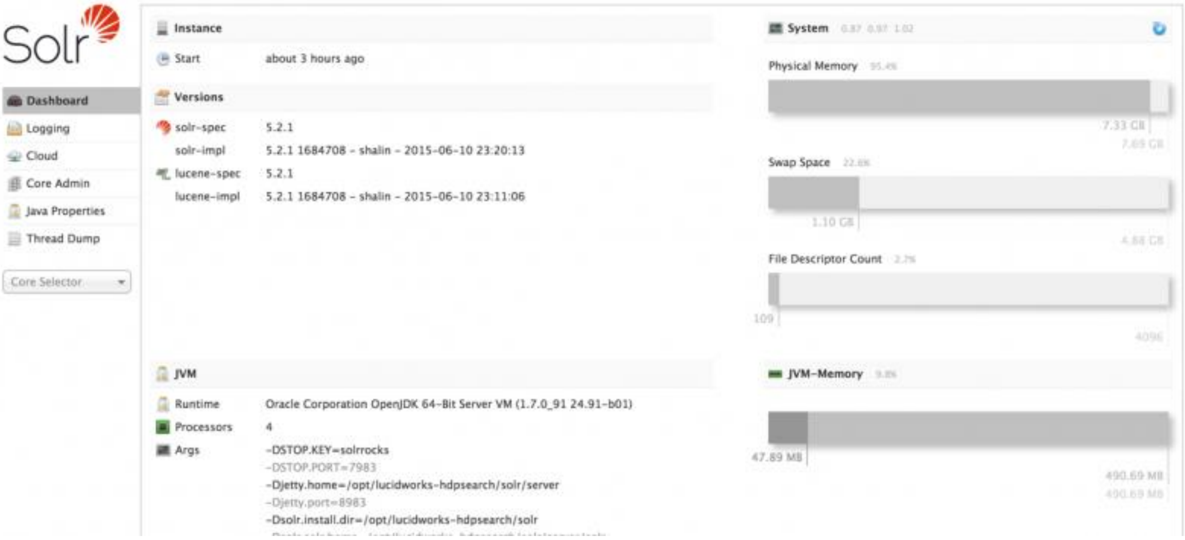
```
exit
```

This will log you out of the `solr` user

Great! Now Solr should be installed and running on your sandbox!

Ensure that you can access the Solr UI by navigating

to http://sandbox.hortonworks.com:8983/solr/



Instance
Start about 3 hours ago

Versions
solr-spec 5.2.1
solr-impl 5.2.1 1684708 - shalin - 2015-06-10 23:20:13
lucene-spec 5.2.1
lucene-impl 5.2.1 1684708 - shalin - 2015-06-10 23:11:06

JVM
Runtime Oracle Corporation OpenJDK 64-Bit Server VM (1.7.0_91 24.91-b01)
Processors 4
Args -DSTOP.KEY=solrrocks
-DSTOP.PORT=7983
-Djetty.home=/opt/lucidworks-hdpsearch/solr/server
-Djetty.port=8983
-Dsolr.install.dir=/opt/lucidworks-hdpsearch/solr

GENERATING RANDOM TWEET DATA FOR HIVE AND SOLR

First you'll need to SSH into the sandbox execute the following command

```
wget https://raw.githubusercontent.com/hortonworks/data-tutorials/master/tutorials/hdp/analyzing-social-media-and-customer-sentiment-with-apache-nifi-and-hdp-search/assets/twitter-gen.sh
```

Then run the command with your specified number of tweets that you would like to generate.

```
bash twitter-gen.sh {NUMBER_OF_TWEETS}
```

Example:

```
bash twitter-gen.sh 2000
```

The script will generate the data and put it in the directory `/tmp/data/`

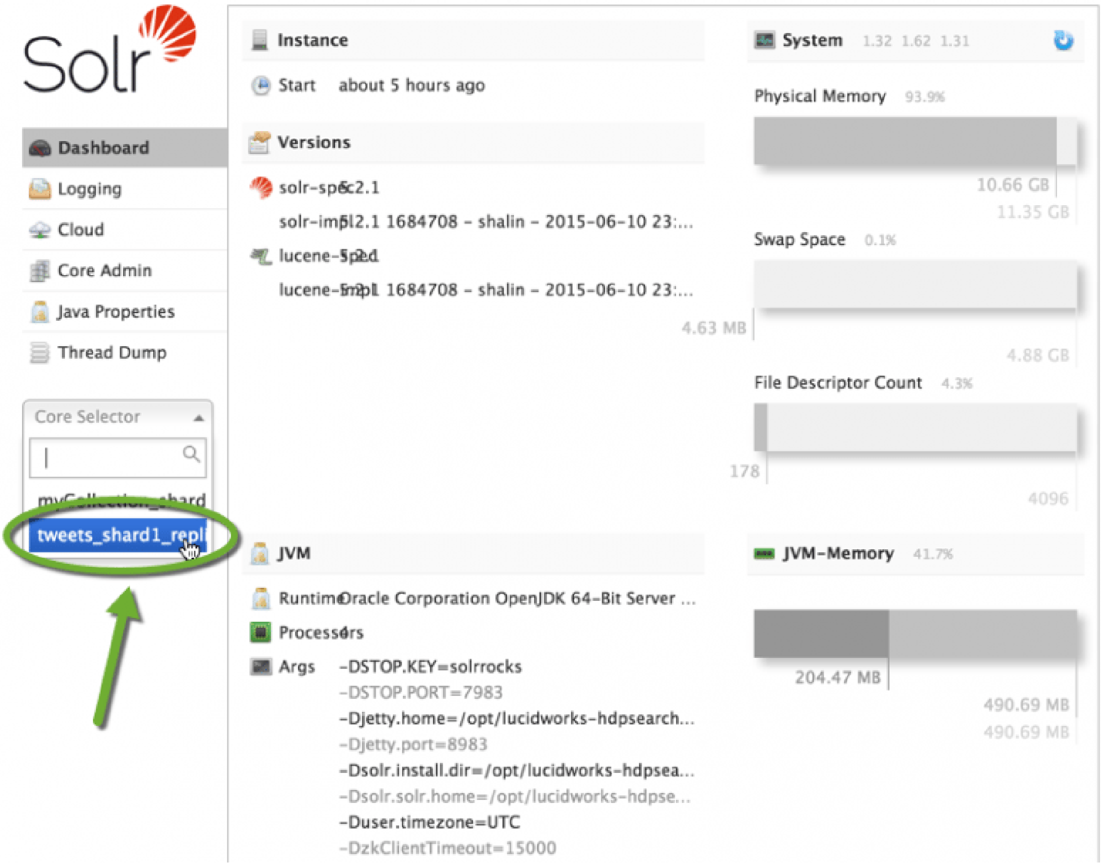
You can now continue with the rest of the tutorial.

ANALYZE AND SEARCH DATA WITH SOLR

Now that we have our data in HDP-Search/Solr we can go ahead and start searching through our data.

Let's go do some custom search on the data! Head back to the normal Solr dashboard at http://sandbox.hortonworks.com:8983/solr

Select the **tweets shard** that we created before from the `Core Selector` menu on the bottom left of the screen.

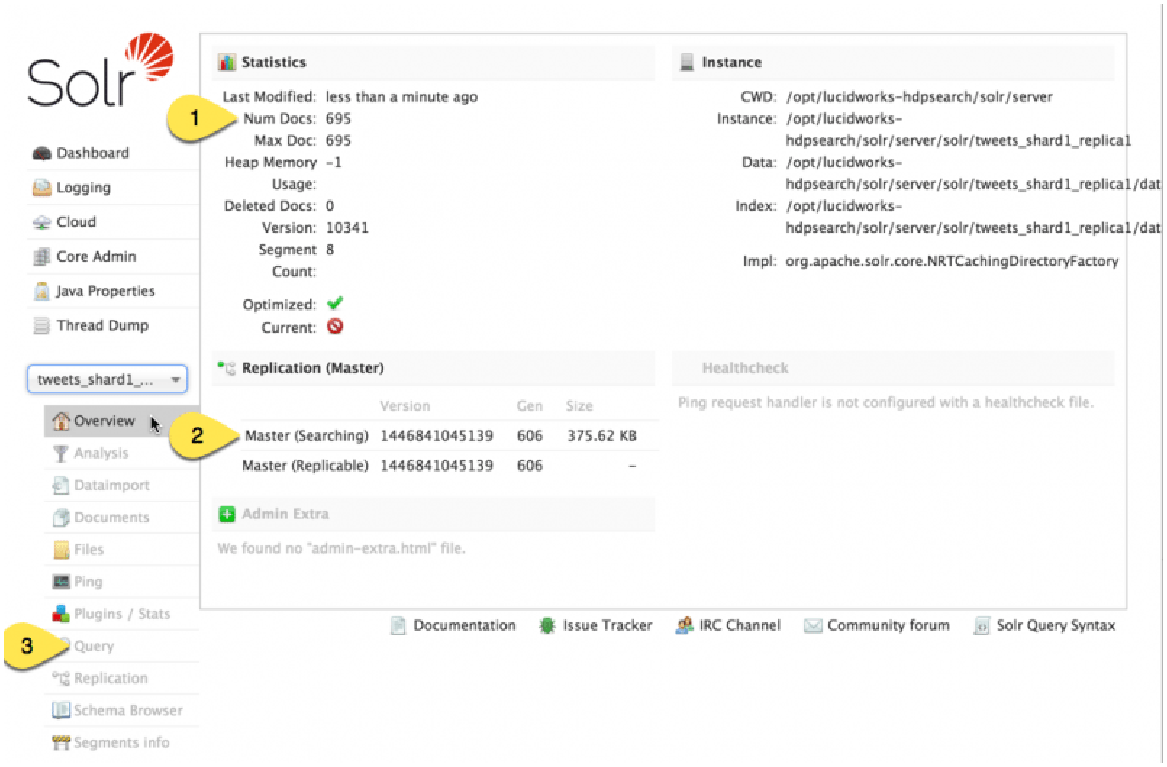


Start about 5 hours ago

solr-spec 5.2.1
solr-impl 5.2.1 1684708 - shalin - 2015-06-10 23:...
lucene-spec 5.2.1
lucene-impl 5.2.1 1684708 - shalin - 2015-06-10 23:...

Runtime Oracle Corporation OpenJDK 64-Bit Server ...
Processors 4
Args -DSTOP.KEY=solrrocks
-DSTOP.PORT=7983
-Djetty.home=/opt/lucidworks-hdpsearch...
-Djetty.port=8983
-Dsolr.install.dir=/opt/lucidworks-hdpsea...
-Dsolr.solr.home=/opt/lucidworks-hdpse...
-Duser.timezone=UTC
-DzkClientTimeout=15000

Once you’ve selected the tweets shard we can look to see what Solr has done with our data.



Statistics

Last Modified: less than a minute ago
Num Docs: 695
Max Doc: 695
Heap Memory Usage: -1
Deleted Docs: 0
Version: 10341
Segment Count: 8
Optimized: ✓
Current: ⊘

Instance

CWD: /opt/lucidworks-hdpsearch/solr/server
Instance: /opt/lucidworks-hdpsearch/solr/server/solr/tweets_shard1_replica1
Data: /opt/lucidworks-hdpsearch/solr/server/solr/tweets_shard1_replica1/dat
Index: /opt/lucidworks-hdpsearch/solr/server/solr/tweets_shard1_replica1/dat
Impl: org.apache.solr.core.NRTCachingDirectoryFactory

Replication (Master)

	Version	Gen	Size
Master (Searching)	1446841045139	606	375.62 KB
Master (Replicable)	1446841045139	606	-

Healthcheck

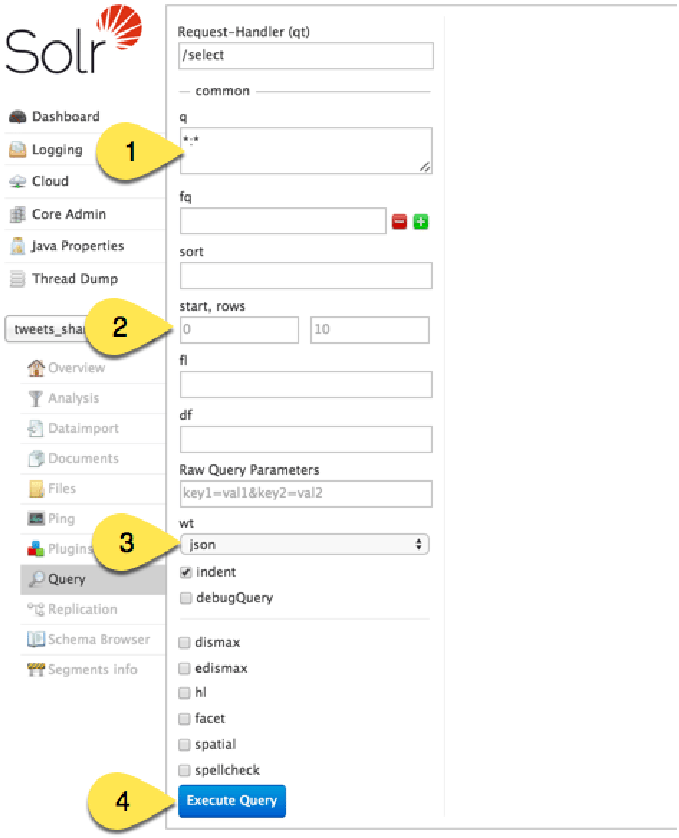
Ping request handler is not configured with a healthcheck file.

Admin Extra

We found no "admin-extra.html" file.

1. If you used the `twitter-gen.sh` script then this number should be close to the amount of tweets that you generated.
2. Here we can see the size on the disk that the data is taking up in Solr. We don’t have many tweets collected yet, so this number is quite small.
3. On the left side bar there are a number of different tabs to view the data that’s stored within Solr. We’re going to focus on the **Query** one, but you should explore the others as well.

   Click on the query tab, and you should be brought to screen like the following:



Request-Handler (qt)
/select

common

q
:

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☑ indent
□ debugQuery

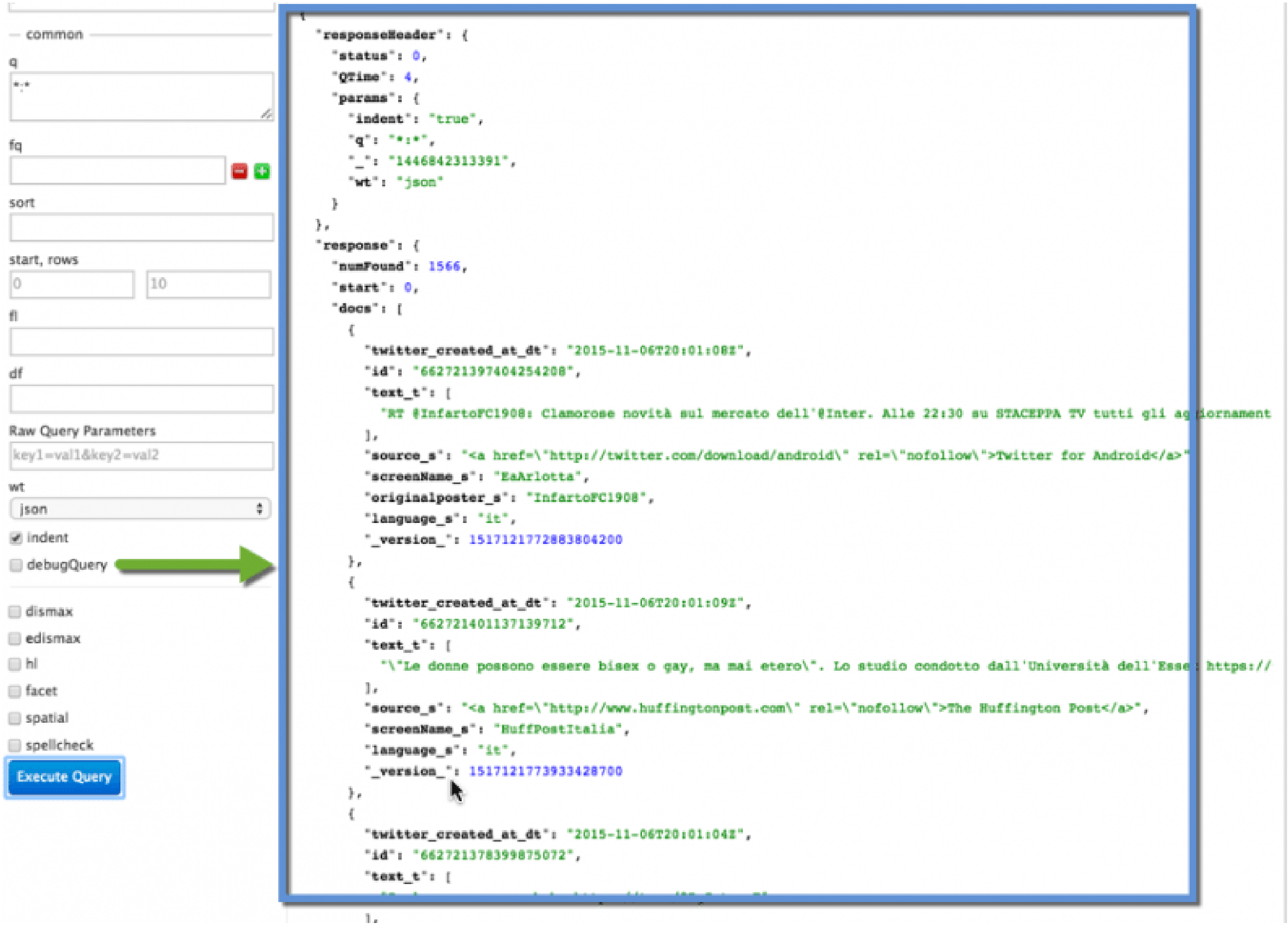
□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

We’re only going to be using 3 of these fields before we execute any queries, but let’s quickly outline the different query parameters

- **fq**: This is a filter query parameter it lets us retrieve data that only contains certain values that we’re looking for. Example: we can specify that we only want tweets after a certain time to be returned.
- **sort**: self-explanatory. You can sort by a specified field in ascending or descending order. we could return all tweets by alphabetical order of Twitter handles, or possible by the time they were tweeted as well.
- **start, rows**: This tells us where exactly in the index we should start searching, and how many rows should be returned when we execute the query. The defaults for each of these is `0` and `10` respectively.
- **fl**: Short for *field list* specify which fields you want to be returned. If the data many, many fields, you can choose to specify only a few that are returned in the query.
- **df**: Short for *default fields* you can tell which fields solr should be searching in. You will not need this if the query fields are already defined.
- **Raw Query Params**: These will be added directly the the url that is requested when Solr send the request with all of the query information.

- **wt**: This is the type of data that solr will return. We can specify many things such as JSON, XML, or CSV formatting.

  We are not going to worry about the rest of the flags. Without entering any parameters click **Execute Query**.



q: *:*

fq:

sort:

start, rows: 0 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent
☐ debugQuery

☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

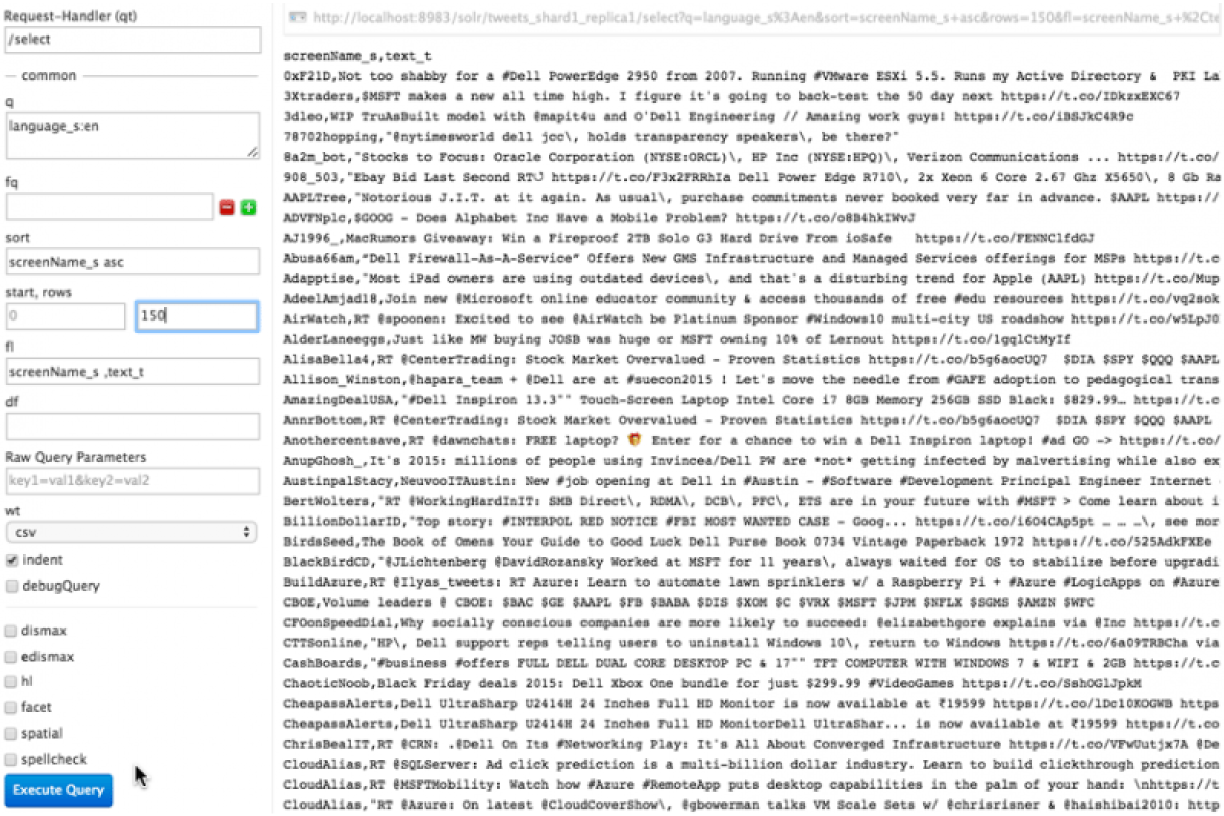
```
{
  "responseHeader": {
    "status": 0,
    "QTime": 4,
    "params": {
      "indent": "true",
      "q": "*:*",
      "_": "1446842313391",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 1566,
    "start": 0,
    "docs": [
      {
        "twitter_created_at_dt": "2015-11-06T20:01:08Z",
        "id": "662721397404254208",
        "text_t": [
          "RT @InfartoFC1908: Clamorose novità sul mercato dell'@Inter. Alle 22:30 su STACEPPA TV tutti gli aggiornament
        ],
        "source_s": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>"
        "screenName_s": "EaArlotta",
        "originalposter_s": "InfartoFC1908",
        "language_s": "it",
        "_version_": 1517121772883804200
      },
      {
        "twitter_created_at_dt": "2015-11-06T20:01:09Z",
        "id": "662721401137139712",
        "text_t": [
          "\"Le donne possono essere bisex o gay, ma mai etero\". Lo studio condotto dall'Università dell'Esse: https://
        ],
        "source_s": "<a href=\"http://www.huffingtonpost.com\" rel=\"nofollow\">The Huffington Post</a>",
        "screenName_s": "HuffPostItalia",
        "language_s": "it",
        "_version_": 1517121773933428700
      },
      {
        "twitter_created_at_dt": "2015-11-06T20:01:04Z",
        "id": "662721378399875072",
        "text_t": [

        ],
```

From this you should be able to view all the tweet data that is collected. Try playing with some of the parameters and add more to the **rows** value in the query to see how many results you can obtain.

Now let's do a real query and see if we can find some valuable data.

- For **q** type `language_s:en`
- For **sort** type `screenName_s asc`
- For **rows** type `150`
- For **fl** type `screenName_s, text_t`
- For **wt** choose `csv`



Request-Handler (qt)
/select

common

q
language_s:en

fq

sort
screenName_s asc

start, rows
0 150

fl
screenName_s ,text_t

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☑ indent
□ debugQuery

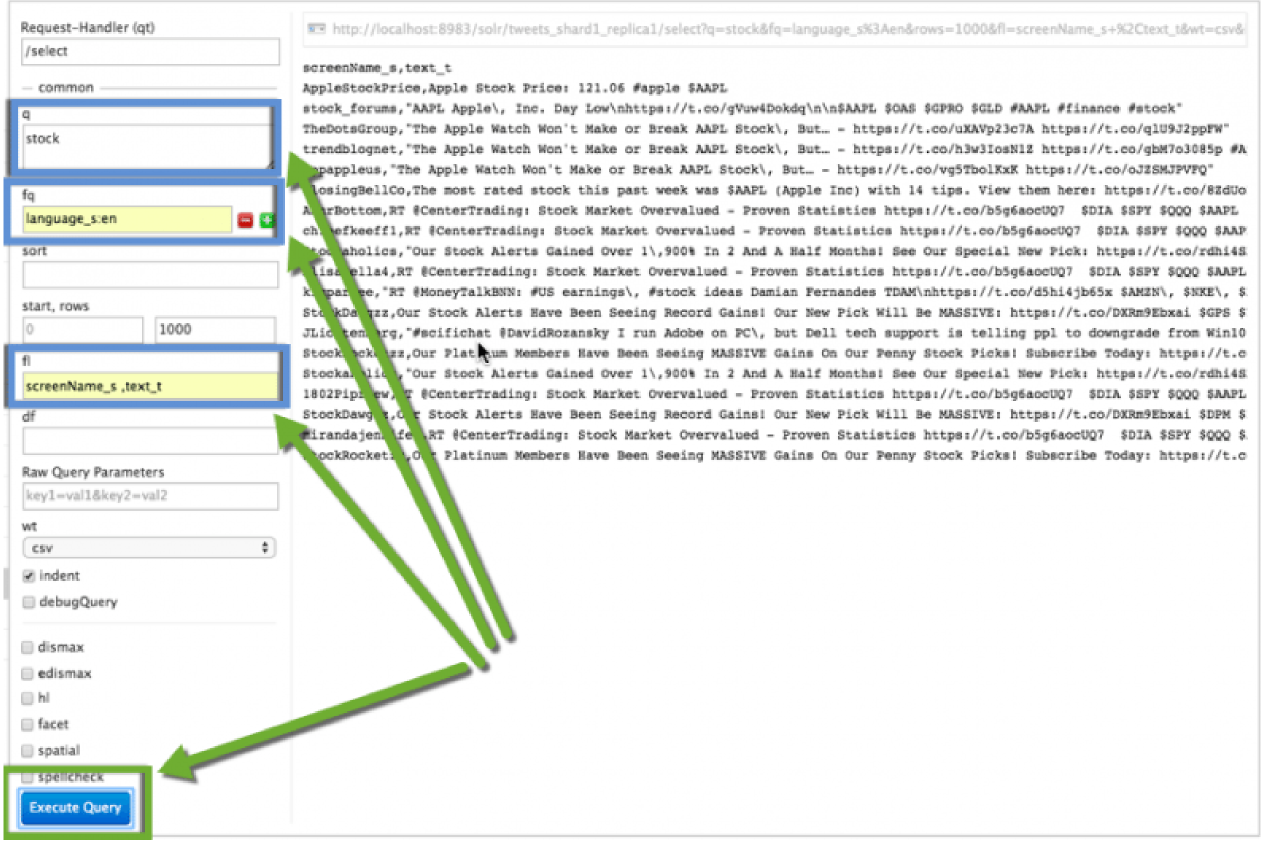
□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://localhost:8983/solr/tweets_shard1_replica1/select?q=language_s%3Aen&sort=screenName_s+asc&rows=150&fl=screenName_s+%2Cte

```
screenName_s,text_t
0xF21D,Not too shabby for a #Dell PowerEdge 2950 from 2007. Running #VMware ESXi 5.5. Runs my Active Directory &  PKI La
3Xtraders,$MSFT makes a new all time high. I figure it's going to back-test the 50 day next https://t.co/IDkzxEXC67
3dleo,WIP TruAsBuilt model with @mapit4u and O'Dell Engineering // Amazing work guys! https://t.co/iBSJkC4R9c
78702hopping,"@nytimesworld dell jcc\, holds transparency speakers\, be there?"
8a2m_bot,"Stocks to Focus: Oracle Corporation (NYSE:ORCL)\, HP Inc (NYSE:HPQ)\, Verizon Communications ... https://t.co/
908_503,"Ebay Bid Last Second RT↺ https://t.co/F3x2FRRhIa Dell Power Edge R710\, 2x Xeon 6 Core 2.67 Ghz X5650\, 8 Gb Ra
AAPLTree,"Notorious J.I.T. at it again. As usual\, purchase commitments never booked very far in advance. $AAPL https://
ADVFNplc,$GOOG - Does Alphabet Inc Have a Mobile Problem? https://t.co/o8B4hkIWvJ
AJ1996_,MacRumors Giveaway: Win a Fireproof 2TB Solo G3 Hard Drive From ioSafe   https://t.co/FENNClfdGJ
Abusa66am,"Dell Firewall-As-A-Service" Offers New GMS Infrastructure and Managed Services offerings for MSPs https://t.c
Adapptise,"Most iPad owners are using outdated devices\, and that's a disturbing trend for Apple (AAPL) https://t.co/Mup
AdeelAmjad18,Join new @Microsoft online educator community & access thousands of free #edu resources https://t.co/vq2sok
AirWatch,RT @spoonen: Excited to see @AirWatch be Platinum Sponsor #Windows10 multi-city US roadshow https://t.co/w5LpJ0
AlderLaneeggs,Just like MW buying JOSB was huge or MSFT owning 10% of Lernout https://t.co/lgqlCtMyIf
AlisaBella4,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAPL
Allison_Winston,@hapara_team + @Dell are at #suecon2015 ! Let's move the needle from #GAFE adoption to pedagogical trans
AmazingDealUSA,"#Dell Inspiron 13.3"" Touch-Screen Laptop Intel Core i7 8GB Memory 256GB SSD Black: $829.99… https://t.c
AnnrBottom,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAPL
Anothercentsave,RT @dawnchats: FREE laptop? 🎁 Enter for a chance to win a Dell Inspiron laptop! #ad GO -> https://t.co/
AnupGhosh_,It's 2015: millions of people using Invincea/Dell PW are *not* getting infected by malvertising while also ex
AustinpalStacy,NeuvooITAustin: New #job opening at Dell in #Austin - #Software #Development Principal Engineer Internet
BertWolters,"RT @WorkingHardInIT: SMB Direct\, RDMA\, DCB\, PFC\, ETS are in your future with #MSFT > Come learn about i
BillionDollarID,"Top story: #INTERPOL RED NOTICE #FBI MOST WANTED CASE - Goog... https://t.co/i6O4CAp5pt … … …\, see mor
BirdsSeed,The Book of Omens Your Guide to Good Luck Dell Purse Book 0734 Vintage Paperback 1972 https://t.co/525AdkFXEe
BlackBirdCD,"@JLichtenberg @DavidRozansky Worked at MSFT for 11 years\, always waited for OS to stabilize before upgradi
BuildAzure,RT @Ilyas_tweets: RT Azure: Learn to automate lawn sprinklers w/ a Raspberry Pi + #Azure #LogicApps on #Azure
CBOE,Volume leaders @ CBOE: $BAC $GE $AAPL $FB $BABA $DIS $XOM $C $VRX $MSFT $JPM $NFLX $SGMS $AMZN $WFC
CFOonSpeedDial,Why socially conscious companies are more likely to succeed: @elizabethgore explains via @Inc https://t.c
CTTSonline,"HP\, Dell support reps telling users to uninstall Windows 10\, return to Windows https://t.co/6a09TRBCha via
CashBoards,"#business #offers FULL DELL DUAL CORE DESKTOP PC & 17"" TFT COMPUTER WITH WINDOWS 7 & WIFI & 2GB https://t.c
ChaoticNoob,Black Friday deals 2015: Dell Xbox One bundle for just $299.99 #VideoGames https://t.co/Ssh0GlJpkM
CheapassAlerts,Dell UltraSharp U2414H 24 Inches Full HD Monitor is now available at ₹19599 https://t.co/1Dc10KOGWB https
CheapassAlerts,Dell UltraSharp U2414H 24 Inches Full HD MonitorDell UltraShar... is now available at ₹19599 https://t.co
ChrisBealIT,RT @CRN: .@Dell On Its #Networking Play: It's All About Converged Infrastructure https://t.co/VFwUutjx7A @De
CloudAlias,RT @SQLServer: Ad click prediction is a multi-billion dollar industry. Learn to build clickthrough prediction
CloudAlias,RT @MSFTMobility: Watch how #Azure #RemoteApp puts desktop capabilities in the palm of your hand: \nhttps://t
CloudAlias,"RT @Azure: On latest @CloudCoverShow\, @gbowerman talks VM Scale Sets w/ @chrisrisner & @haishibai2010: http
```

Let's try one last query. This time you can omit the **sort** field and chooses whichever **wt** format you like. Keep the **fl** parameter as is though.

- Specify an **fq** parameter as `language_s:en`
- In the query box, pick any keyword. I am going to use `stock`



Request-Handler (qt)
/select

common

q
stock

fq
language_s:en

sort

start, rows
0 1000

fl
screenName_s ,text_t

df

Raw Query Parameters
key1=val1&key2=val2

wt
csv

☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://localhost:8983/solr/tweets_shard1_replica1/select?q=stock&fq=language_s%3Aen&rows=1000&fl=screenName_s+%2Ctext_t&wt=csv&

```
screenName_s,text_t
AppleStockPrice,Apple Stock Price: 121.06 #apple $AAPL
stock_forums,"AAPL Apple\, Inc. Day Low\nhttps://t.co/gVuw4Dokdq\n\n$AAPL $OAS $GPRO $GLD #AAPL #finance #stock"
TheDotsGroup,"The Apple Watch Won't Make or Break AAPL Stock\, But… - https://t.co/uXAVp23c7A https://t.co/q1U9J2ppFW"
trendblognet,"The Apple Watch Won't Make or Break AAPL Stock\, But… - https://t.co/h3w3IosN1Z https://t.co/gbM7o3085p #A
 pappleus,"The Apple Watch Won't Make or Break AAPL Stock\, But… - https://t.co/vg5TbolKxK https://t.co/oJZSMJPVFQ"
ClosingBellCo,The most rated stock this past week was $AAPL (Apple Inc) with 14 tips. View them here: https://t.co/8ZdUo
A  rBottom,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAPL
ch  efkeeff1,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAP
Sto  aholics,"Our Stock Alerts Gained Over 1\,900% In 2 And A Half Months! See Our Special New Pick: https://t.co/rdhi4S
 isa ella4,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAPL
k  par  ee,"RT @MoneyTalkBNN: #US earnings\, #stock ideas Damian Fernandes TDAM\nhttps://t.co/d5hi4jb65x $AMZN\, $NKE\, $
StockDawgzz,Our Stock Alerts Have Been Seeing Record Gains! Our New Pick Will Be MASSIVE: https://t.co/DXRm9Ebxai $GPS $
JLic  ten  rg,"#scifichat @DavidRozansky I run Adobe on PC\, but Dell tech support is telling ppl to downgrade from Win10
Stock  ck  zz,Our Platinum Members Have Been Seeing MASSIVE Gains On Our Penny Stock Picks! Subscribe Today: https://t.c
Stocka  lic ,"Our Stock Alerts Gained Over 1\,900% In 2 And A Half Months! See Our Special New Pick: https://t.co/rdhi4S
1802Pip  ew, T @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $AAPL
StockDawg  ,O r Stock Alerts Have Been Seeing Record Gains! Our New Pick Will Be MASSIVE: https://t.co/DXRm9Ebxai $DPM $
Mirandajen  fe ,RT @CenterTrading: Stock Market Overvalued - Proven Statistics https://t.co/b5g6aocUQ7  $DIA $SPY $QQQ $
 ockRocket  ,O r Platinum Members Have Been Seeing MASSIVE Gains On Our Penny Stock Picks! Subscribe Today: https://t.c
```

ANALYZE TWEET DATA IN HIVE

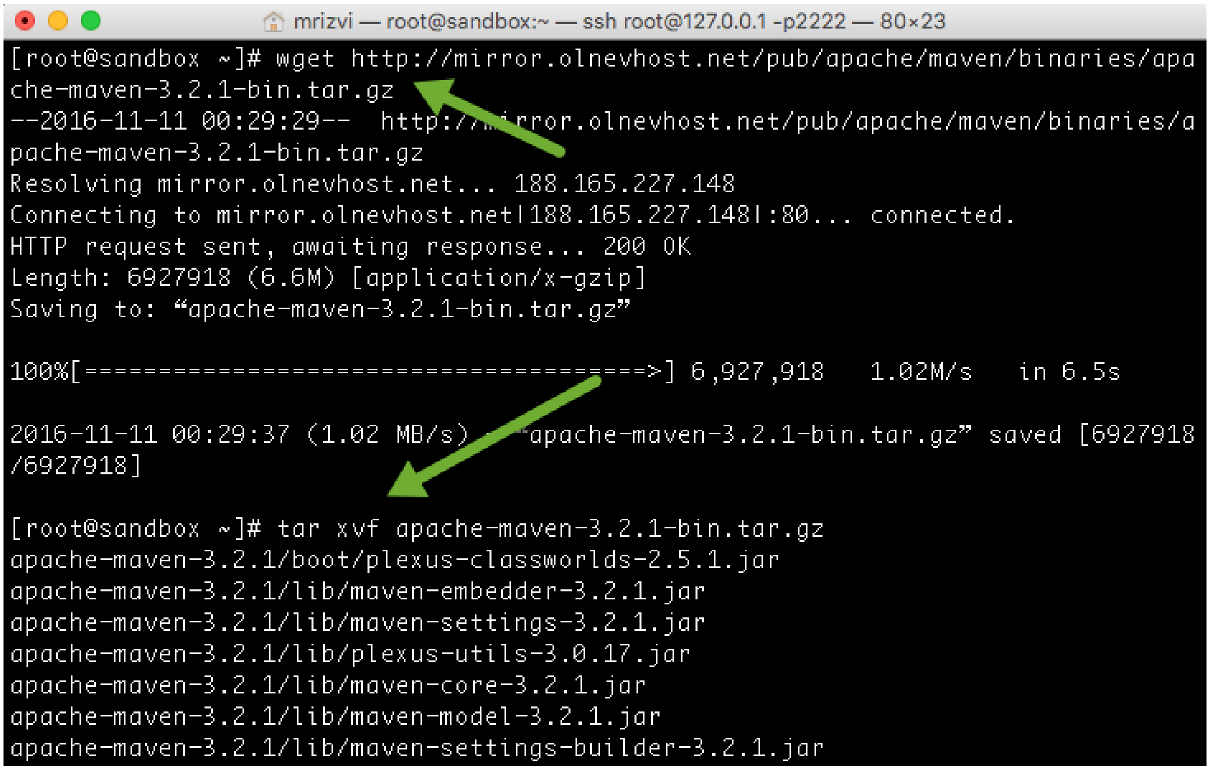
Now that we've looked at some of our data and searched it with Solr, let's see if we can refine it a bit more.

But before moving ahead, let us setup **Hive-JSON-Serde** to read the data in **JSON** format. We must use the maven to compile the serde. Go back to the terminal and follow the below steps to setup the maven:

```
wget http://mirror.olnevhost.net/pub/apache/maven/binaries/apache-maven-3.2.1-bin.tar.gz
```

Now, extract this file:

```
tar xvf apache-maven-3.2.1-bin.tar.gz
```



```
[root@sandbox ~]# wget http://mirror.olnevhost.net/pub/apache/maven/binaries/apa
che-maven-3.2.1-bin.tar.gz
--2016-11-11 00:29:29--  http://mirror.olnevhost.net/pub/apache/maven/binaries/a
pache-maven-3.2.1-bin.tar.gz
Resolving mirror.olnevhost.net... 188.165.227.148
Connecting to mirror.olnevhost.net|188.165.227.148|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 6927918 (6.6M) [application/x-gzip]
Saving to: “apache-maven-3.2.1-bin.tar.gz”

100%[======================================>] 6,927,918   1.02M/s   in 6.5s

2016-11-11 00:29:37 (1.02 MB/s) - “apache-maven-3.2.1-bin.tar.gz” saved [6927918
/6927918]

[root@sandbox ~]# tar xvf apache-maven-3.2.1-bin.tar.gz
apache-maven-3.2.1/boot/plexus-classworlds-2.5.1.jar
apache-maven-3.2.1/lib/maven-embedder-3.2.1.jar
apache-maven-3.2.1/lib/maven-settings-3.2.1.jar
apache-maven-3.2.1/lib/plexus-utils-3.0.17.jar
apache-maven-3.2.1/lib/maven-core-3.2.1.jar
apache-maven-3.2.1/lib/maven-model-3.2.1.jar
apache-maven-3.2.1/lib/maven-settings-builder-3.2.1.jar
```

Now since our maven is installed, let us download the Hive-JSON-Serde. Type the following command:

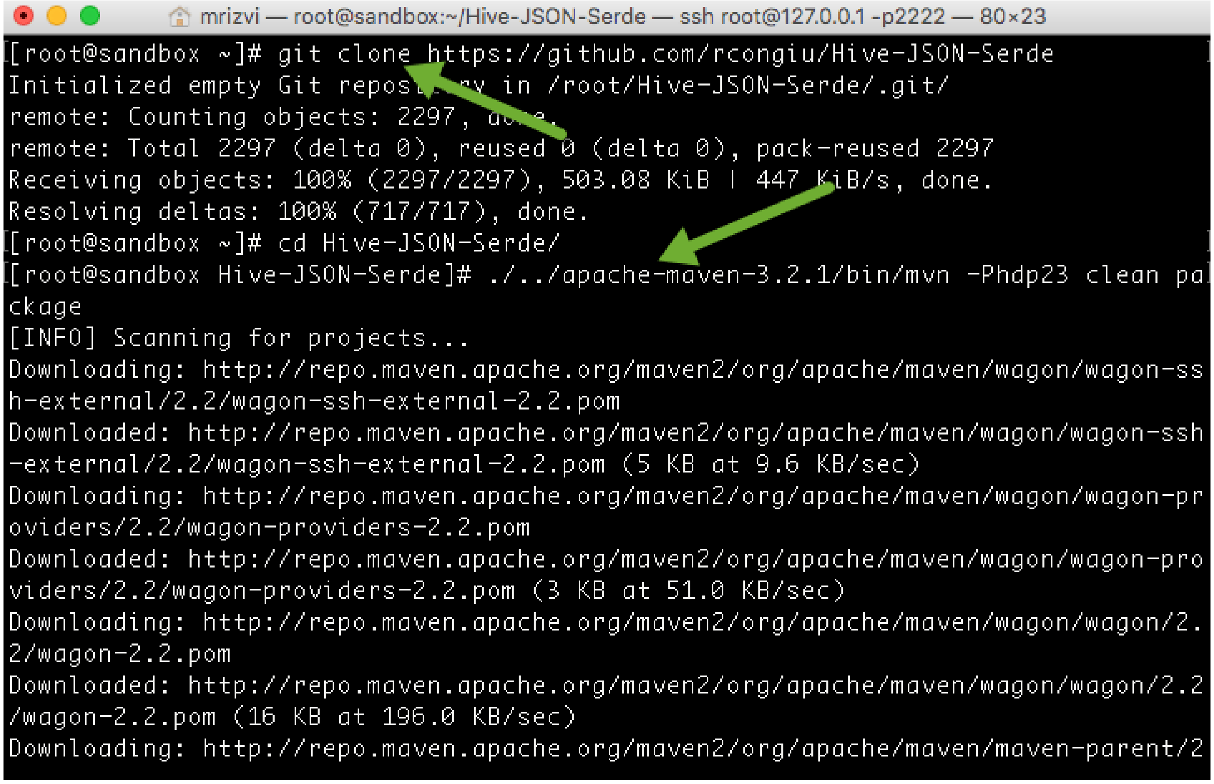
```
git clone https://github.com/rcongiu/Hive-JSON-Serde
```

This command must have created the new directory, go inside to that directory using cd:

```
cd Hive-JSON-Serde
```

Next, run the command to compile the serde:

```
./../apache-maven-3.2.1/bin/mvn -Phdp23 clean package
```



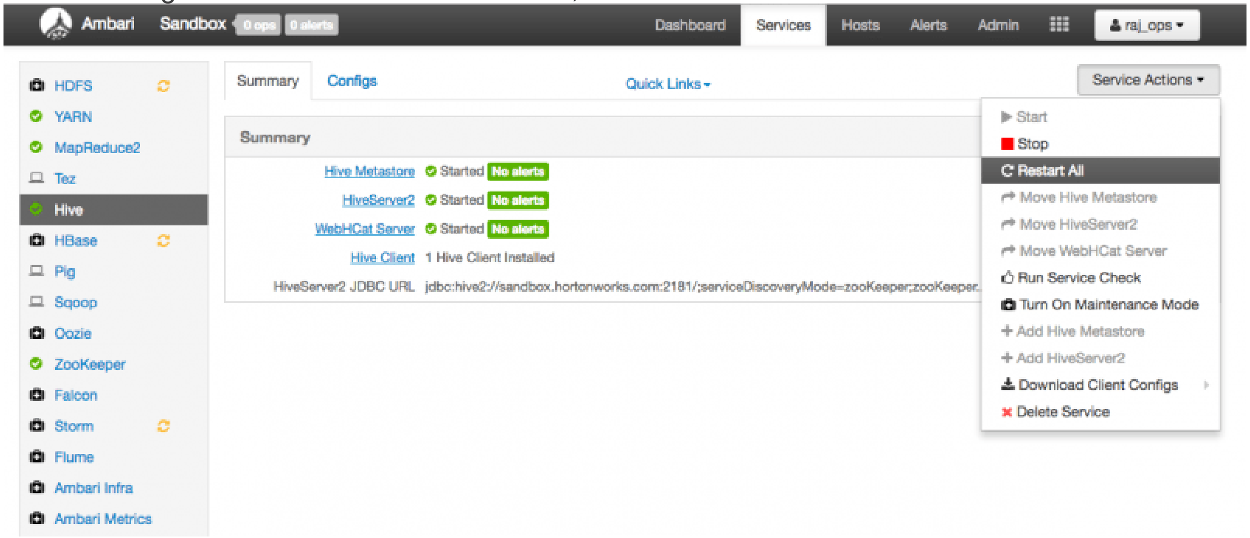
mrizvi — root@sandbox:~/Hive-JSON-Serde — ssh root@127.0.0.1 -p2222 — 80×23

```
[root@sandbox ~]# git clone https://github.com/rcongiu/Hive-JSON-Serde
Initialized empty Git repository in /root/Hive-JSON-Serde/.git/
remote: Counting objects: 2297, done.
remote: Total 2297 (delta 0), reused 0 (delta 0), pack-reused 2297
Receiving objects: 100% (2297/2297), 503.08 KiB | 447 KiB/s, done.
Resolving deltas: 100% (717/717), done.
[root@sandbox ~]# cd Hive-JSON-Serde/
[root@sandbox Hive-JSON-Serde]# ./../apache-maven-3.2.1/bin/mvn -Phdp23 clean pa
ckage
[INFO] Scanning for projects...
Downloading: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon-ss
h-external/2.2/wagon-ssh-external-2.2.pom
Downloaded: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon-ssh
-external/2.2/wagon-ssh-external-2.2.pom (5 KB at 9.6 KB/sec)
Downloading: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon-pr
oviders/2.2/wagon-providers-2.2.pom
Downloaded: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon-pro
viders/2.2/wagon-providers-2.2.pom (3 KB at 51.0 KB/sec)
Downloading: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon/2.
2/wagon-2.2.pom
Downloaded: http://repo.maven.apache.org/maven2/org/apache/maven/wagon/wagon/2.2
/wagon-2.2.pom (16 KB at 196.0 KB/sec)
Downloading: http://repo.maven.apache.org/maven2/org/apache/maven/maven-parent/2
```

Wait for its completion, and then you must copy the serde jar to the Hive lib:

```
cp json-serde/target/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar
/usr/hdp/2.6.0.3-8/hive/lib
cp json-serde/target/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar
/usr/hdp/2.6.0.3-8/hive2/lib
```

DO NOT forget to Restart Hive from Ambari,



Summary

Hive Metastore ✓ Started No alerts
HiveServer2 ✓ Started No alerts
WebHCat Server ✓ Started No alerts
Hive Client 1 Hive Client Installed
HiveServer2 JDBC URL jdbc:hive2://sandbox.hortonworks.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeper

We’re going to attempt to get the sentiment of each tweet by matching the words in the tweets with a sentiment dictionary. From this we can determine the sentiment of each tweet and analyze it from there.

Next, you’ll need to SSH into the sandbox again and run the following two commands

```
# Virtualbox
        sudo -u hdfs hdfs dfs -chown -R maria_dev /tmp/tweets_staging
        sudo -u hdfs hdfs dfs -chmod -R 777 /tmp/tweets_staging
```

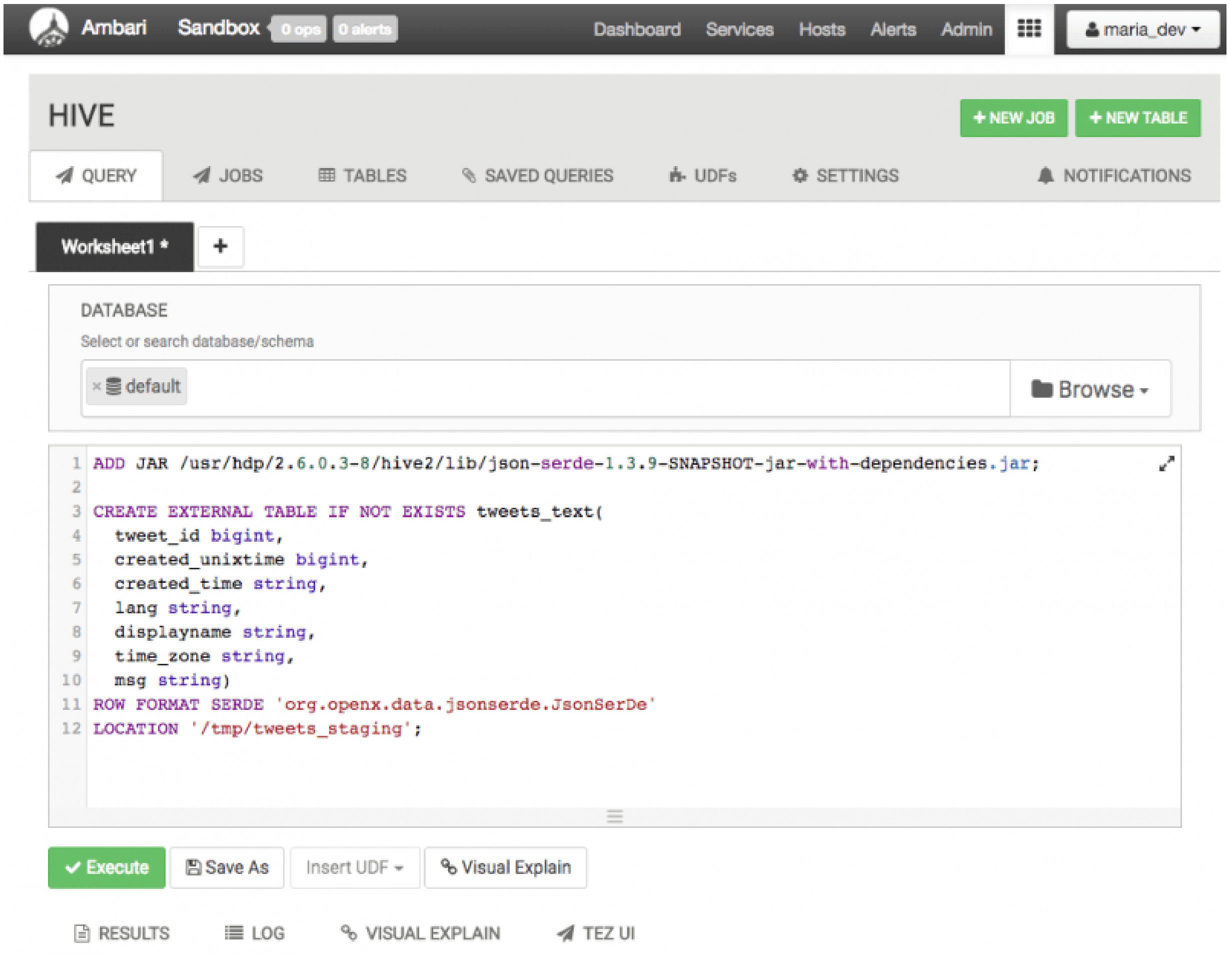
After the commands complete let’s go to the Hive view. Head over to http://sandbox.hortonworks.com:8080. Login into Ambari. Refer to Learning the Ropes of the Hortonworks Sandbox if you need assistance with logging into Ambari.

Note: login credentials are `maria_dev/maria_dev` (Virtualbox). Use the dropdown menu at the top to get to the Hive view.

Enter **Hive View 2.0**. Execute the following command to create a table for the tweets

```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar;

CREATE EXTERNAL TABLE IF NOT EXISTS tweets_text(
  tweet_id bigint,
  created_unixtime bigint,
  created_time string,
  lang string,
  displayname string,
  time_zone string,
  msg string)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
LOCATION '/tmp/tweets_staging';
```



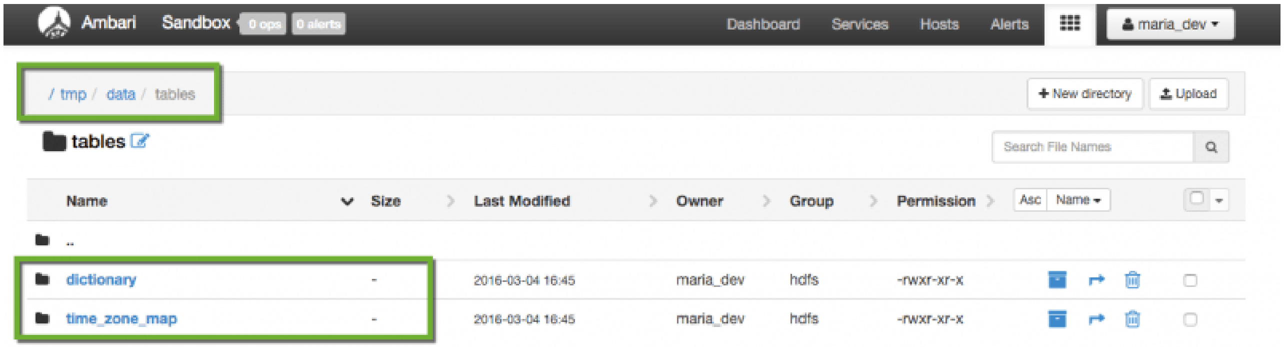
```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar;

CREATE EXTERNAL TABLE IF NOT EXISTS tweets_text(
  tweet_id bigint,
  created_unixtime bigint,
  created_time string,
  lang string,
  displayname string,
  time_zone string,
  msg string)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
LOCATION '/tmp/tweets_staging';
```

Now we're going to need to do some data analysis.

First, you're going to need to head to the **HDFS Files View** and create a new directory in `/tmp/data/tables`

Then create two new directories inside of `/tmp/data/tables`. One named **time_zone_map** and another named **dictionary**



Name	Size	Last Modified	Owner	Group	Permission
..					
dictionary	-	2016-03-04 16:45	maria_dev	hdfs	-rwxr-xr-x
time_zone_map	-	2016-03-04 16:45	maria_dev	hdfs	-rwxr-xr-x

In each of the folders respectively you'll need to upload the `dictionary.tsv` file, and the `time_zone_map.tsv` file to each of their respective directories.

After doing so, you'll need to run the following command on the Sandbox:

```
sudo -u hdfs hdfs dfs -chmod -R 777 /tmp/data/tables
```

Finally, run the following two commands in **Hive View 2.0**:

The first table created is **dictionary** and the dataset loaded into the table is in this path: `/tmp/data/tables/dictionary`.

```
CREATE EXTERNAL TABLE if not exists dictionary (
        type string,
        length int,
        word string,
        pos string,
        stemmed string,
        polarity string )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY 't'
STORED AS TEXTFILE
LOCATION '/tmp/data/tables/dictionary';
```

The second table created is **time_zone_map** and the dataset loaded into the table is in this path: `/tmp/data/tables/time_zone_map`.

```
CREATE EXTERNAL TABLE if not exists time_zone_map (
    time_zone string,
    country string,
    notes string )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY 't'
STORED AS TEXTFILE
LOCATION '/tmp/data/tables/time_zone_map';
```

Next, we'll need to create two table views from our tweets which will simplify the columns the data we have access to.

tweets_simple view:

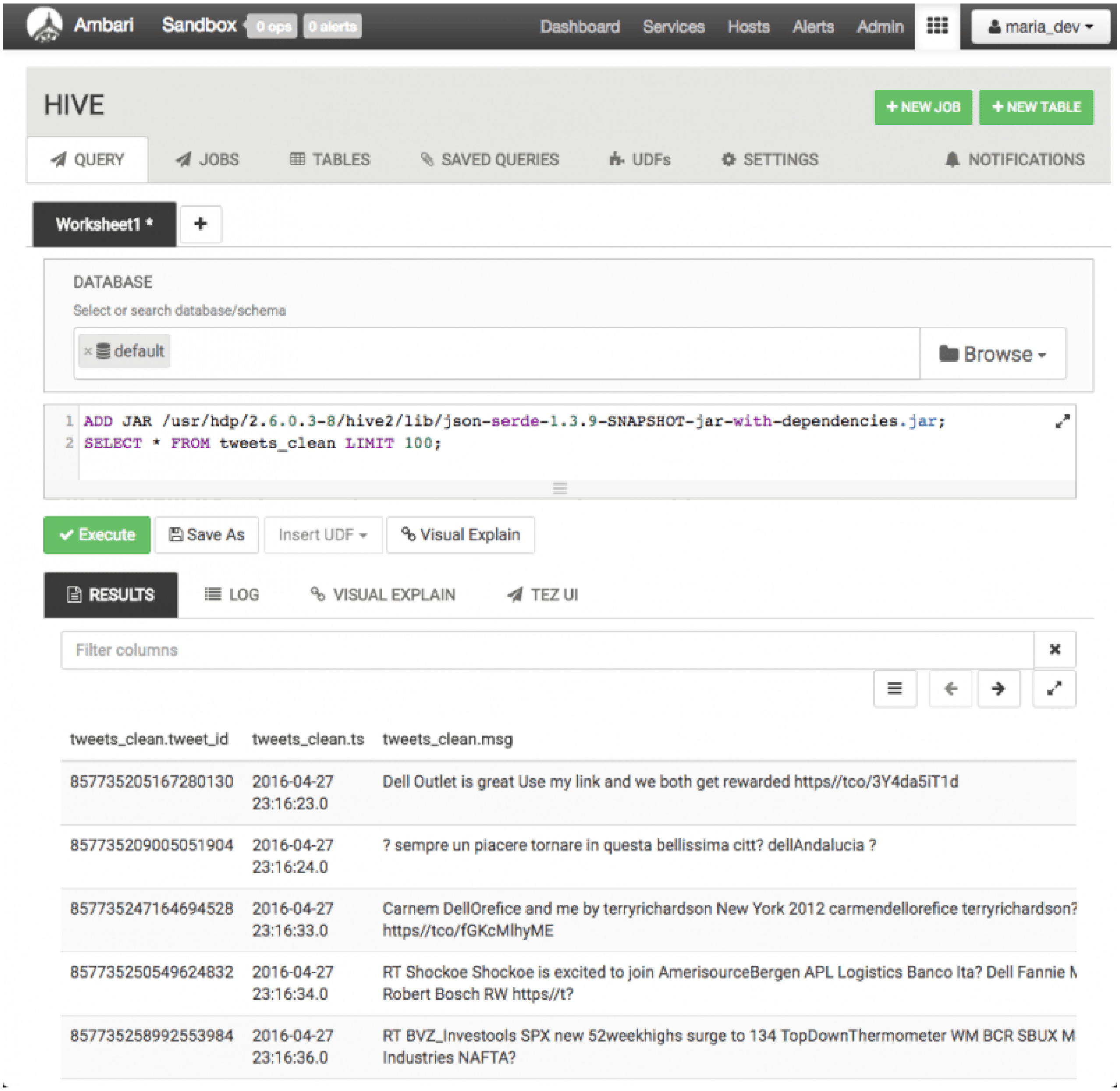
```
CREATE VIEW IF NOT EXISTS tweets_simple AS
SELECT
  tweet_id,
  cast ( from_unixtime( unix_timestamp(concat( '2016 ',
substring(created_time,5,15)), 'yyyy MMM dd hh:mm:ss')) as timestamp) ts,
  msg,
  time_zone
FROM tweets_text;
```

tweets_clean view:

```
CREATE VIEW IF NOT EXISTS tweets_clean AS
SELECT
  t.tweet_id,
  t.ts,
  t.msg,
  m.country
 FROM tweets_simple t LEFT OUTER JOIN time_zone_map m ON t.time_zone =
m.time_zone;
```

After running the above commands, you should be able to run:

```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-
dependencies.jar;
SELECT * FROM tweets_clean LIMIT 100;
```



```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar;
SELECT * FROM tweets_clean LIMIT 100;
```

tweets_clean.tweet_id	tweets_clean.ts	tweets_clean.msg
857735205167280130	2016-04-27 23:16:23.0	Dell Outlet is great Use my link and we both get rewarded https//tco/3Y4da5iT1d
857735209005051904	2016-04-27 23:16:24.0	? sempre un piacere tornare in questa bellissima citt? dellAndalucia ?
857735247164694528	2016-04-27 23:16:33.0	Carnem DellOrefice and me by terryrichardson New York 2012 carmendellorefice terryrichardson? https//tco/fGKcMlhyME
857735250549624832	2016-04-27 23:16:34.0	RT Shockoe Shockoe is excited to join AmerisourceBergen APL Logistics Banco Ita? Dell Fannie M Robert Bosch RW https//t?
857735258992553984	2016-04-27 23:16:36.0	RT BVZ_Investools SPX new 52weekhighs surge to 134 TopDownThermometer WM BCR SBUX M Industries NAFTA?

Now that we’ve cleaned our data we can get around to computing the sentiment. Use the following Hive commands to create some views that will allow us to do that.

l1 view, **l2** view, **l3** view:

```
-- Compute sentiment
create view IF NOT EXISTS l1 as select tweet_id, words from tweets_text
lateral view explode(sentences(lower(msg))) dummy as words;

create view IF NOT EXISTS l2 as select tweet_id, word from l1 lateral view
explode( words ) dummy as word;

create view IF NOT EXISTS l3 as select
    tweet_id,
    l2.word,
    case d.polarity
      when  'negative' then -1
      when 'positive' then 1
      else 0 end as polarity
from l2 l2 left outer join dictionary d on l2.word = d.word;
```

Now that we could compute some sentiment values we can assign whether a tweet was **positive**, **neutral**, or **negative**. Use this next Hive command to do that.

tweets_sentiment table:

```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-
dependencies.jar;

create table IF NOT EXISTS tweets_sentiment stored as orc as select
  tweet_id,
  case
    when sum( polarity ) > 0 then 'positive'
    when sum( polarity ) < 0 then 'negative'
    else 'neutral' end as sentiment
from l3 group by tweet_id;
```

Note: We will need to specify the location of the json-serde library JAR file since this table references another table that works with json data.

Lastly, to make our analysis somewhat easier we are going to turn those ‘positive’, ‘negative’, and ‘neutral’ values into numerical values using the next Hive command

tweetsbi table:

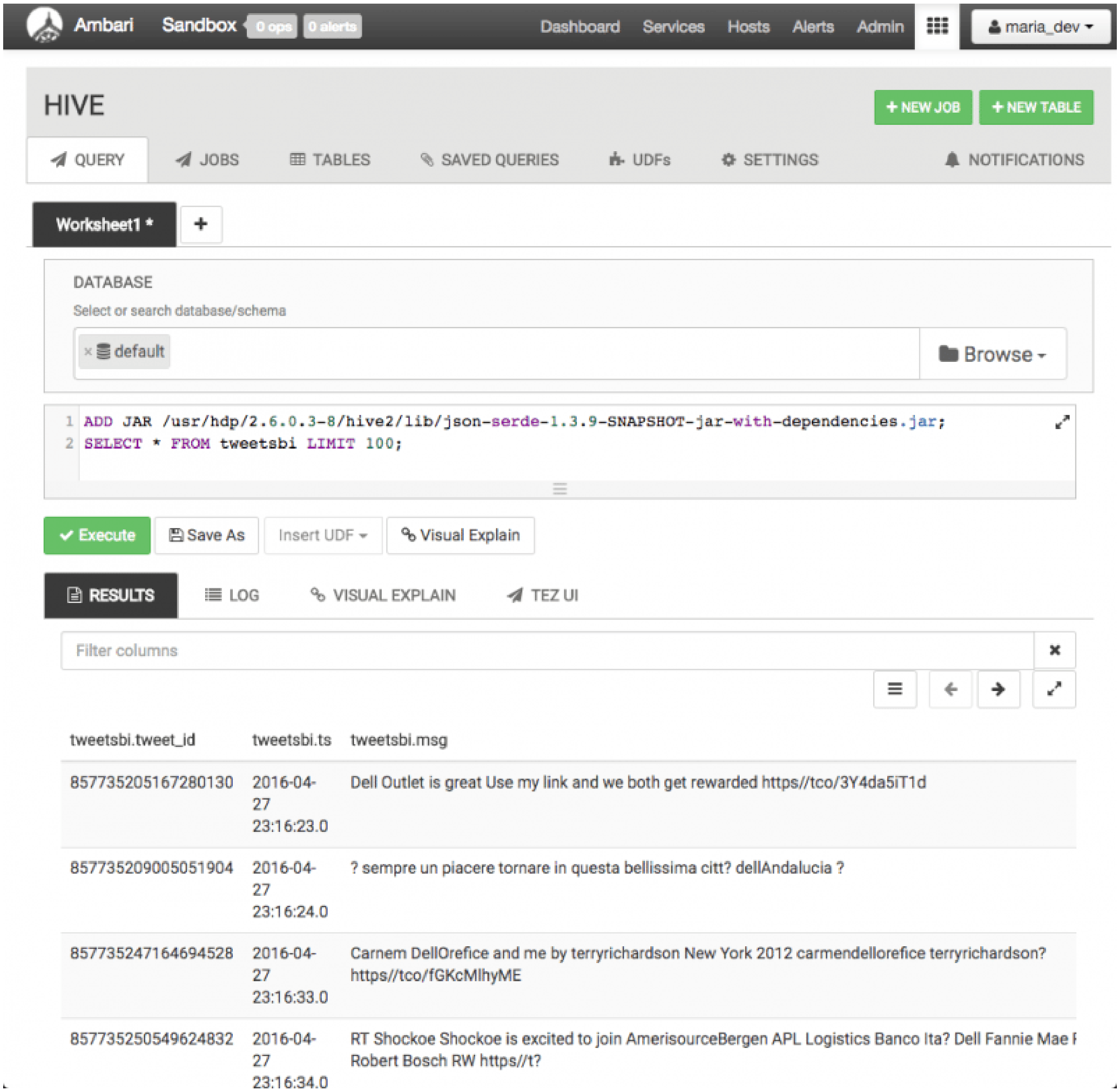
```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-
dependencies.jar;

CREATE TABLE IF NOT EXISTS tweetsbi
STORED AS ORC
AS SELECT
  t.*,
  case s.sentiment
    when 'positive' then 2
    when 'neutral' then 1
    when 'negative' then 0
  end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.tweet_id =
s.tweet_id;
```

Load the tweetsbi data:

```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-
dependencies.jar;
SELECT * FROM tweetsbi LIMIT 100;
```

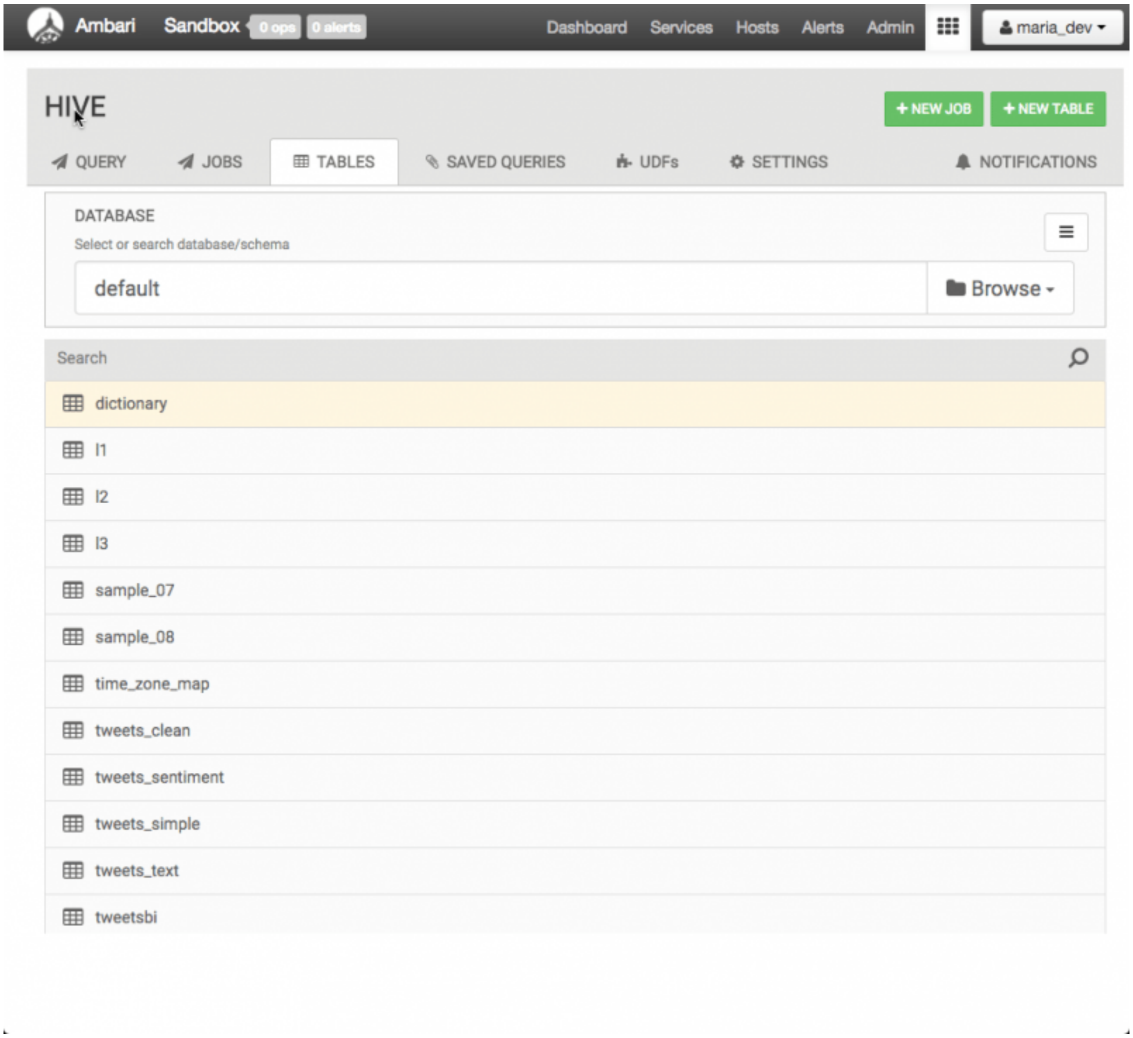
This command should yield our results table as shown below.



```
ADD JAR /usr/hdp/2.6.0.3-8/hive2/lib/json-serde-1.3.9-SNAPSHOT-jar-with-dependencies.jar;
SELECT * FROM tweetsbi LIMIT 100;
```

tweetsbi.tweet_id	tweetsbi.ts	tweetsbi.msg
857735205167280130	2016-04-27 23:16:23.0	Dell Outlet is great Use my link and we both get rewarded https//tco/3Y4da5iT1d
857735209005051904	2016-04-27 23:16:24.0	? sempre un piacere tornare in questa bellissima citt? dellAndalucia ?
857735247164694528	2016-04-27 23:16:33.0	Carnem DellOrefice and me by terryrichardson New York 2012 carmendellorefice terryrichardson? https//tco/fGKcMlhyME
857735250549624832	2016-04-27 23:16:34.0	RT Shockoe Shockoe is excited to join AmerisourceBergen APL Logistics Banco Ita? Dell Fannie Mae F Robert Bosch RW https//t?

Now we have created all our hive tables and views. They should appear in the **TABLES** tab where you can see all tables and views in your current database: as shown below:



DATABASE

Select or search database/schema

default

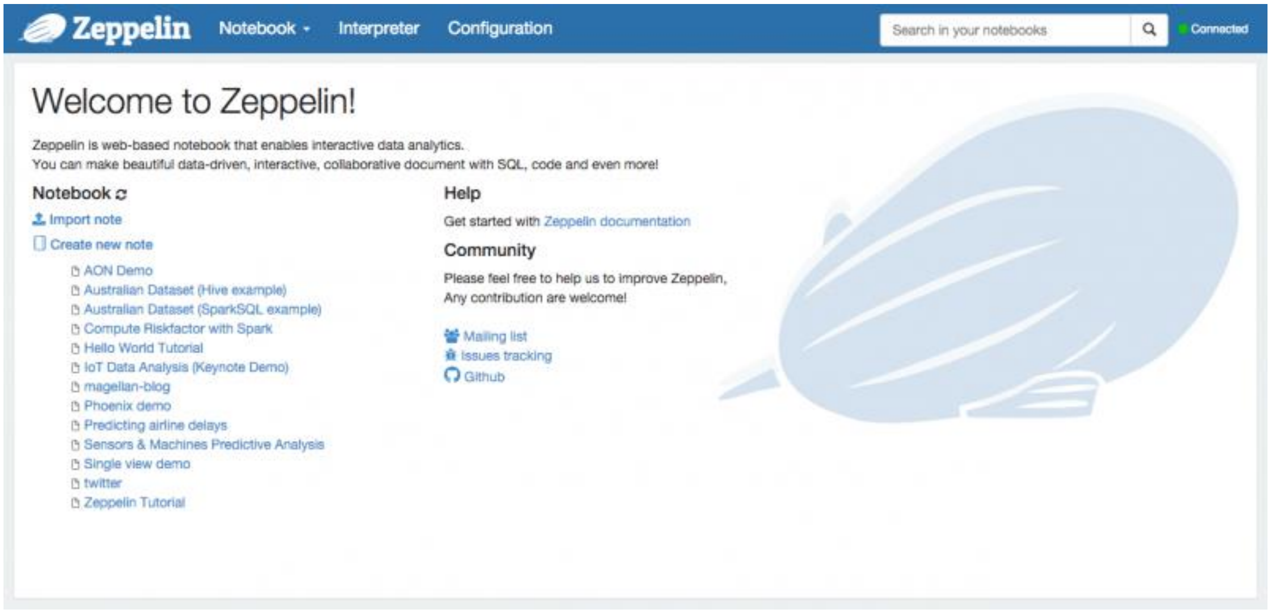
dictionary
l1
l2
l3
sample_07
sample_08
time_zone_map
tweets_clean
tweets_sentiment
tweets_simple
tweets_text
tweetsbi

Try the new Hive Visualization tab!

On the right-hand side of the screen try clicking the **graph icon** in the column located in row 3. It will bring up a new tab where you can directly create charts using your query results in Hive! Now that we can access the sentiment data in our Hive table let's do some visualization on the analysis using Apache Zeppelin.

VISUALIZE SENTIMENT WITH ZEPPELIN

Make sure your Zeppelin service is started in Ambari, then head over to the Zeppelin at http://sandbox.hortonworks.com:9995.



Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Notebook

- Import note
- Create new note
  - AON Demo
  - Australian Dataset (Hive example)
  - Australian Dataset (SparkSQL example)
  - Compute Riskfactor with Spark
  - Hello World Tutorial
  - IoT Data Analysis (Keynote Demo)
  - magellan-blog
  - Phoenix demo
  - Predicting airline delays
  - Sensors & Machines Predictive Analysis
  - Single view demo
  - twitter
  - Zeppelin Tutorial

Help

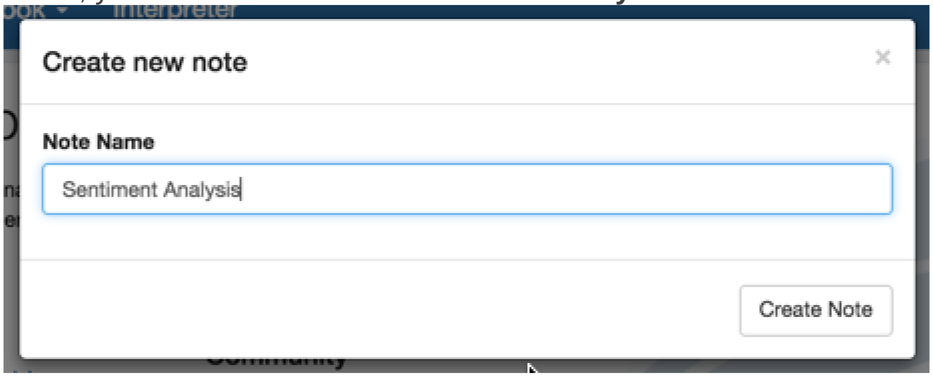
Get started with Zeppelin documentation

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

- Mailing list
- Issues tracking
- Github

Use the **Notebook** dropdown menu at the top of the screen and click **+ Create New Note**. After which, you can name the note **Sentiment Analysis**.



Create new note

Note Name

Sentiment Analysis

After creating the note, open it up to the blank Notebook screen and type the following command.

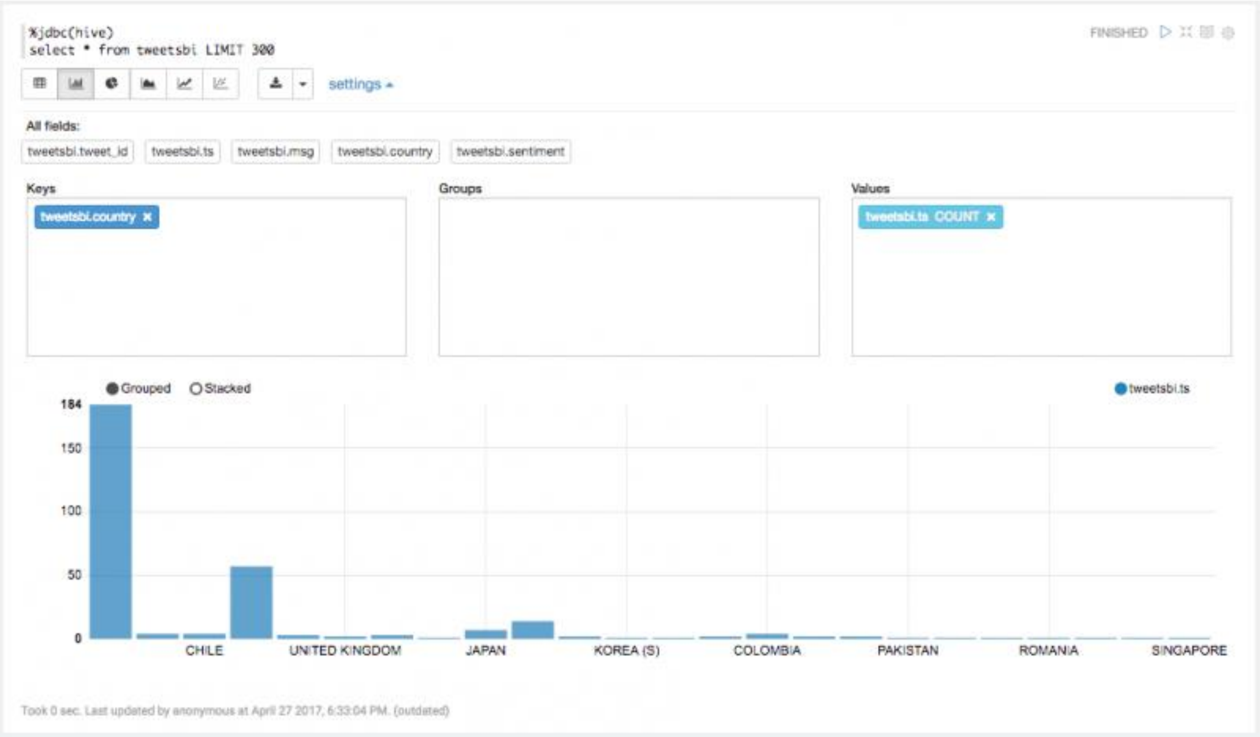
```
%jdbc(hive)
select * from tweetsbi LIMIT 300
```

We're limiting our query to just `300` results because right now we won't need to see everything.

And if you've collected a lot of data from NiFi, then it could slow down your computer.

- Arrange your results so that your chart is a **bar graph**.
- The `tweetsbi.country` column is a **key** and the `tweetsbi.sentiment` as the **value**.
- Make sure that **sentiment** is labeled as **COUNT**.
- Run the query by **clicking the arrow on the right hand side**, or by pressing **Shift+Enter**.

Your results should look like the following:



```
%jdbc(hive)
select * from tweetsbi LIMIT 300
```

All fields: tweetsbi.tweet_id, tweetsbi.ts, tweetsbi.msg, tweetsbi.country, tweetsbi.sentiment

Keys: tweetsbi.country

Groups:

Values: tweetsbi.ts COUNT

Took 0 sec. Last updated by anonymous at April 27 2017, 6:33:04 PM. (outdated)

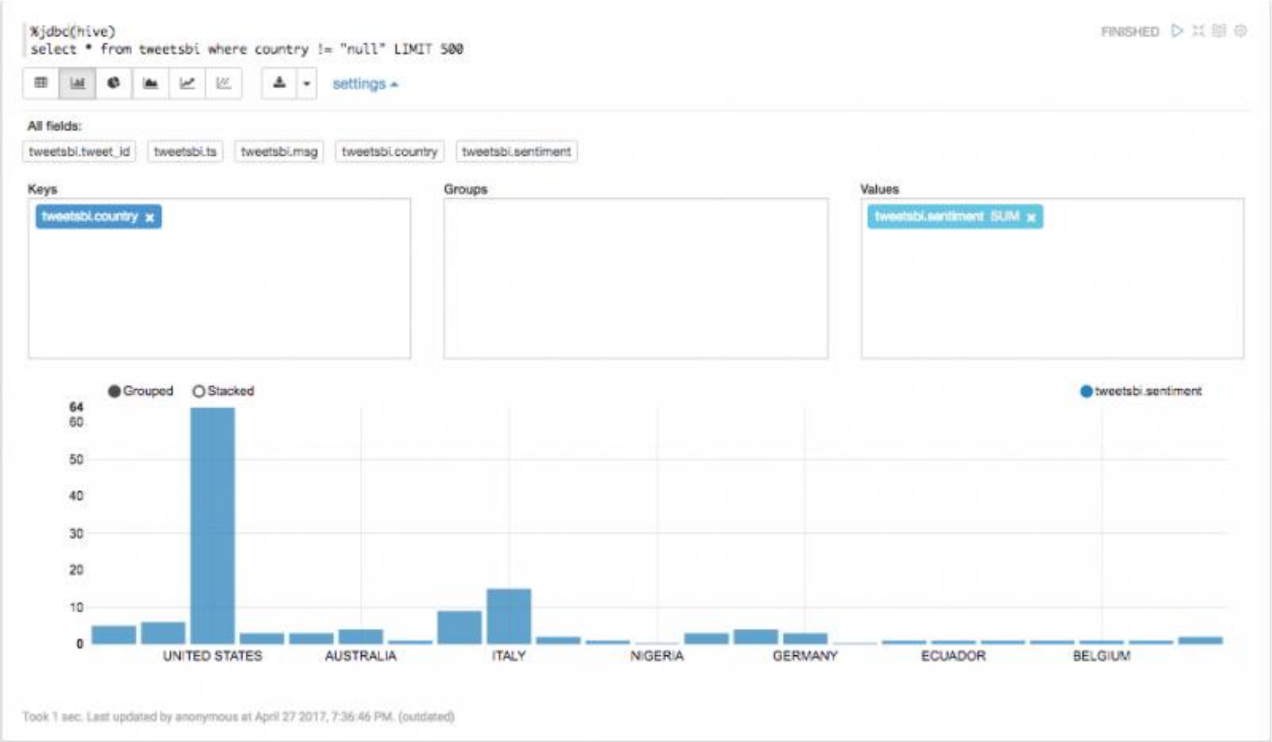
After looking at the results we see that if we group by country that many tweets are labeled as null.

For the sake of visualization let’s remove any tweets that might appear in our select statement that have a country value of “null”, as well as increase our result limit to 500.

Scroll down to the next note and create run the following query, and set up the results the same way as above.

Note Before running Hive queries, restart the Spark Interpreter since Spark jobs take up cluster resources. Click the **Interpreter** tab located near Zeppelin logo at the top of the page, under **Spark** click on the button that says **restart**.

```
%jdbc(hive)
select * from tweetsbi where country != "null" LIMIT 500
```



```
%jdbc(hive)
select * from tweetsbi where country != "null" LIMIT 500
```

FINISHED

All fields: tweetsbi.tweet_id tweetsbi.ts tweetsbi.msg tweetsbi.country tweetsbi.sentiment

Keys: tweetsbi.country

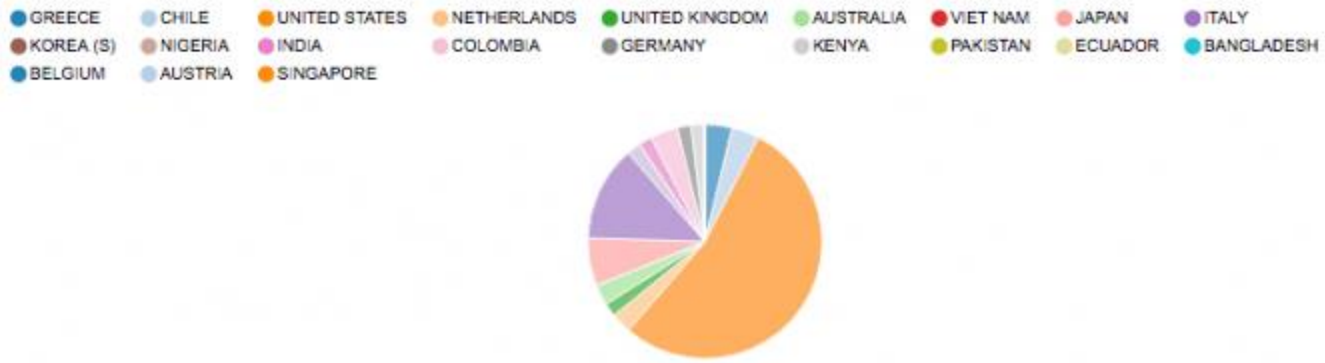
Groups:

Values: tweetsbi.sentiment SUM

Took 1 sec. Last updated by anonymous at April 27 2017, 7:36:46 PM. (outdated)

Great! Now given the data we have, we can at least have an idea of the distribution of users whose tweets come from certain countries!

You can also experiment with this and try a pie chart as well.

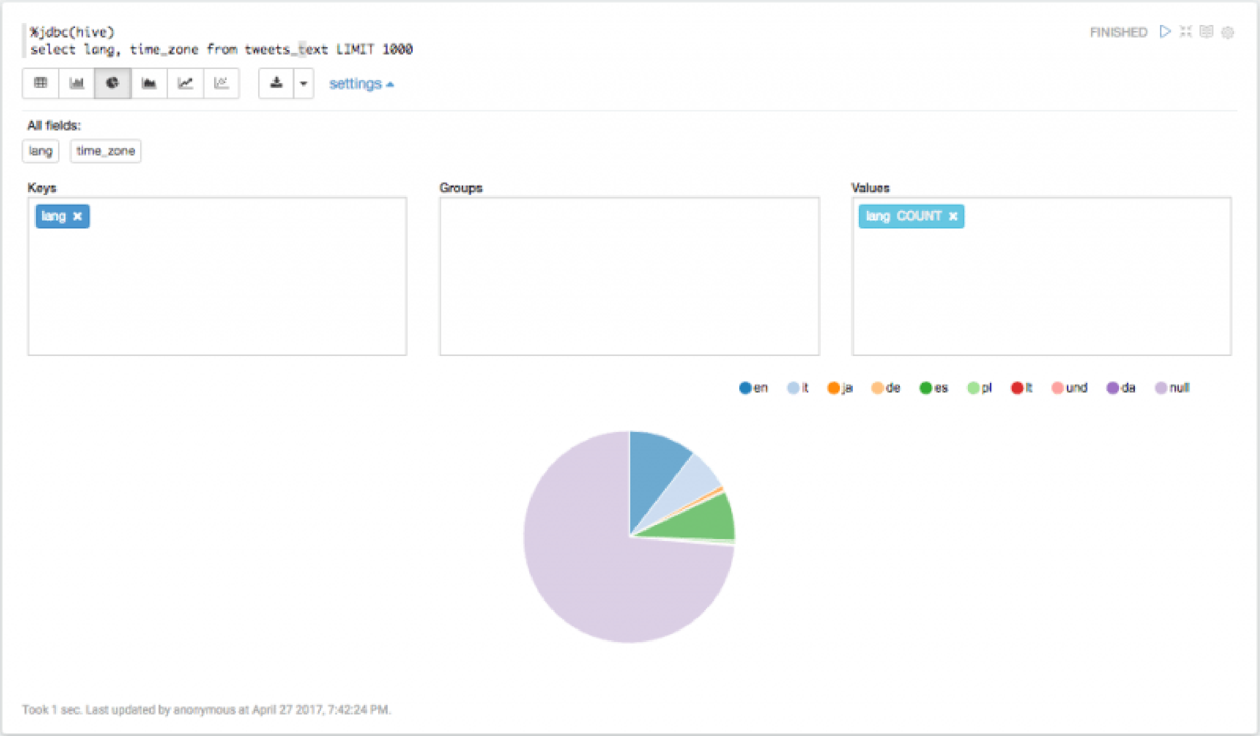


In our original raw tweet data from NiFi we also collected the language from our users as well. So we can also get an idea of the distribution of languages!

Run the following query and make

- **lang** as the **Key**
- **COUNT** for **lang** in **values**

```
%jdbc(hive)
select lang, time_zone from tweets_text LIMIT 1000
```



```
%jdbc(hive)
select lang, time_zone from tweets_text LIMIT 1000
```

All fields: lang time_zone

Keys: lang

Groups:

Values: lang COUNT

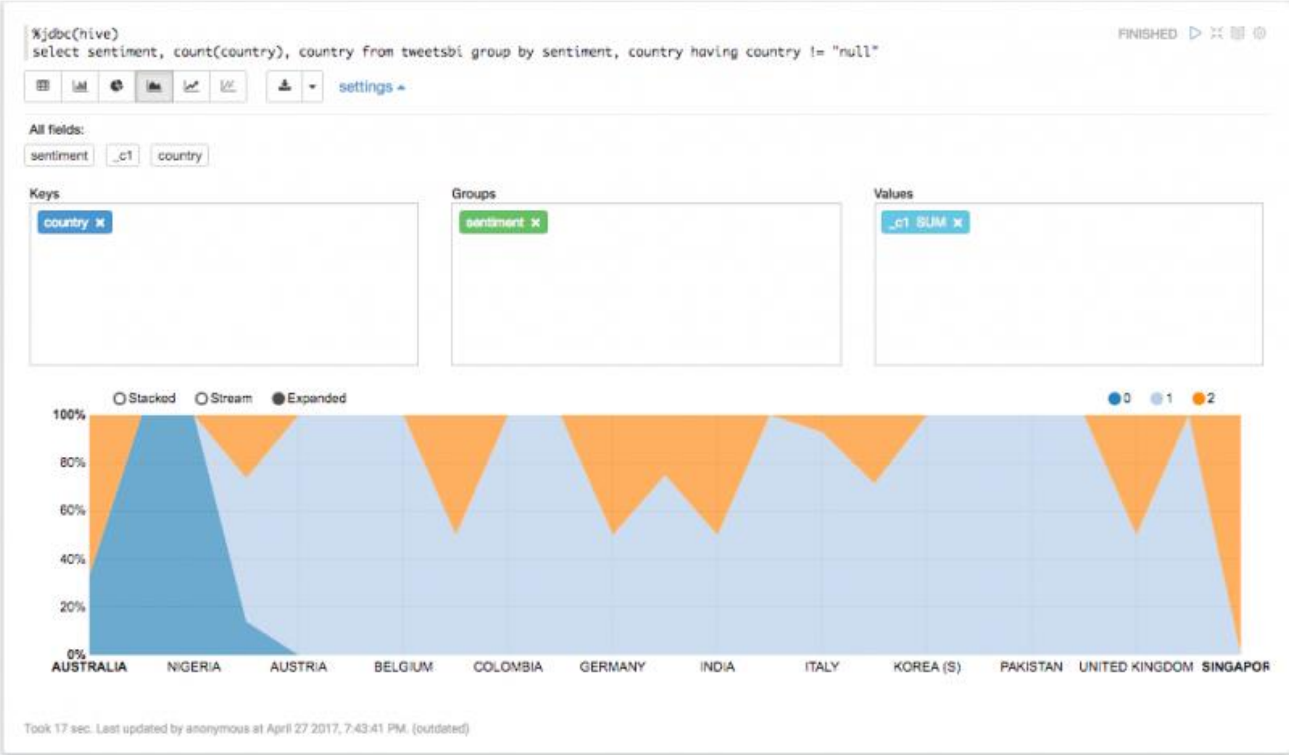
Took 1 sec. Last updated by anonymous at April 27 2017, 7:42:24 PM.

If you have not seen from our earlier analysis in Hive

- A bad or negative sentiment is **0**
- A neutral sentiment value is **1**.
- A positive sentiment value is **2**.

Using this we can now look at individual countries and see the sentiment distributions of each.

```
%jdbc(hive)
select sentiment, count(country), country from tweetsbi group by sentiment,
country having country != "null"
```



```
%jdbc(hive)
select sentiment, count(country), country from tweetsbi group by sentiment, country having country != "null"
```

FINISHED

settings

All fields: sentiment, _c1, country

Keys: country

Groups: sentiment

Values: _c1 SUM

Took 17 sec. Last updated by anonymous at April 27 2017, 7:43:41 PM. (outdated)

Using this data you can determine how you might want to market your products to different countries!