Analyzing Twitter data using Apache Hive

The goal is to do basic analysis on Twitter data set and help people that are initiating the use of Apache Hive.

Apache Hive is a language very similar to SQL and

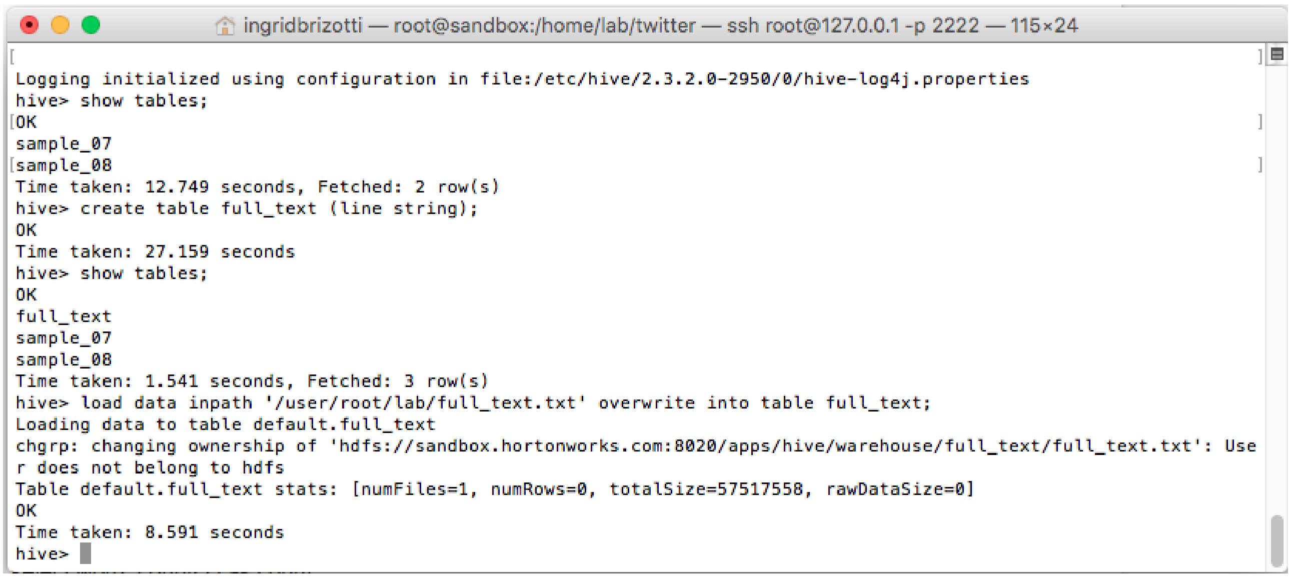
- is part of the Hadoop ecosystem
- is used to explore data in HDFS (Hadoop Distributed File System)
- only works in structured data

To do the analysis I'd used Apache Hive from Hortonworks(sandbox) installed in a virtual machine. Here, some useful sites to take a look at:
https://hive.apache.org/
http://br.hortonworks.com/products/sandbox/

Steps:

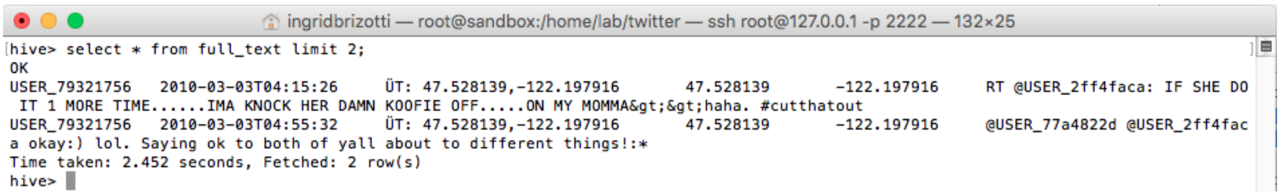
1) **Load data**
2) **Find the most popular words**
3) **Extract new variables**
4) **Find top 10 users who tweet longest tweets**
5) **Show all tweets made by an specific user**
6) **Frequency the number of tweets in an specific date**
7) **Search for tweets in an specific area (using BoundingBox)**
8) **Calculate the number of tweets per user**
9) **Find top 10 tweeters (number of tweets/user) in New York area:**
10) **Find the hour of the day that generated most number of tweets on an specific date**
11) **Find the most tweeter that tweeted from most distinct location**
12) **Find 5 most popular topics (hashtags)**
13) **Find most frequently mentioned twitter (@)**

Load data:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 115×24

```
Logging initialized using configuration in file:/etc/hive/2.3.2.0-2950/0/hive-log4j.properties
hive> show tables;
OK
sample_07
sample_08
Time taken: 12.749 seconds, Fetched: 2 row(s)
hive> create table full_text (line string);
OK
Time taken: 27.159 seconds
hive> show tables;
OK
full_text
sample_07
sample_08
Time taken: 1.541 seconds, Fetched: 3 row(s)
hive> load data inpath '/user/root/lab/full_text.txt' overwrite into table full_text;
Loading data to table default.full_text
chgrp: changing ownership of 'hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/full_text/full_text.txt': Use
r does not belong to hdfs
Table default.full_text stats: [numFiles=1, numRows=0, totalSize=57517558, rawDataSize=0]
OK
Time taken: 8.591 seconds
hive> 
```

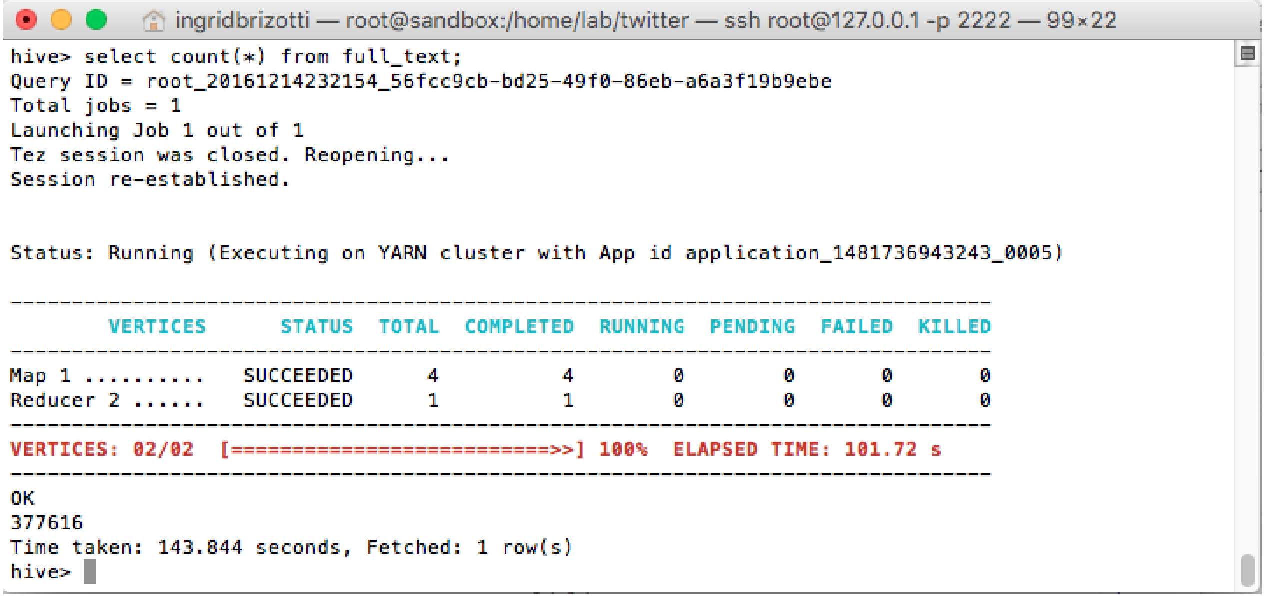
Display the first 2 rows:



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 132×25

```
hive> select * from full_text limit 2;
OK
USER_79321756   2010-03-03T04:15:26     ÜT: 47.528139,-122.197916       47.528139       -122.197916     RT @USER_2ff4faca: IF SHE DO
 IT 1 MORE TIME......IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&gt;&gt;haha. #cutthatout
USER_79321756   2010-03-03T04:55:32     ÜT: 47.528139,-122.197916       47.528139       -122.197916     @USER_77a4822d @USER_2ff4fac
a okay:) lol. Saying ok to both of yall about to different things!:*
Time taken: 2.452 seconds, Fetched: 2 row(s)
hive> 
```

Check the number of rows:

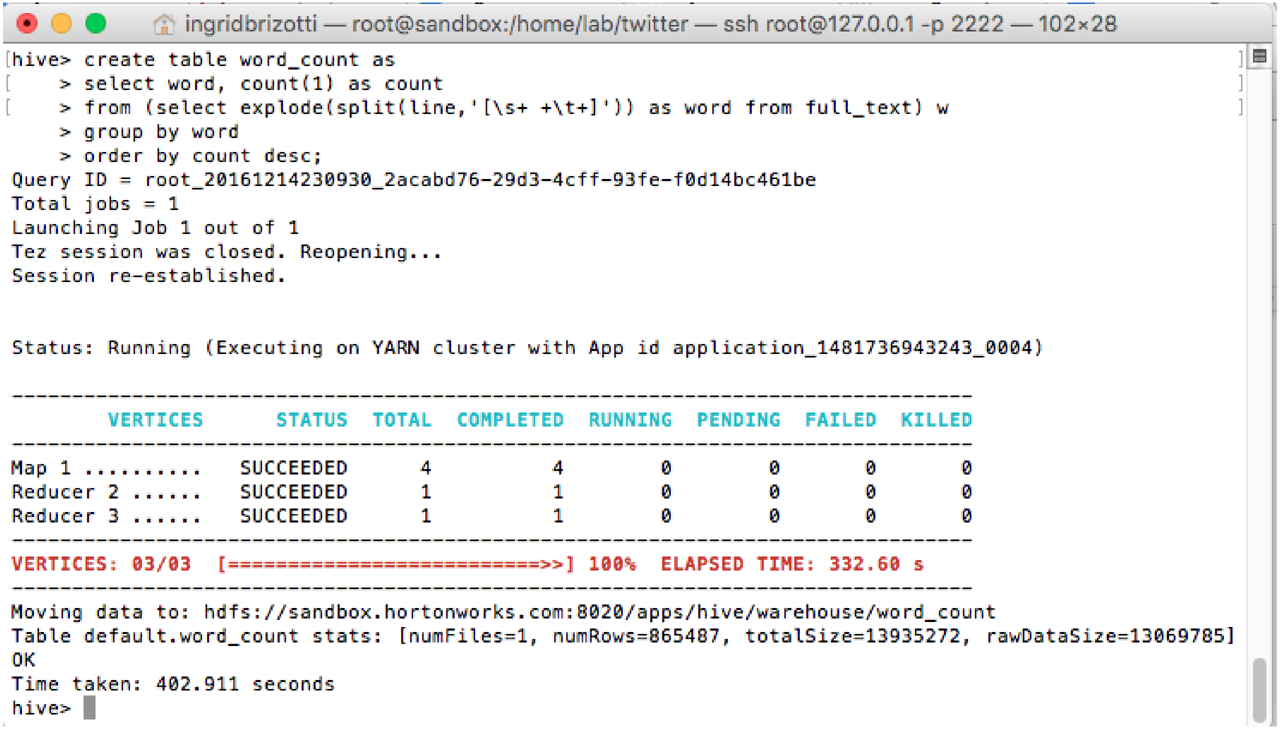


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 99×22

```
hive> select count(*) from full_text;
Query ID = root_20161214232154_56fcc9cb-bd25-49f0-86eb-a6a3f19b9ebe
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481736943243_0005)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 101.72 s
--------------------------------------------------------------------------------
OK
377616
Time taken: 143.844 seconds, Fetched: 1 row(s)
hive> 
```

The data set has 377.616 rows

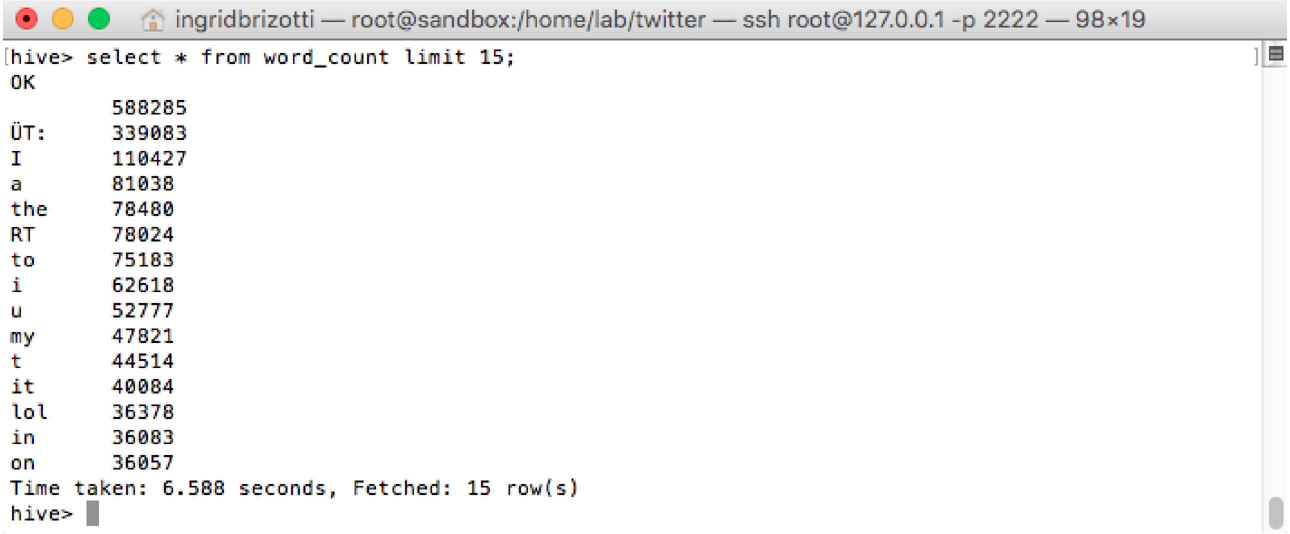
Counts how many times a word appears and put the result in descending order:



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 102×28

```
hive> create table word_count as
    > select word, count(1) as count
    > from (select explode(split(line,'[\s+ +\t+]')) as word from full_text) w
    > group by word
    > order by count desc;
Query ID = root_20161214230930_2acabd76-29d3-4cff-93fe-f0d14bc461be
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481736943243_0004)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 332.60 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/word_count
Table default.word_count stats: [numFiles=1, numRows=865487, totalSize=13935272, rawDataSize=13069785]
OK
Time taken: 402.911 seconds
hive> 
```

Shows the first 15 more popular words:

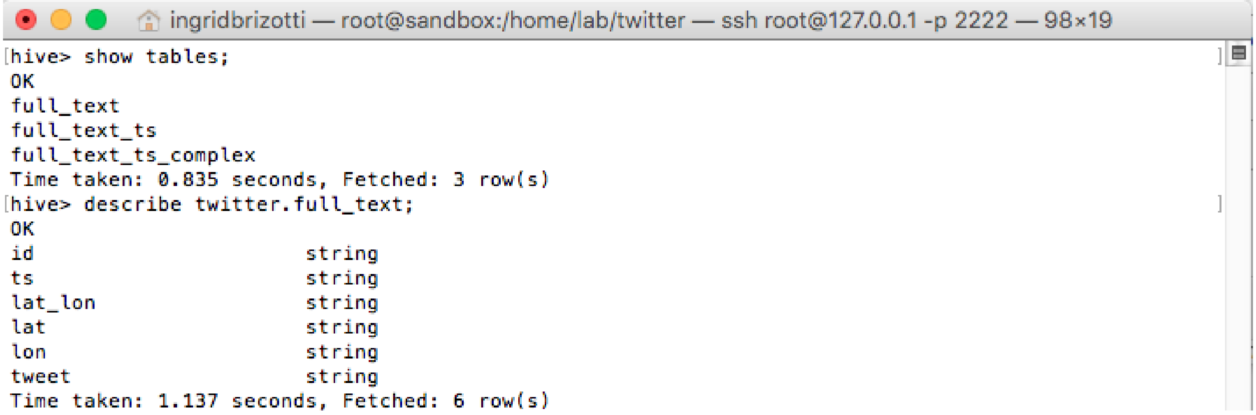

ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 98×19

```
[hive> select * from word_count limit 15;
OK
        588285
ÜT:     339083
I       110427
a       81038
the     78480
RT      78024
to      75183
i       62618
u       52777
my      47821
t       44514
it      40084
lol     36378
in      36083
on      36057
Time taken: 6.588 seconds, Fetched: 15 row(s)
hive>
```

The full_text was load inside twitter database is using this format:

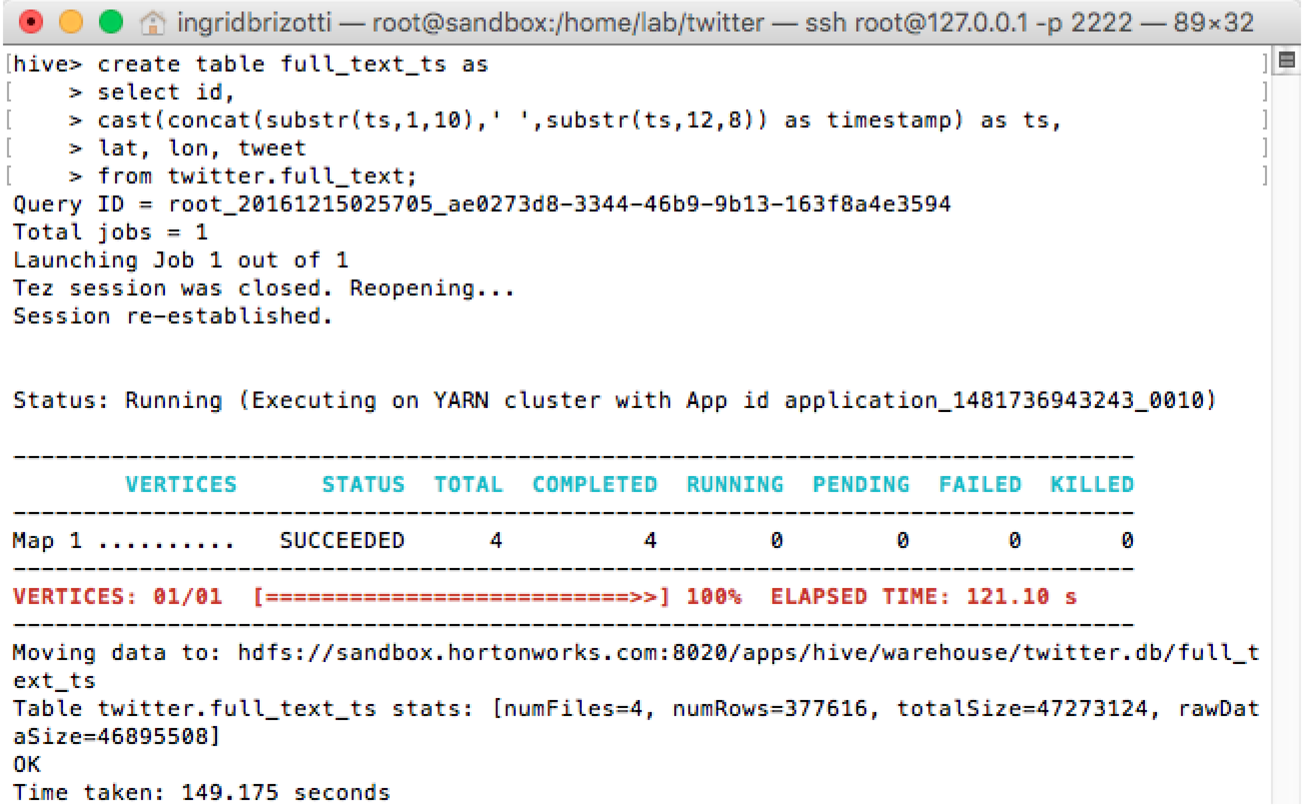
Variable	Format	Description
id	string	Twitter id
ts	string	Timestamp
lat_lon	string	Latitude and longitude
tweet	string	Tweet full text

Check full_text created on twitter database


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 98×19

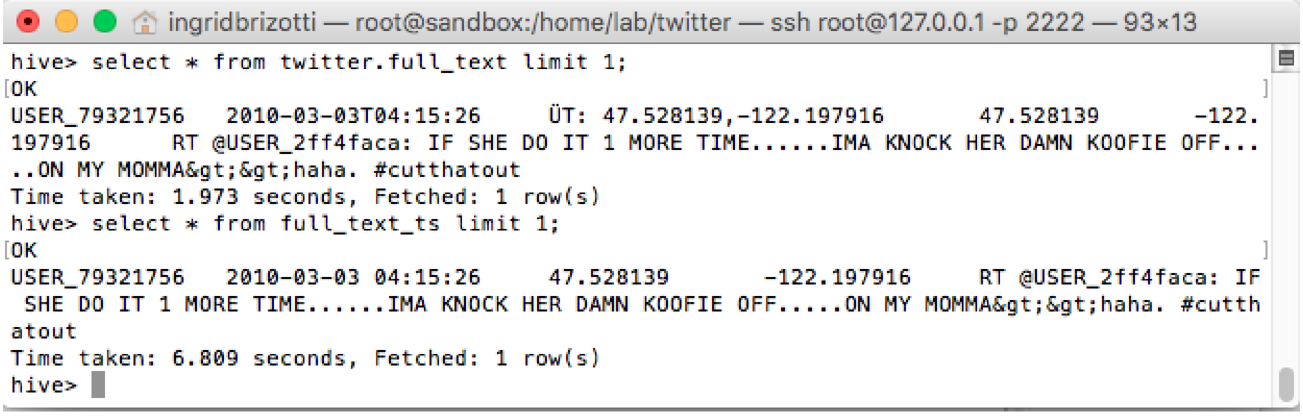
```
[hive> show tables;
OK
full_text
full_text_ts
full_text_ts_complex
Time taken: 0.835 seconds, Fetched: 3 row(s)
[hive> describe twitter.full_text;
OK
id                      string
ts                      string
lat_lon                 string
lat                     string
lon                     string
tweet                   string
Time taken: 1.137 seconds, Fetched: 6 row(s)
```

Put a space between the date and time on the timestamp(ts) variable:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 89×32

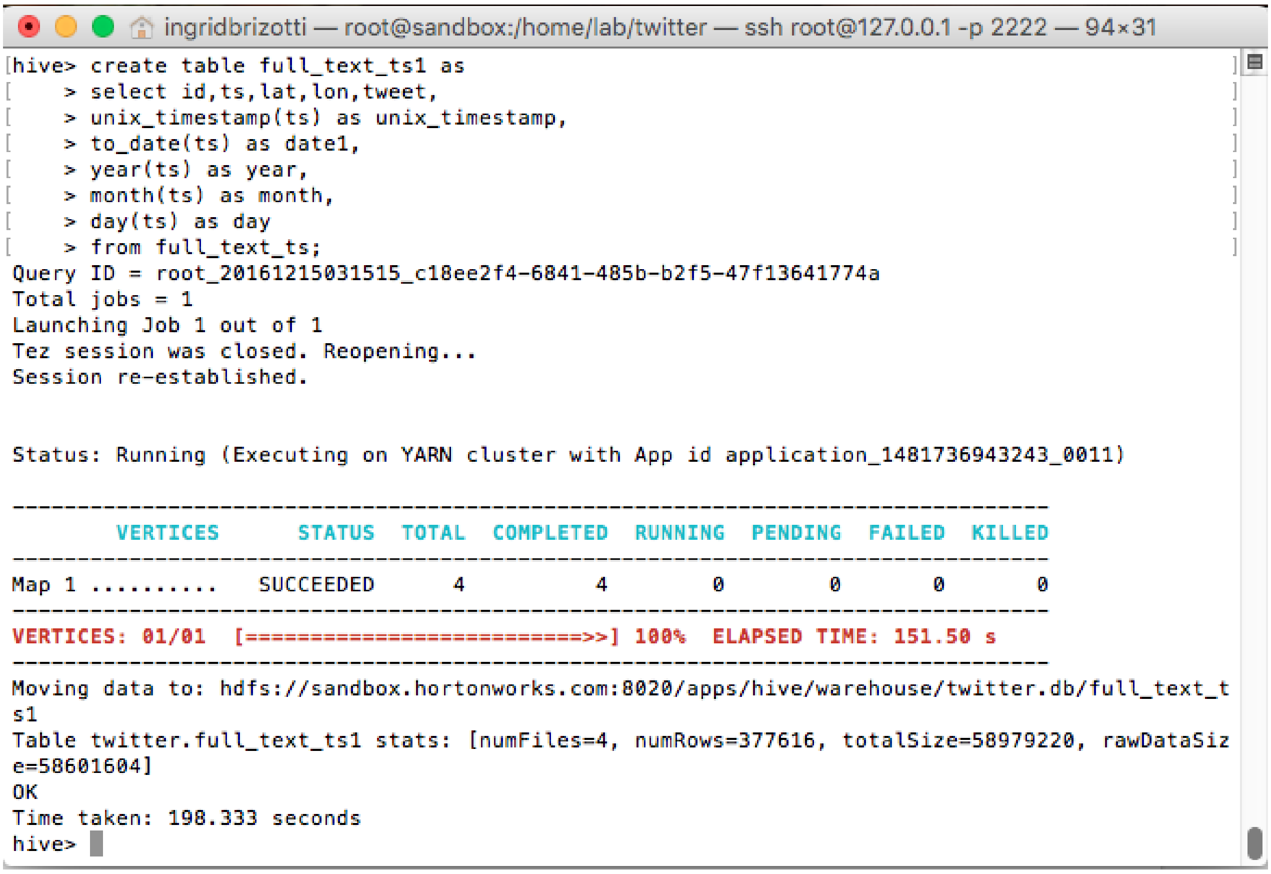
```
[hive> create table full_text_ts as
[    > select id,
[    > cast(concat(substr(ts,1,10),' ',substr(ts,12,8)) as timestamp) as ts,
[    > lat, lon, tweet
[    > from twitter.full_text;
Query ID = root_20161215025705_ae0273d8-3344-46b9-9b13-163f8a4e3594
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481736943243_0010)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 121.10 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/full_t
ext_ts
Table twitter.full_text_ts stats: [numFiles=4, numRows=377616, totalSize=47273124, rawDat
aSize=46895508]
OK
Time taken: 149.175 seconds
```

Check before and after:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 93×13

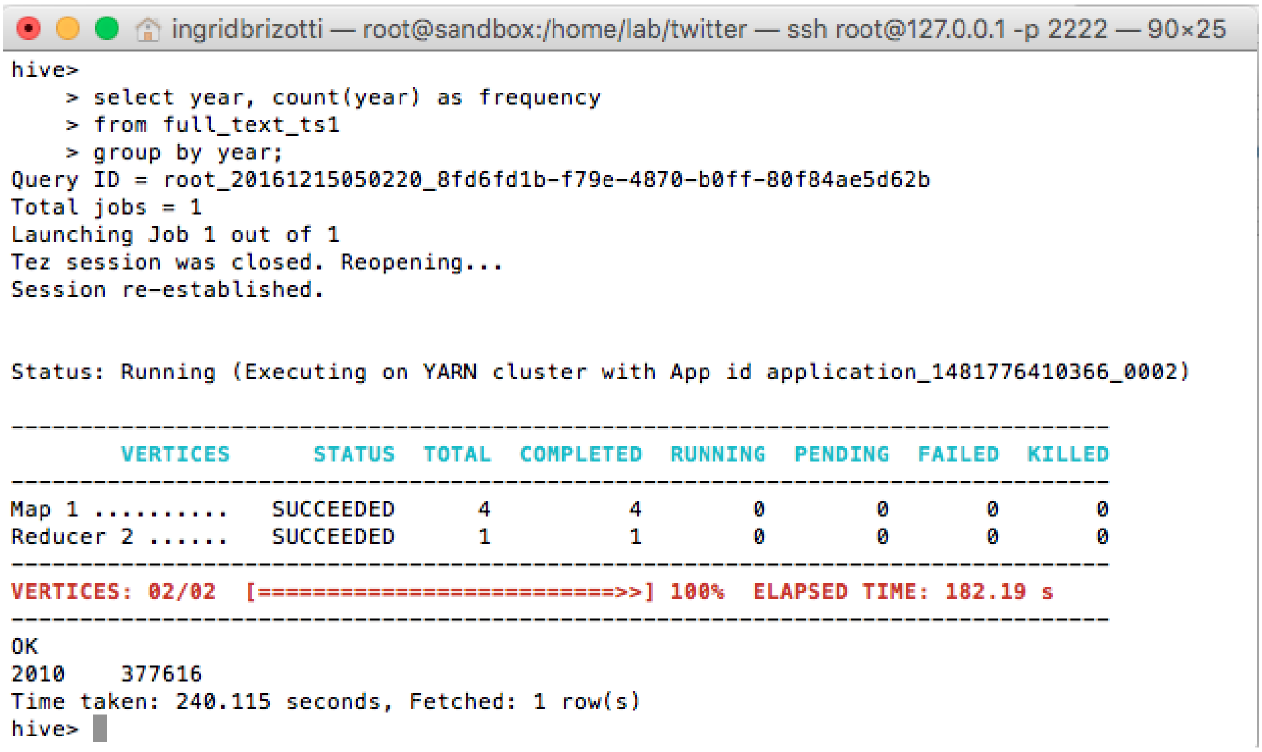
```
hive> select * from twitter.full_text limit 1;
[OK
USER_79321756   2010-03-03T04:15:26     ÜT: 47.528139,-122.197916       47.528139       -122.
197916      RT @USER_2ff4faca: IF SHE DO IT 1 MORE TIME......IMA KNOCK HER DAMN KOOFIE OFF...
..ON MY MOMMA&gt;&gt;haha. #cutthatout
Time taken: 1.973 seconds, Fetched: 1 row(s)
hive> select * from full_text_ts limit 1;
[OK
USER_79321756   2010-03-03 04:15:26     47.528139       -122.197916     RT @USER_2ff4faca: IF
 SHE DO IT 1 MORE TIME......IMA KNOCK HER DAMN KOOFIE OFF.....ON MY MOMMA&gt;&gt;haha. #cutth
atout
Time taken: 6.809 seconds, Fetched: 1 row(s)
hive> 
```

Create new variables from timestamp(ts): day, month and year


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 94×31

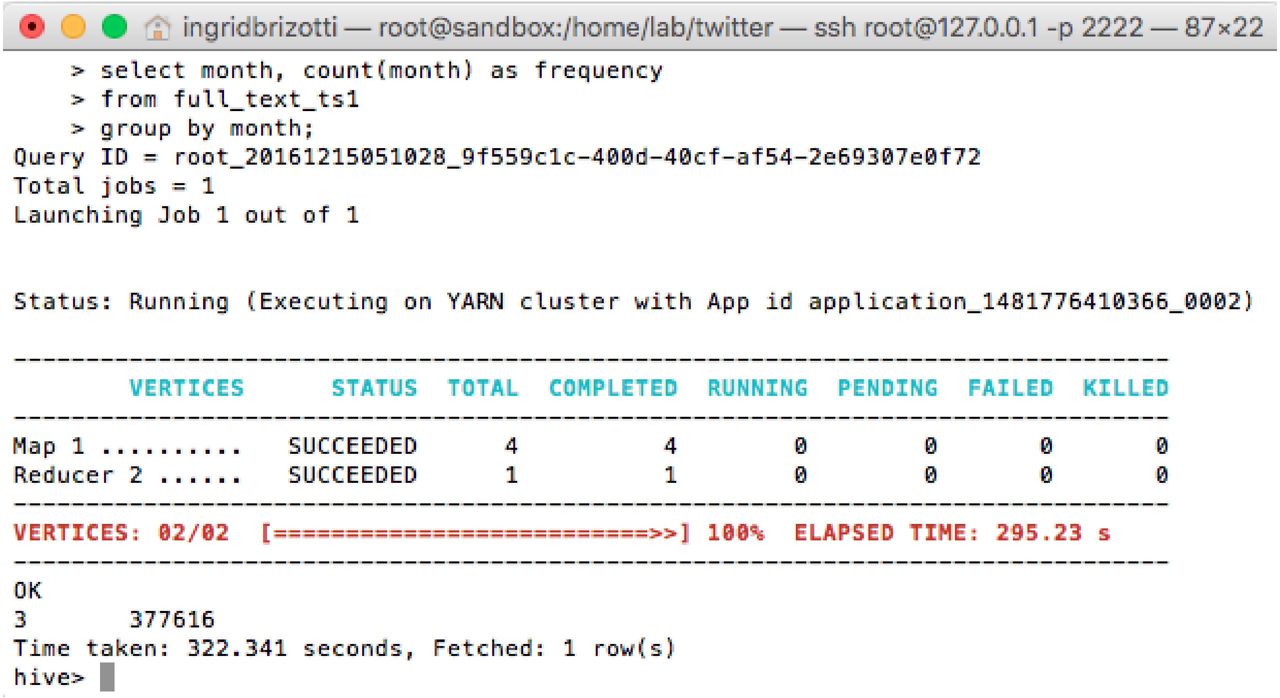
```
hive> create table full_text_ts1 as
    > select id,ts,lat,lon,tweet,
    > unix_timestamp(ts) as unix_timestamp,
    > to_date(ts) as date1,
    > year(ts) as year,
    > month(ts) as month,
    > day(ts) as day
    > from full_text_ts;
Query ID = root_20161215031515_c18ee2f4-6841-485b-b2f5-47f13641774a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481736943243_0011)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 151.50 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/full_text_t
s1
Table twitter.full_text_ts1 stats: [numFiles=4, numRows=377616, totalSize=58979220, rawDataSiz
e=58601604]
OK
Time taken: 198.333 seconds
hive> 
```

Frequency on year:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 90×25

```
hive>
    > select year, count(year) as frequency
    > from full_text_ts1
    > group by year;
Query ID = root_20161215050220_8fd6fd1b-f79e-4870-b0ff-80f84ae5d62b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481776410366_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 182.19 s
--------------------------------------------------------------------------------
OK
2010    377616
Time taken: 240.115 seconds, Fetched: 1 row(s)
hive> 
```

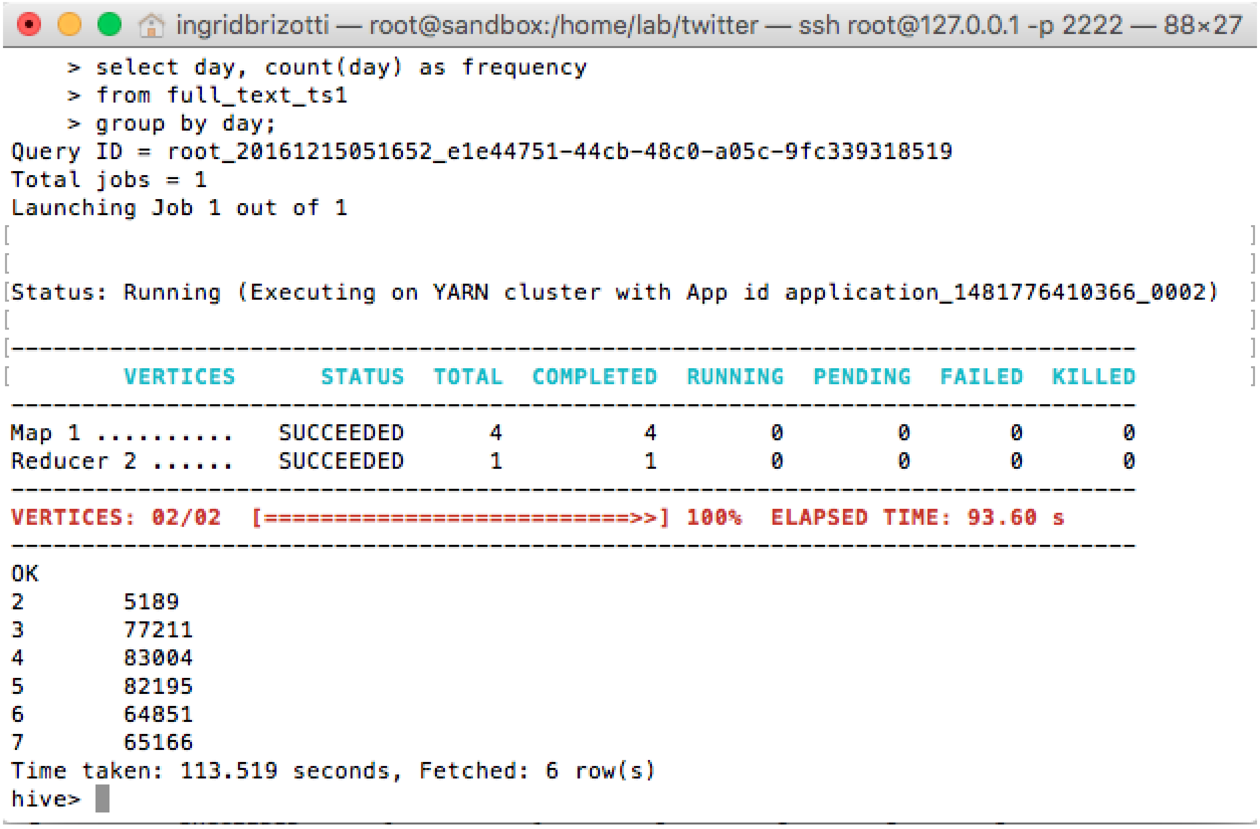
Frequency on month:



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 87×22

```
    > select month, count(month) as frequency
    > from full_text_ts1
    > group by month;
Query ID = root_20161215051028_9f559c1c-400d-40cf-af54-2e69307e0f72
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1481776410366_0002)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 295.23 s
--------------------------------------------------------------------------------
OK
3       377616
Time taken: 322.341 seconds, Fetched: 1 row(s)
hive> 
```

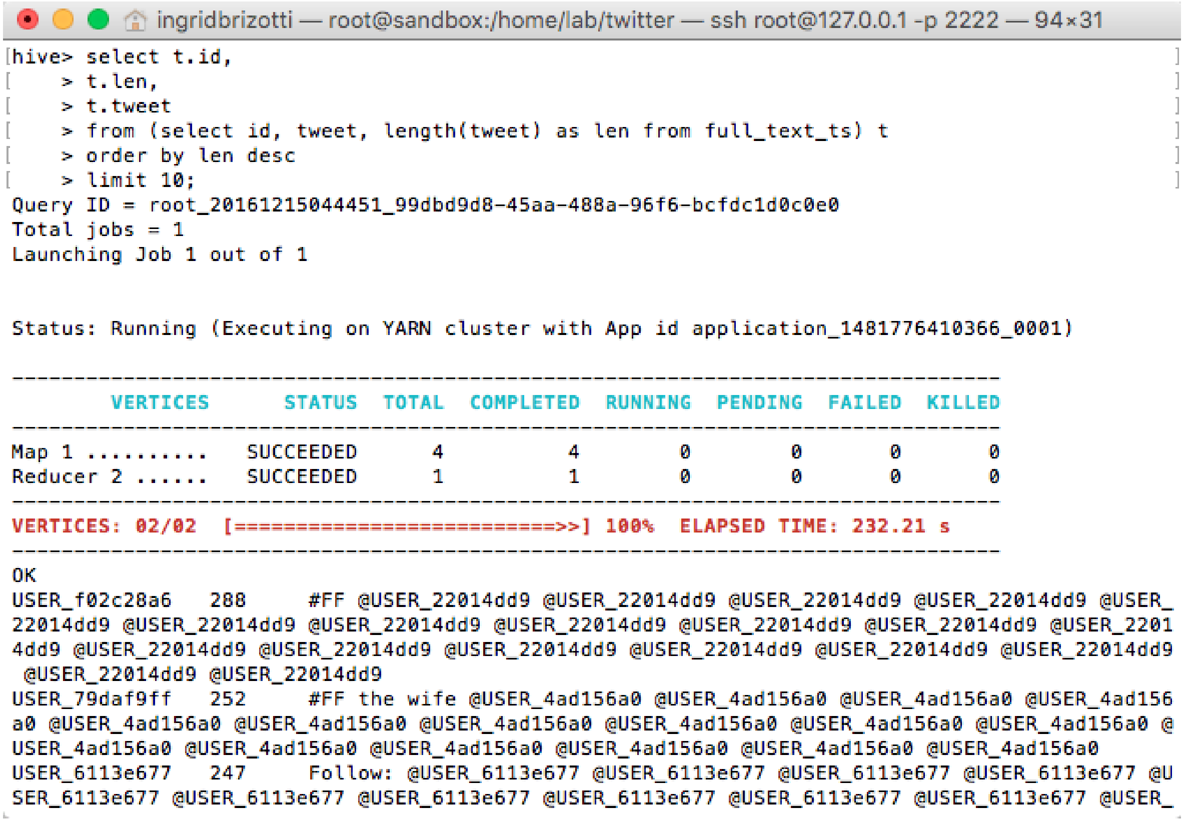
Frequency on day:



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 88×27

```
    > select day, count(day) as frequency
    > from full_text_ts1
    > group by day;
Query ID = root_20161215051652_e1e44751-44cb-48c0-a05c-9fc339318519
Total jobs = 1
Launching Job 1 out of 1
[                                                                                      ]
[                                                                                      ]
[Status: Running (Executing on YARN cluster with App id application_1481776410366_0002) ]
[                                                                                      ]
[--------------------------------------------------------------------------------      ]
[        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED      ]
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 93.60 s
--------------------------------------------------------------------------------
OK
2       5189
3       77211
4       83004
5       82195
6       64851
7       65166
Time taken: 113.519 seconds, Fetched: 6 row(s)
hive> 
```

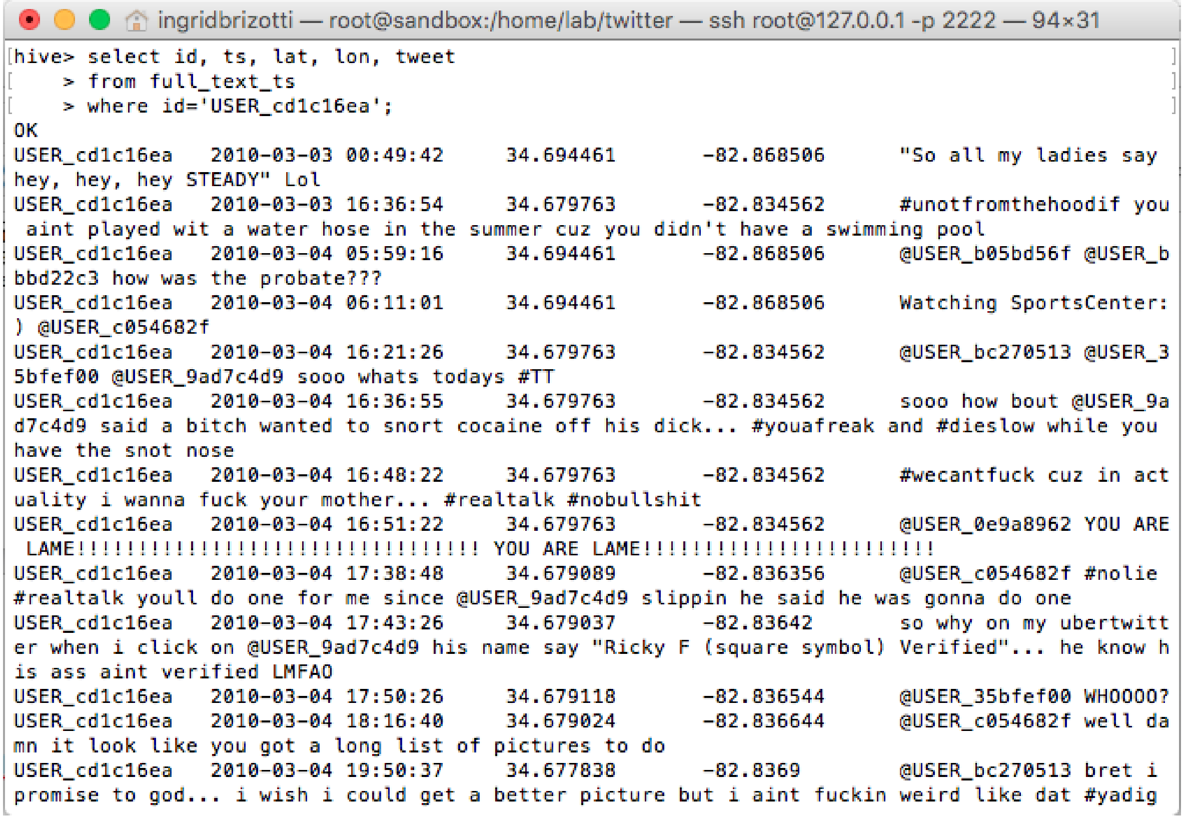
Show top 10 users who tweet long tweets and the length of tweets:



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 94×31

```
[hive> select t.id,
[    > t.len,
[    > t.tweet
[    > from (select id, tweet, length(tweet) as len from full_text_ts) t
[    > order by len desc
[    > limit 10;
Query ID = root_20161215044451_99dbd9d8-45aa-488a-96f6-bcfdc1d0c0e0
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1481776410366_0001)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 232.21 s
--------------------------------------------------------------------------------
OK
USER_f02c28a6   288     #FF @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_
22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_2201
4dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9 @USER_22014dd9
 @USER_22014dd9 @USER_22014dd9
USER_79daf9ff   252     #FF the wife @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156
a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @
USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0 @USER_4ad156a0
USER_6113e677   247     Follow: @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @U
SER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_6113e677 @USER_
```

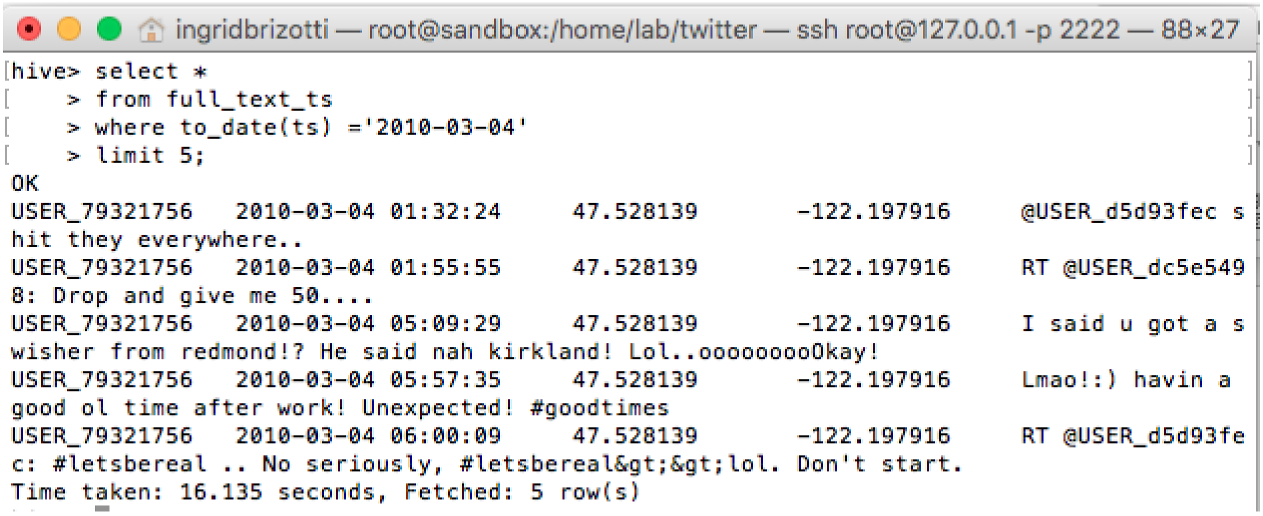
Shows all tweets made by an specific user (USER_cd1c15ea):



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 94×31

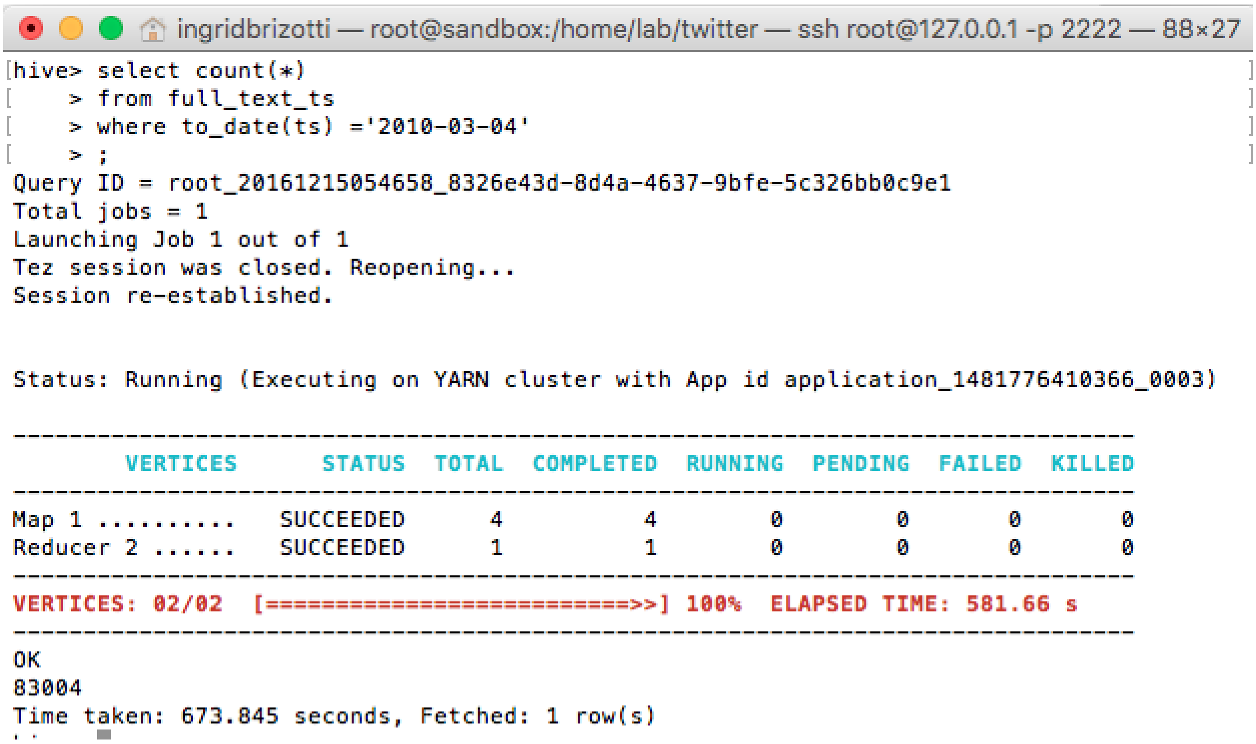
```
[hive> select id, ts, lat, lon, tweet
[    > from full_text_ts
[    > where id='USER_cd1c16ea';
OK
USER_cd1c16ea   2010-03-03 00:49:42     34.694461       -82.868506      "So all my ladies say
hey, hey, hey STEADY" Lol
USER_cd1c16ea   2010-03-03 16:36:54     34.679763       -82.834562      #unotfromthehoodif you
 aint played wit a water hose in the summer cuz you didn't have a swimming pool
USER_cd1c16ea   2010-03-04 05:59:16     34.694461       -82.868506      @USER_b05bd56f @USER_b
bbd22c3 how was the probate???
USER_cd1c16ea   2010-03-04 06:11:01     34.694461       -82.868506      Watching SportsCenter:
) @USER_c054682f
USER_cd1c16ea   2010-03-04 16:21:26     34.679763       -82.834562      @USER_bc270513 @USER_3
5bfef00 @USER_9ad7c4d9 sooo whats todays #TT
USER_cd1c16ea   2010-03-04 16:36:55     34.679763       -82.834562      sooo how bout @USER_9a
d7c4d9 said a bitch wanted to snort cocaine off his dick... #youafreak and #dieslow while you
have the snot nose
USER_cd1c16ea   2010-03-04 16:48:22     34.679763       -82.834562      #wecantfuck cuz in act
uality i wanna fuck your mother... #realtalk #nobullshit
USER_cd1c16ea   2010-03-04 16:51:22     34.679763       -82.834562      @USER_0e9a8962 YOU ARE
 LAME!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! YOU ARE LAME!!!!!!!!!!!!!!!!!!!!!!
USER_cd1c16ea   2010-03-04 17:38:48     34.679089       -82.836356      @USER_c054682f #nolie
#realtalk youll do one for me since @USER_9ad7c4d9 slippin he said he was gonna do one
USER_cd1c16ea   2010-03-04 17:43:26     34.679037       -82.83642       so why on my ubertwitt
er when i click on @USER_9ad7c4d9 his name say "Ricky F (square symbol) Verified"... he know h
is ass aint verified LMFAO
USER_cd1c16ea   2010-03-04 17:50:26     34.679118       -82.836544      @USER_35bfef00 WHOOOO?
USER_cd1c16ea   2010-03-04 18:16:40     34.679024       -82.836644      @USER_c054682f well da
mn it look like you got a long list of pictures to do
USER_cd1c16ea   2010-03-04 19:50:37     34.677838       -82.8369        @USER_bc270513 bret i
promise to god... i wish i could get a better picture but i aint fuckin weird like dat #yadig
```

Bring 5 tweets in an specific date:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 88×27

```
[hive> select *
[    > from full_text_ts
[    > where to_date(ts) ='2010-03-04'
[    > limit 5;
OK
USER_79321756   2010-03-04 01:32:24     47.528139       -122.197916     @USER_d5d93fec s
hit they everywhere..
USER_79321756   2010-03-04 01:55:55     47.528139       -122.197916     RT @USER_dc5e549
8: Drop and give me 50....
USER_79321756   2010-03-04 05:09:29     47.528139       -122.197916     I said u got a s
wisher from redmond!? He said nah kirkland! Lol..ooooooooOkay!
USER_79321756   2010-03-04 05:57:35     47.528139       -122.197916     Lmao!:) havin a
good ol time after work! Unexpected! #goodtimes
USER_79321756   2010-03-04 06:00:09     47.528139       -122.197916     RT @USER_d5d93fe
c: #letsbereal .. No seriously, #letsbereal&gt;&gt;lol. Don't start.
Time taken: 16.135 seconds, Fetched: 5 row(s)
```

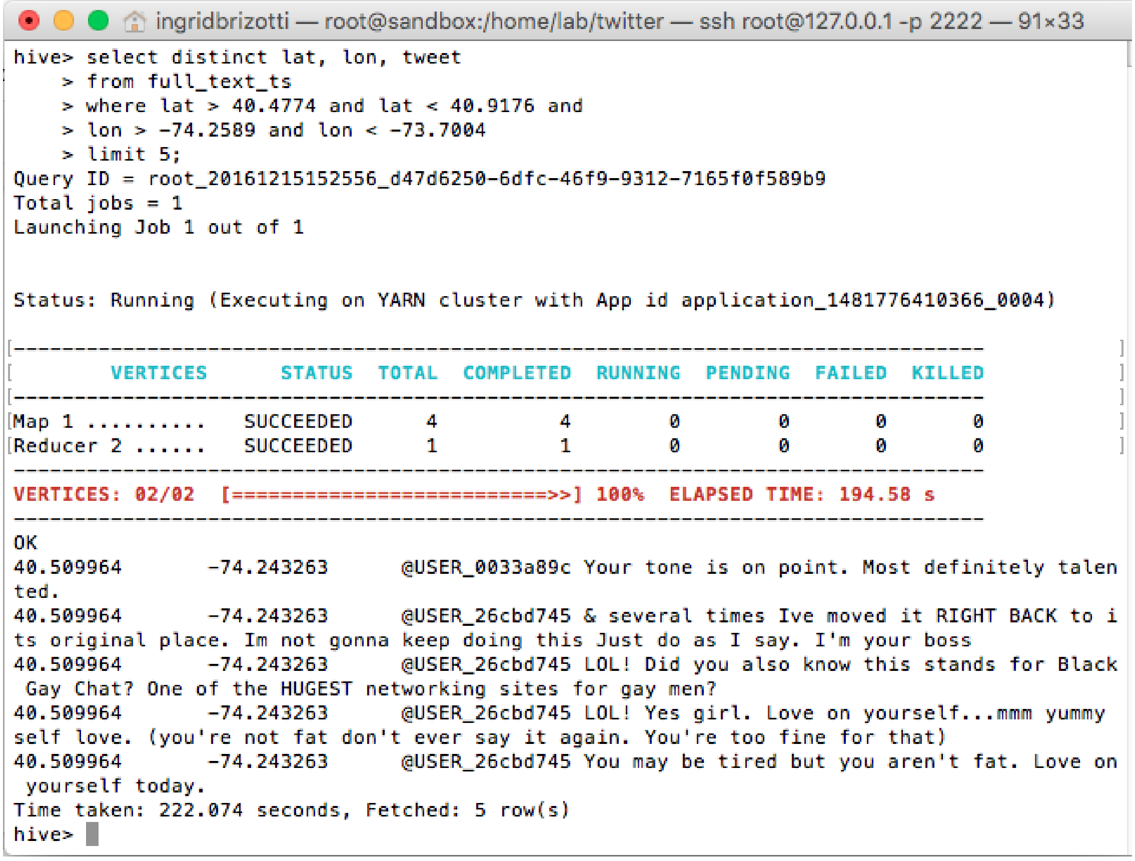
Count the number of tweets in an specific date:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 88×27

```
[hive> select count(*)
[    > from full_text_ts
[    > where to_date(ts) ='2010-03-04'
[    > ;
Query ID = root_20161215054658_8326e43d-8d4a-4637-9bfe-5c326bb0c9e1
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481776410366_0003)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 581.66 s
--------------------------------------------------------------------------------
OK
83004
Time taken: 673.845 seconds, Fetched: 1 row(s)
```

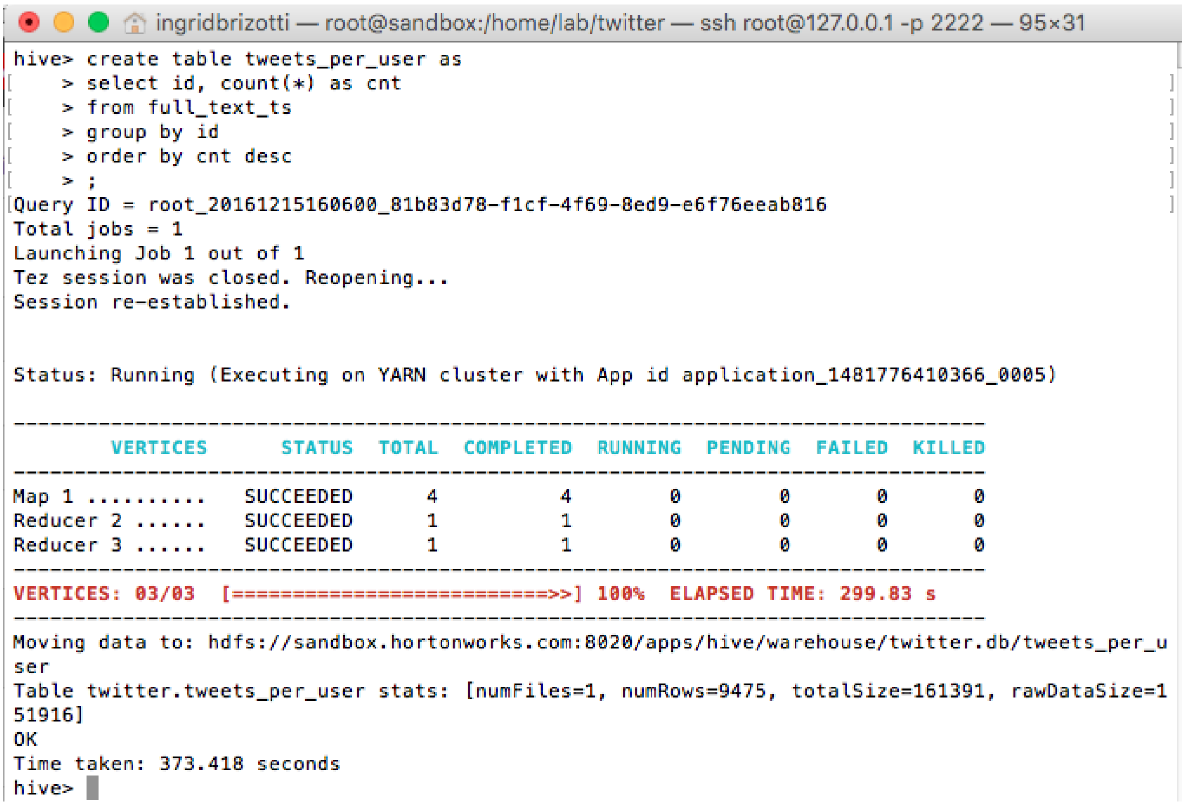
There are 83.004 tweets on 2010-03-04

Show 5 tweets from New York area (using BoundingBox: W 74°25'89"--W 73°70'04"/N 40°47'74"--N 40°91'76") - http://boundingbox.klokantech.com/

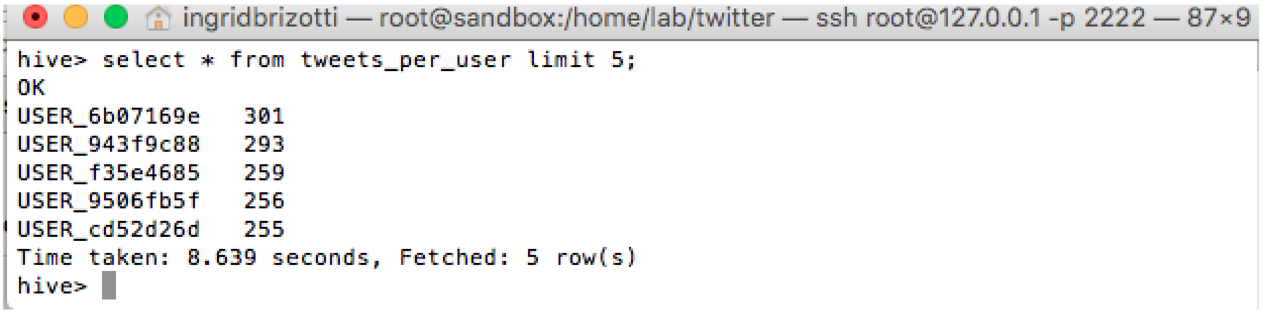

ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 91×33

```
hive> select distinct lat, lon, tweet
    > from full_text_ts
    > where lat > 40.4774 and lat < 40.9176 and
    > lon > -74.2589 and lon < -73.7004
    > limit 5;
Query ID = root_20161215152556_d47d6250-6dfc-46f9-9312-7165f0f589b9
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1481776410366_0004)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 194.58 s
--------------------------------------------------------------------------------
OK
40.509964       -74.243263      @USER_0033a89c Your tone is on point. Most definitely talen
ted.
40.509964       -74.243263      @USER_26cbd745 & several times Ive moved it RIGHT BACK to i
ts original place. Im not gonna keep doing this Just do as I say. I'm your boss
40.509964       -74.243263      @USER_26cbd745 LOL! Did you also know this stands for Black
 Gay Chat? One of the HUGEST networking sites for gay men?
40.509964       -74.243263      @USER_26cbd745 LOL! Yes girl. Love on yourself...mmm yummy
self love. (you're not fat don't ever say it again. You're too fine for that)
40.509964       -74.243263      @USER_26cbd745 You may be tired but you aren't fat. Love on
 yourself today.
Time taken: 222.074 seconds, Fetched: 5 row(s)
hive> 
```

Calculate the number of tweets per user:

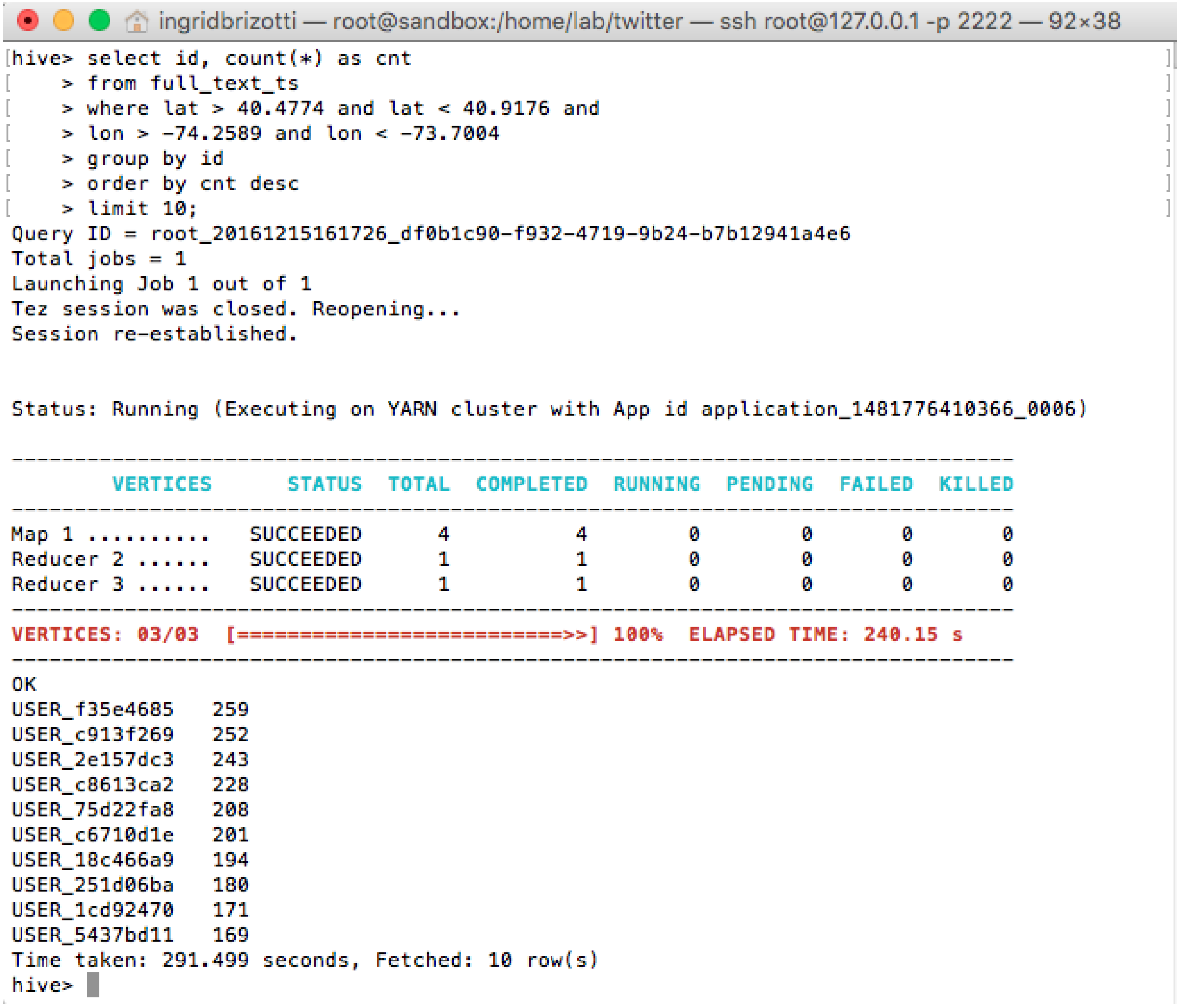

ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 95×31

```
hive> create table tweets_per_user as
    > select id, count(*) as cnt
    > from full_text_ts
    > group by id
    > order by cnt desc
    > ;
Query ID = root_20161215160600_81b83d78-f1cf-4f69-8ed9-e6f76eeab816
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481776410366_0005)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 299.83 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/twitter.db/tweets_per_u
ser
Table twitter.tweets_per_user stats: [numFiles=1, numRows=9475, totalSize=161391, rawDataSize=1
51916]
OK
Time taken: 373.418 seconds
hive> 
```


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 87×9

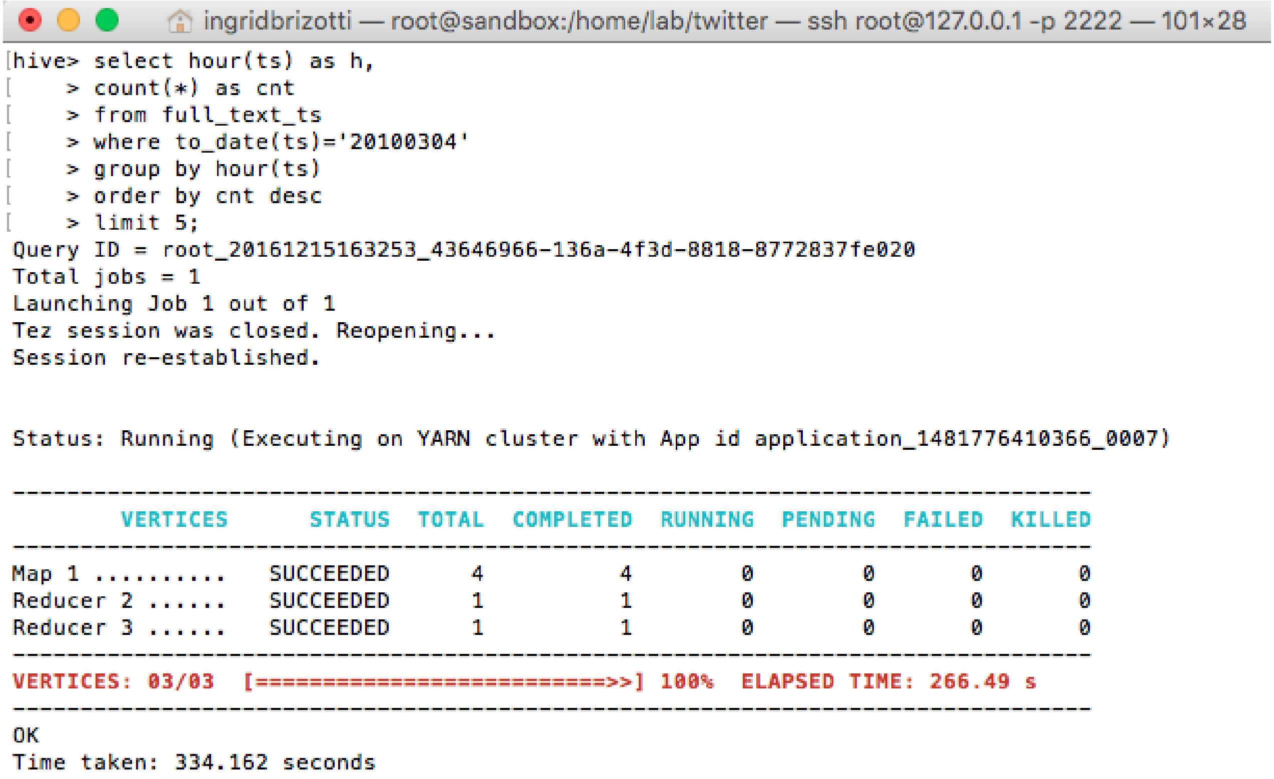
```
hive> select * from tweets_per_user limit 5;
OK
USER_6b07169e   301
USER_943f9c88   293
USER_f35e4685   259
USER_9506fb5f   256
USER_cd52d26d   255
Time taken: 8.639 seconds, Fetched: 5 row(s)
hive> 
```

Find top 10 tweeters (number of tweets/user) in New York area:


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 92×38

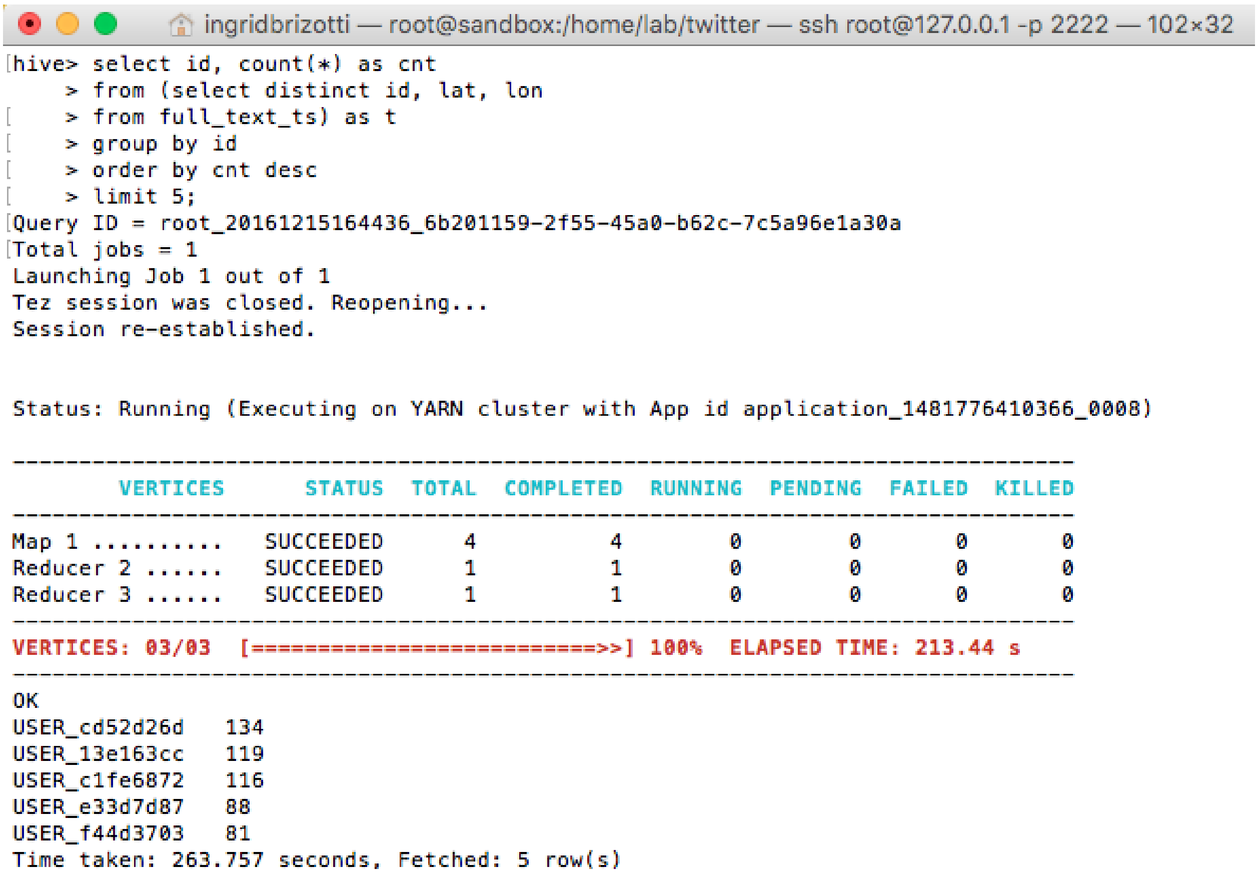
```
hive> select id, count(*) as cnt
    > from full_text_ts
    > where lat > 40.4774 and lat < 40.9176 and
    > lon > -74.2589 and lon < -73.7004
    > group by id
    > order by cnt desc
    > limit 10;
Query ID = root_20161215161726_df0b1c90-f932-4719-9b24-b7b12941a4e6
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481776410366_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 240.15 s
--------------------------------------------------------------------------------
OK
USER_f35e4685   259
USER_c913f269   252
USER_2e157dc3   243
USER_c8613ca2   228
USER_75d22fa8   208
USER_c6710d1e   201
USER_18c466a9   194
USER_251d06ba   180
USER_1cd92470   171
USER_5437bd11   169
Time taken: 291.499 seconds, Fetched: 10 row(s)
hive> 
```

Find hour of the day that generated most number of tweets on March 4, 2010


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 101×28

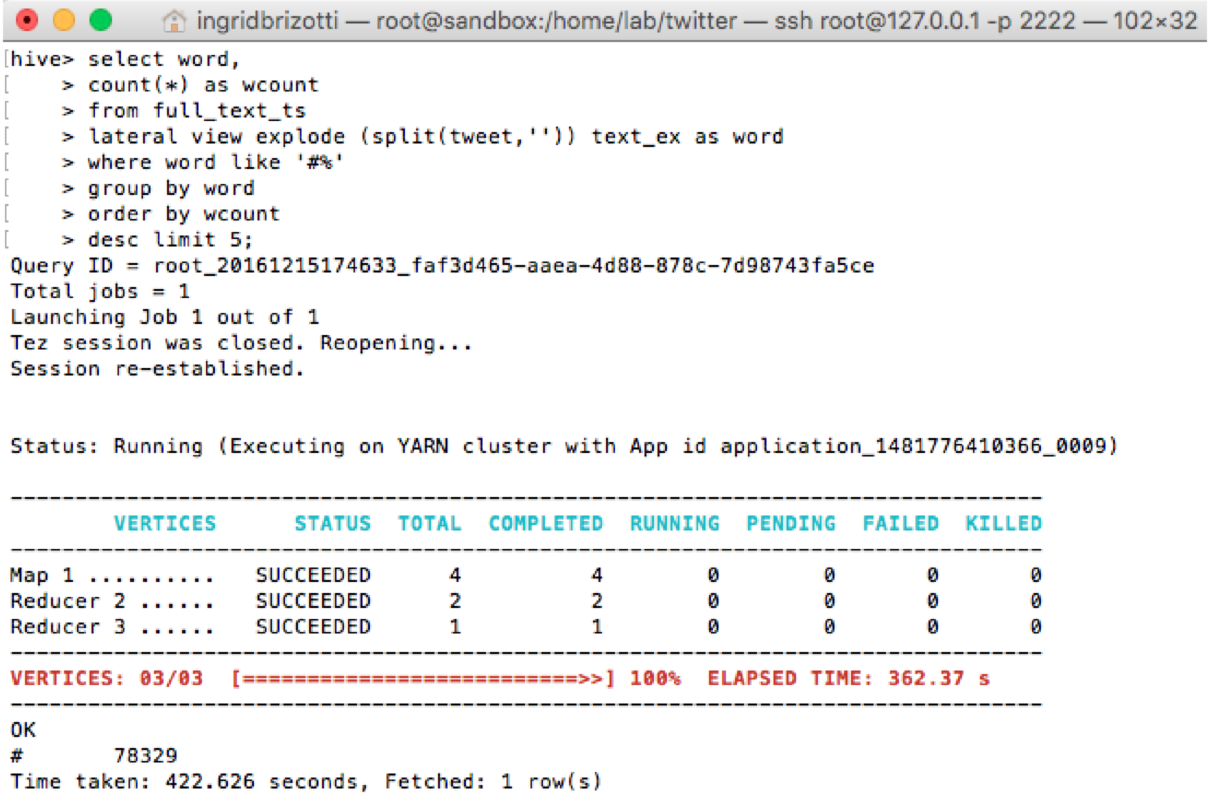
```
[hive> select hour(ts) as h,
[    > count(*) as cnt
[    > from full_text_ts
[    > where to_date(ts)='20100304'
[    > group by hour(ts)
[    > order by cnt desc
[    > limit 5;
 Query ID = root_20161215163253_43646966-136a-4f3d-8818-8772837fe020
 Total jobs = 1
 Launching Job 1 out of 1
 Tez session was closed. Reopening...
 Session re-established.


 Status: Running (Executing on YARN cluster with App id application_1481776410366_0007)

 --------------------------------------------------------------------------------
         VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
 --------------------------------------------------------------------------------
 Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
 Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
 Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
 --------------------------------------------------------------------------------
 VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 266.49 s
 --------------------------------------------------------------------------------
 OK
 Time taken: 334.162 seconds
```

Find the most tweeter that tweeted from most distinct location


ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 102×32

```
[hive> select id, count(*) as cnt
     > from (select distinct id, lat, lon
[    > from full_text_ts) as t
[    > group by id
[    > order by cnt desc
[    > limit 5;
[Query ID = root_20161215164436_6b201159-2f55-45a0-b62c-7c5a96e1a30a
[Total jobs = 1
 Launching Job 1 out of 1
 Tez session was closed. Reopening...
 Session re-established.


 Status: Running (Executing on YARN cluster with App id application_1481776410366_0008)

 --------------------------------------------------------------------------------
         VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
 --------------------------------------------------------------------------------
 Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
 Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
 Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
 --------------------------------------------------------------------------------
 VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 213.44 s
 --------------------------------------------------------------------------------
 OK
 USER_cd52d26d   134
 USER_13e163cc   119
 USER_c1fe6872   116
 USER_e33d7d87   88
 USER_f44d3703   81
 Time taken: 263.757 seconds, Fetched: 5 row(s)
```

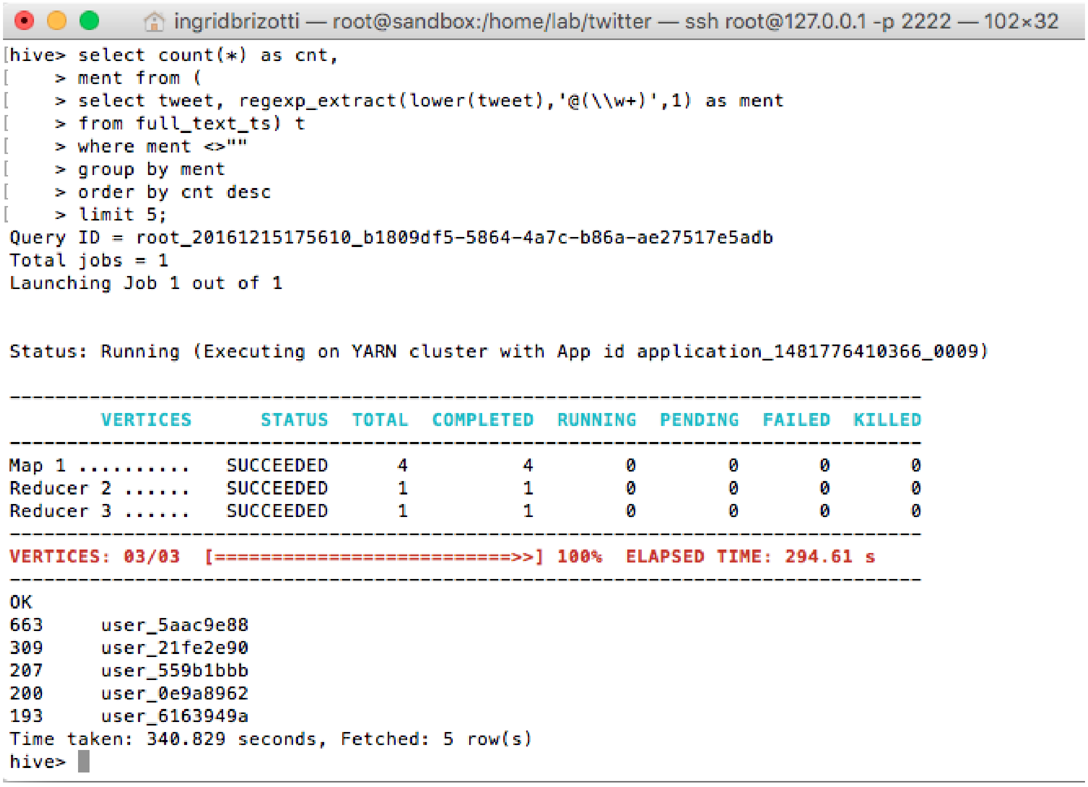
Find 5 most popular topics (hashtags)



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 102×32

```
[hive> select word,
[    > count(*) as wcount
[    > from full_text_ts
[    > lateral view explode (split(tweet,'')) text_ex as word
[    > where word like '#%'
[    > group by word
[    > order by wcount
[    > desc limit 5;
Query ID = root_20161215174633_faf3d465-aaea-4d88-878c-7d98743fa5ce
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1481776410366_0009)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 362.37 s
--------------------------------------------------------------------------------
OK
#       78329
Time taken: 422.626 seconds, Fetched: 1 row(s)
```

Find most frequently mentioned twitter



ingridbrizotti — root@sandbox:/home/lab/twitter — ssh root@127.0.0.1 -p 2222 — 102×32

```
[hive> select count(*) as cnt,
[    > ment from (
[    > select tweet, regexp_extract(lower(tweet),'@(\\w+)',1) as ment
[    > from full_text_ts) t
[    > where ment <>""
[    > group by ment
[    > order by cnt desc
[    > limit 5;
Query ID = root_20161215175610_b1809df5-5864-4a7c-b86a-ae27517e5adb
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1481776410366_0009)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 294.61 s
--------------------------------------------------------------------------------
OK
663     user_5aac9e88
309     user_21fe2e90
207     user_559b1bbb
200     user_0e9a8962
193     user_6163949a
Time taken: 340.829 seconds, Fetched: 5 row(s)
hive> 
```