Ciencia de Datos II

Carmen Jackeline Fernández Cruz
Javier Rosales Bañón
Benjamín Vega Herrera

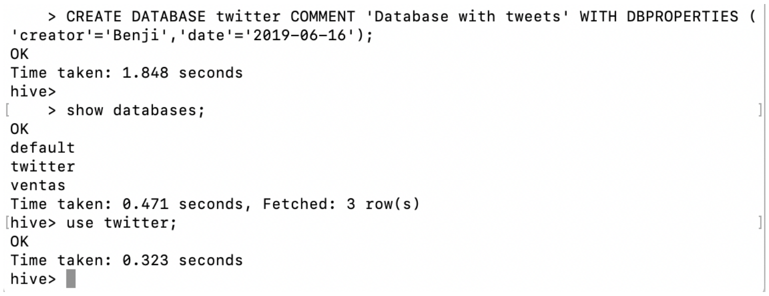
Junio 2019

1 Hive

1.1 Creación de la base de datos

Para crear la base de datos usamos los siguientes comandos:

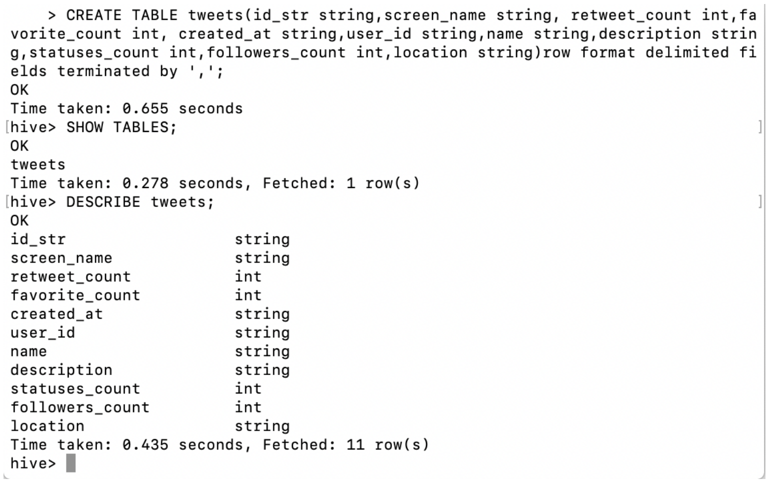
- CREATE DATABASE twitter COMMENT 'Database with tweets' WITH DBPROPERTIES ('creator'='Benji','date'='2019-06-16');



```
    > CREATE DATABASE twitter COMMENT 'Database with tweets' WITH DBPROPERTIES (
'creator'='Benji','date'='2019-06-16');
OK
Time taken: 1.848 seconds
hive>
    > show databases;
OK
default
twitter
ventas
Time taken: 0.471 seconds, Fetched: 3 row(s)
hive> use twitter;
OK
Time taken: 0.323 seconds
hive>
```

Figure 1: Creación de la base de datos.

- USE twitter;
- CREATE TABLE tweets(id_str string,text string,screen_name string, retweet_count int,favorite_count int, created_at string,user_id string,name string,description string,statuses_count int,followers_count int,location string)row format delimited fields terminated by ',';

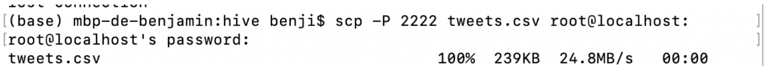


```
    > CREATE TABLE tweets(id_str string,screen_name string, retweet_count int,fa
vorite_count int, created_at string,user_id string,name string,description strin
g,statuses_count int,followers_count int,location string)row format delimited fi
elds terminated by ',';
OK
Time taken: 0.655 seconds
hive> SHOW TABLES;
OK
tweets
Time taken: 0.278 seconds, Fetched: 1 row(s)
hive> DESCRIBE tweets;
OK
id_str                  string
screen_name             string
retweet_count           int
favorite_count          int
created_at              string
user_id                 string
name                    string
description             string
statuses_count          int
followers_count         int
location                string
Time taken: 0.435 seconds, Fetched: 11 row(s)
hive> 
```

Figure 2: Creación de la base de datos.

Una vez creada la base de datos importamos el fichero csv con los datos a la máquina virtual y una vez allí importamos el fichero a los sistemas de ficheros de hadoop.

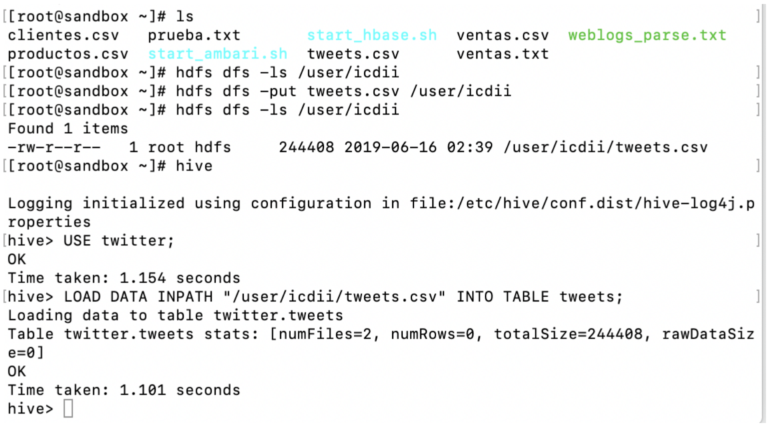
- scp -P 2222 tweets.csv root@localhost:



```
(base) mbp-de-benjamin:hive benji$ scp -P 2222 tweets.csv root@localhost:
root@localhost's password:
tweets.csv                                    100%  239KB  24.8MB/s   00:00
```

Figure 3: Carga del fichero csv a la máquina virtual.

- hdfs dfs -ls /user/icdii
- hdfs dfs -put tweets.csv /user/icdii
- hdfs dfs -ls /user/icdii
- LOAD DATA INPATH ”/user/icdii/tweets.csv” INTO TABLE tweets;
- hdfs dfs -ls /user/icdii



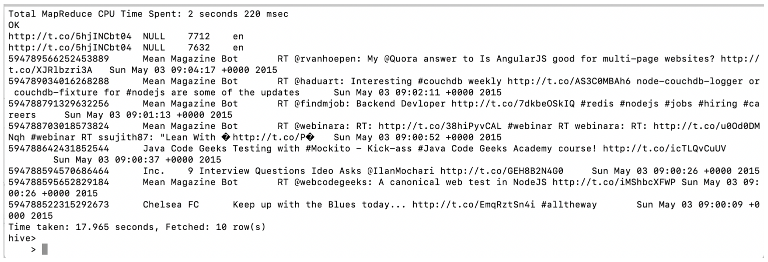
```
[[root@sandbox ~]# ls
clientes.csv   prueba.txt       start_hbase.sh   ventas.csv   weblogs_parse.txt
productos.csv  start_ambari.sh  tweets.csv       ventas.txt
[[root@sandbox ~]# hdfs dfs -ls /user/icdii
[[root@sandbox ~]# hdfs dfs -put tweets.csv /user/icdii
[[root@sandbox ~]# hdfs dfs -ls /user/icdii
Found 1 items
-rw-r--r--   1 root hdfs     244408 2019-06-16 02:39 /user/icdii/tweets.csv
[[root@sandbox ~]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
[hive> USE twitter;
OK
Time taken: 1.154 seconds
[hive> LOAD DATA INPATH "/user/icdii/tweets.csv" INTO TABLE tweets;
Loading data to table twitter.tweets
Table twitter.tweets stats: [numFiles=2, numRows=0, totalSize=244408, rawDataSiz
e=0]
OK
Time taken: 1.101 seconds
hive> 
```

Figure 4: Carga del fichero en Hive.

1.2 Consultas

1.2.1 Los 10 tweets más recientes

SELECT id_str, name, text, created_at FROM tweets where created_at IS NOT NULL ORDER BY created_at DESC LIMIT 10;

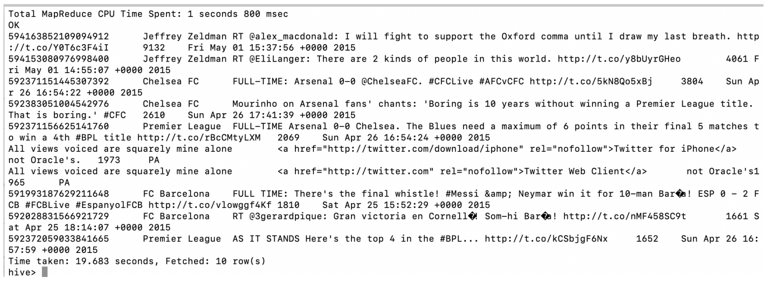


```
Total MapReduce CPU Time Spent: 2 seconds 220 msec
OK
http://t.co/5hjINCbt04  NULL    7712    en
http://t.co/5hjINCbt04  NULL    7632    en
594789566252453889      Mean Magazine Bot       RT @rvanhoepen: My @Quora answer to Is AngularJS good for multi-page websites? http://
t.co/XJRlbzri3A   Sun May 03 09:04:17 +0000 2015
594789034016268288      Mean Magazine Bot       RT @haduart: Interesting #couchdb weekly http://t.co/AS3C0MBAh6 node-couchdb-logger or
 couchdb-fixture for #nodejs are some of the updates      Sun May 03 09:02:11 +0000 2015
594788791329632256      Mean Magazine Bot       RT @findmjob: Backend Devloper http://t.co/7dkbeOSkIQ #redis #nodejs #jobs #hiring #ca
reers     Sun May 03 09:01:13 +0000 2015
594788703018573824      Mean Magazine Bot       RT @webinara: RT: http://t.co/38hiPyvCAL #webinar RT webinara: RT: http://t.co/u0Od0DM
Nqh #webinar RT ssujith87: "Lean With �http://t.co/P�    Sun May 03 09:00:52 +0000 2015
594788642431852544      Java Code Geeks Testing with #Mockito - Kick-ass #Java Code Geeks Academy course! http://t.co/icTLQvCuUV
        Sun May 03 09:00:37 +0000 2015
594788594570686464      Inc.    9 Interview Questions Ideo Asks @IlanMochari http://t.co/GEH8B2N4G0     Sun May 03 09:00:26 +0000 2015
594788595652829184      Mean Magazine Bot       RT @webcodegeeks: A canonical web test in NodeJS http://t.co/iMShbcXFWP Sun May 03 09:
00:26 +0000 2015
594788522315292673      Chelsea FC      Keep up with the Blues today... http://t.co/EmqRztSn4i #alltheway       Sun May 03 09:00:09 +0
000 2015
Time taken: 17.965 seconds, Fetched: 10 row(s)
hive>
    > 
```

Figure 5: Los 10 tweets más recientes.

1.2.2 Los 10 tweets con más retweets

SELECT id_str, name, text, retweet_count, created_at FROM tweets ORDER BY retweet_count DESC LIMIT 10;

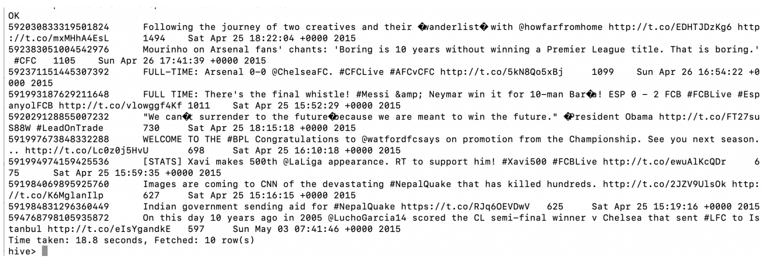


```
Total MapReduce CPU Time Spent: 1 seconds 800 msec
OK
594163852109094912      Jeffrey Zeldman RT @alex_macdonald: I will fight to support the Oxford comma until I draw my last breath. http
://t.co/Y0T6c3F4iI      9132    Fri May 01 15:37:56 +0000 2015
594153080976998400      Jeffrey Zeldman RT @EliLanger: There are 2 kinds of people in this world. http://t.co/y8bUyrGHeo        4061 F
ri May 01 14:55:07 +0000 2015
592371151445307392      Chelsea FC      FULL-TIME: Arsenal 0-0 @ChelseaFC. #CFCLive #AFCvCFC http://t.co/5kN8Qo5xBj     3804    Sun Ap
r 26 16:54:22 +0000 2015
592383051004542976      Chelsea FC      Mourinho on Arsenal fans' chants: 'Boring is 10 years without winning a Premier League title.
That is boring.' #CFC   2610    Sun Apr 26 17:41:39 +0000 2015
592371156625141760      Premier League  FULL-TIME Arsenal 0-0 Chelsea. The Blues need a maximum of 6 points in their final 5 matches t
o win a 4th #BPL title http://t.co/rBcCMtyLXM   2069    Sun Apr 26 16:54:24 +0000 2015
All views voiced are squarely mine alone        <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
not Oracle's.   1973     PA
All views voiced are squarely mine alone        <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>       not Oracle's1
965      PA
591993187629211648      FC Barcelona    FULL TIME: There's the final whistle! #Messi &amp; Neymar win it for 10-man Bar�a! ESP 0 - 2 F
CB #FCBLive #EspanyolFCB http://t.co/vlowggf4Kf 1810    Sat Apr 25 15:52:29 +0000 2015
592028831566921729      FC Barcelona    RT @3gerardpique: Gran victoria en Cornell�! Som-hi Bar�a! http://t.co/nMF458SC9t       1661 S
at Apr 25 18:14:07 +0000 2015
592372059033841665      Premier League  AS IT STANDS Here's the top 4 in the #BPL... http://t.co/kCSbjgF6Nx     1652    Sun Apr 26 16:
57:59 +0000 2015
Time taken: 19.683 seconds, Fetched: 10 row(s)
hive> █
```

Figure 6: Los 10 tweets con más retweets.

1.2.3 Los 10 tweets con más favoritos

SELECT id_str name, text, favorite_count, created_at FROM tweets ORDER BY favorite_count DESC LIMIT 10;

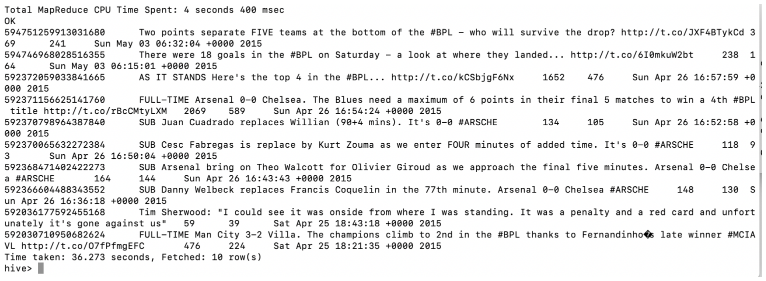


```
OK
592030833319501824      Following the journey of two creatives and their �wanderlist� with @howfarfromhome http://t.co/EDHTJDzKg6 http
://t.co/mxMHhA4EsL      1494    Sat Apr 25 18:22:04 +0000 2015
592383051004542976      Mourinho on Arsenal fans' chants: 'Boring is 10 years without winning a Premier League title. That is boring.'
 #CFC   1105    Sun Apr 26 17:41:39 +0000 2015
592371151445307392      FULL-TIME: Arsenal 0-0 @ChelseaFC. #CFCLive #AFCvCFC http://t.co/5kN8Qo5xBj     1099    Sun Apr 26 16:54:22 +0
000 2015
591993187629211648      FULL TIME: There's the final whistle! #Messi &amp; Neymar win it for 10-man Bar�a! ESP 0 - 2 FCB #FCBLive #Esp
anyolFCB http://t.co/vlowggf4Kf 1011    Sat Apr 25 15:52:29 +0000 2015
592029128855007232      "We can�t surrender to the future�because we are meant to win the future." �President Obama http://t.co/FT27su
S88W #LeadOnTrade       730     Sat Apr 25 18:15:18 +0000 2015
591997673848332288      WELCOME TO THE #BPL Congratulations to @watfordfcsays on promotion from the Championship. See you next season.
.. http://t.co/Lc0z0j5HvU       698     Sat Apr 25 16:10:18 +0000 2015
591994974159425536      [STATS] Xavi makes 500th @LaLiga appearance. RT to support him! #Xavi500 #FCBLive http://t.co/ewuAlKcQDr     6
75      Sat Apr 25 15:59:35 +0000 2015
591984069895925760      Images are coming to CNN of the devastating #NepalQuake that has killed hundreds. http://t.co/2JZV9UlsOk http:
//t.co/K6MglanIlp       627     Sat Apr 25 15:16:15 +0000 2015
591984831296360449      Indian government sending aid for #NepalQuake https://t.co/RJq6OEVDwV   625     Sat Apr 25 15:19:16 +0000 2015
594768798105935872      On this day 10 years ago in 2005 @LuchoGarcia14 scored the CL semi-final winner v Chelsea that sent #LFC to Is
tanbul http://t.co/eIsYgandkE   597     Sun May 03 07:41:46 +0000 2015
Time taken: 18.8 seconds, Fetched: 10 row(s)
hive> █
```

Figure 7: Los 10 tweets con más favoritos.

1.2.4 Los tweets del usuario @premierleague

SELECT id_str, text, retweet_count, favorite_count, created_at FROM tweets WHERE description = 'premierleague' ORDER BY created_at DESC LIMIT 10;

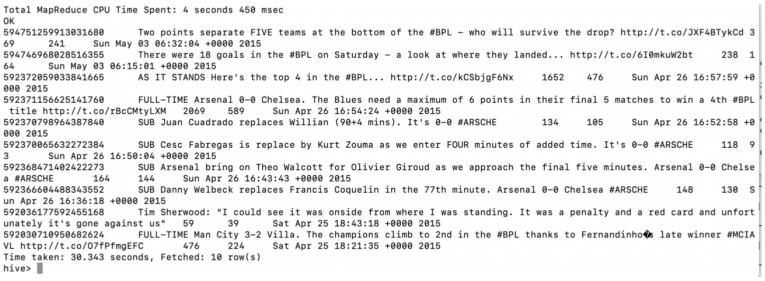


```
Total MapReduce CPU Time Spent: 4 seconds 400 msec
OK
594751259913031680      Two points separate FIVE teams at the bottom of the #BPL - who will survive the drop? http://t.co/JXF4BTykCd 3
69      241     Sun May 03 06:32:04 +0000 2015
594746968028516355      There were 18 goals in the #BPL on Saturday - a look at where they landed... http://t.co/6I0mkuW2bt     238  1
64      Sun May 03 06:15:01 +0000 2015
592372059033841665      AS IT STANDS Here's the top 4 in the #BPL... http://t.co/kCSbjgF6Nx     1652    476     Sun Apr 26 16:57:59 +0
000 2015
592371156625141760      FULL-TIME Arsenal 0-0 Chelsea. The Blues need a maximum of 6 points in their final 5 matches to win a 4th #BPL
 title http://t.co/rBcCMtyLXM   2069    589     Sun Apr 26 16:54:24 +0000 2015
592370798964387840      SUB Juan Cuadrado replaces Willian (90+4 mins). It's 0-0 #ARSCHE        134     105     Sun Apr 26 16:52:58 +0
000 2015
592370065632272384      SUB Cesc Fabregas is replace by Kurt Zouma as we enter FOUR minutes of added time. It's 0-0 #ARSCHE     118  9
3       Sun Apr 26 16:50:04 +0000 2015
592368471402422273      SUB Arsenal bring on Theo Walcott for Olivier Giroud as we approach the final five minutes. Arsenal 0-0 Chelse
a #ARSCHE       164     144     Sun Apr 26 16:43:43 +0000 2015
592366604488343552      SUB Danny Welbeck replaces Francis Coquelin in the 77th minute. Arsenal 0-0 Chelsea #ARSCHE     148     130  S
un Apr 26 16:36:18 +0000 2015
592036177592455168      Tim Sherwood: "I could see it was onside from where I was standing. It was a penalty and a red card and unfort
unately it's gone against us"   59      39      Sat Apr 25 18:43:18 +0000 2015
592030710950682624      FULL-TIME Man City 3-2 Villa. The champions climb to 2nd in the #BPL thanks to Fernandinho�s late winner #MCIA
VL http://t.co/O7fPfmgEFC       476     224     Sat Apr 25 18:21:35 +0000 2015
Time taken: 36.273 seconds, Fetched: 10 row(s)
hive> █
```

Figure 8: Los tweets del usuario @premierleague.

1.2.5 Los 10 usuarios con más seguidores

SELECT id, username, name, description, statuses_count, followers_count FROM tweets ORDER BY followers_count DESC LIMIT 10;



```
Total MapReduce CPU Time Spent: 4 seconds 450 msec
OK
594751259913031680      Two points separate FIVE teams at the bottom of the #BPL – who will survive the drop? http://t.co/JXF4BTykCd 3
69      241     Sun May 03 06:32:04 +0000 2015
594746968028516355      There were 18 goals in the #BPL on Saturday – a look at where they landed... http://t.co/6I0mkuW2bt     238  1
64      Sun May 03 06:15:01 +0000 2015
592372059033841665      AS IT STANDS Here's the top 4 in the #BPL... http://t.co/kCSbjgF6Nx     1652    476     Sun Apr 26 16:57:59 +0
000 2015
592371156625141760      FULL-TIME Arsenal 0-0 Chelsea. The Blues need a maximum of 6 points in their final 5 matches to win a 4th #BPL
 title http://t.co/rBcCMtyLXM   2069    589     Sun Apr 26 16:54:24 +0000 2015
592370798964387840      SUB Juan Cuadrado replaces Willian (90+4 mins). It's 0-0 #ARSCHE        134     105     Sun Apr 26 16:52:58 +0
000 2015
592370065632272384      SUB Cesc Fabregas is replace by Kurt Zouma as we enter FOUR minutes of added time. It's 0-0 #ARSCHE     118  9
3       Sun Apr 26 16:50:04 +0000 2015
592368471402422273      SUB Arsenal bring on Theo Walcott for Olivier Giroud as we approach the final five minutes. Arsenal 0-0 Chelse
a #ARSCHE       164     144     Sun Apr 26 16:43:43 +0000 2015
592366604488343552      SUB Danny Welbeck replaces Francis Coquelin in the 77th minute. Arsenal 0-0 Chelsea #ARSCHE     148     130  S
un Apr 26 16:36:18 +0000 2015
592036177592455168      Tim Sherwood: "I could see it was onside from where I was standing. It was a penalty and a red card and unfort
unately it's gone against us"   59      39      Sat Apr 25 18:43:18 +0000 2015
592030710950682624      FULL-TIME Man City 3-2 Villa. The champions climb to 2nd in the #BPL thanks to Fernandinho�s late winner #MCIA
VL http://t.co/O7fPfmgEFC       476     224     Sat Apr 25 18:21:35 +0000 2015
Time taken: 30.343 seconds, Fetched: 10 row(s)
hive>
```

Figure 9: Los 10 usuarios con más seguidores.