Sqoop-Hadoop

Big Data Aplicado

23/07/16 – I.E.S Fernando Wirtz

Alejandro Regueiro Ruiz

Fecha	Motivo del cambio
	Versión inicial

Índice

CREAR INSTANCIA EN EL CESGA.

Para crear una maquina dentro del cloud del cesga lo que tenemos que hacer es

en el menú ir a la sección de instancias dentro de computación, lanzamos la instancia y la nombramos, ponemos una pequeña descripción si queremos, seleccionamos la versión del sistema operativo que queremos usar, seleccionamos cuantos nucleos y ram queremos en la instancia, en grupos de seguridad añadimos el grupo de openssh y asignamos la clave ssh del dispositivo con el que nos vamos a conectar a la instancia del cesga.

BIGDATA-APLICADO-SQOOP-HADOOP

1. Descargo Docker, otorgo permisos a mi usuario sobre docker y descargo e inicializo los dockers de MariaDB y PostgresSQL

-curl -fsSL https://get.docker.com -o get-docker.sh

sudo sh ./get-docker.sh

-sudo usermod -a -G docker $USER

-sudo su - $USER

-docker volume create datosmariadb

-docker run -p 9907:3306 --name contedor_mariadb -v datosmariadb:/var/lib/mysql --env MARIADB_RANDOM_ROOT_PASSWORD=1 --env MARIADB_DATABASE=demaria --env MARIADB_USER=usuariamaria --env MARIADB_PASSWORD=DonaMaria123456 --restart unless-stopped -d mariadb:latest

-docker run --name de-postre-sql -e POSTGRES_PASSWORD=Cl431Ns3gur4 -p 5432:5432 -p 5433:5433 -d postgres

```
cesgaxuser@xuedua095-alejandro-regueiro:~$ docker ps -a
CONTAINER ID   IMAGE            COMMAND                  CREATED       STATUS              PORTS                                                                                  NAMES
373a721fae24   postgres         "docker-entrypoint.s…"   10 days ago   Up About an hour    0.0.0.0:5432-5433->5432-5433/tcp, :::5432-5433->5432-5433/tcp                          de-postre-sql
c99c61fe7ba8   mariadb:latest   "docker-entrypoint.s…"   10 days ago   Up About an hour    0.0.0.0:9906->3306/tcp, [::]:9906->3306/tcp                                            contedor_mariadb
```

2. Importación/Exportación de Datos

2.1

Con los dockers creados voy a importar los datos de las bases de datos de world y employees en MariaDB desde la terminal local donde están alojados los dockers

Instalando el pack del cliente de MySQL en el local con los dockers:

-apt install mysql-client

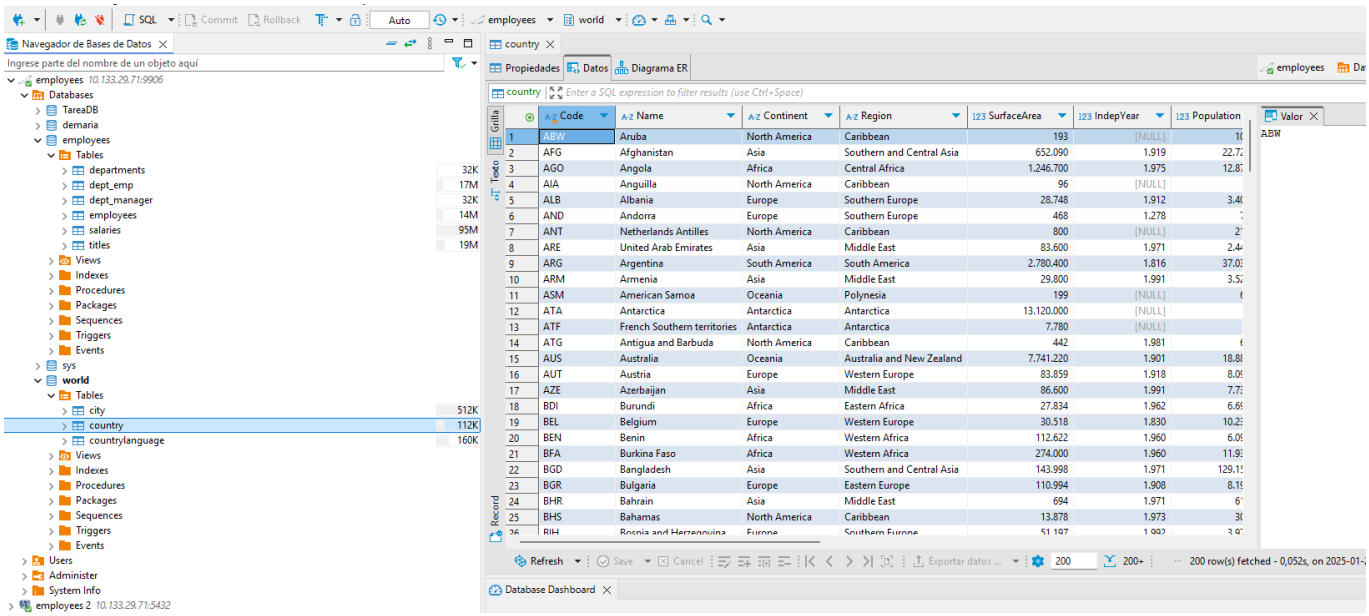
Puedo usar el comando:

mariadb -h(host) -u(user) -p'(contraseña)' < employees.sql

Para importar los datos de la base de datos de empleados(Para poder ejecutar desde el local al dokcer sin necesidad de meterme dentro del docker, tengo que ejecutar el comando dentro de la carpeta en la que reside employees.sql para que tenga acceso al resto de archivos y se ejecute correctamente)

Hacemos los mismo con la base de datos de world:

mariadb -h(host) -u(user) -p'(contraseña)' < world.sql



Como podemos apreciar en la imagen usando dbeaver los datos de employees y world se han importado correctamente

-h(host): Los host usados para importar los datos a los docker son:

MariaDB: 172.17.0.2 por ser el primer docker creado

Postgres: 172.17.0.3

2.2

2.2.1(base de datos world)

Ahora que tenemos los datos importados en MySQL, para importar los datos en postgres vamos a hacerlo mediante ‘SQOOP’

Para conectarse a hadoop usamos el comando:

ssh (usuario)@hadoop.cesga.es

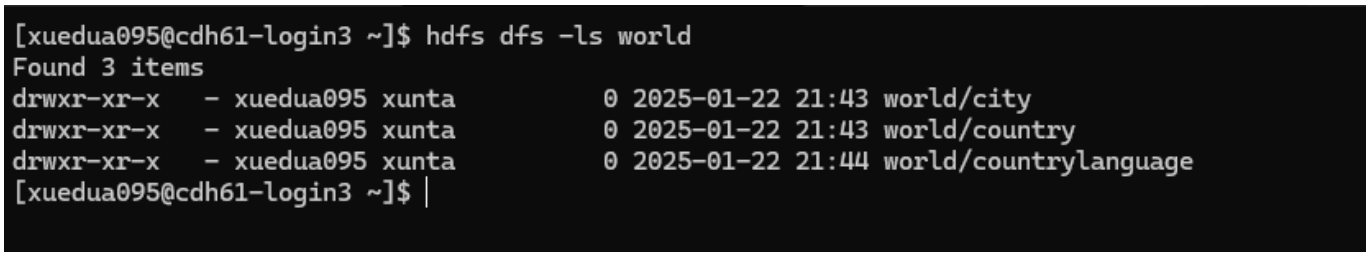
Para saber si hadoop se conecta correctamente con nuestros docker vamos a usar el comando de listar tables y ver si nos devuelve las tablas pertenecientes a la base de datos

-sqoop list-tables --username (usuario) -P --connect jdbc:mysql:///(Ip del docker:Puerto del docker)/(base de datos)

Como podemos apreciar la conexión es correcta ya que nos muestra las tablas que están dentro de la base de datos de ‘world’

Para importar las tablas de MySQL al hadoop tenemos varias opciones, que serían importar todas las tablas de la base de datos a la vez, o importarlas de una en una, yo las voy a importar todas juntas:

-sqoop import-all-tables --connect jdbc:mysql://10.133.29.112:9906/world --username root --password 'iR*-`ySoKm8bq"g[W;mb`:w-rW3DR%$V' --warehouse-dir /user/xuedua095/world --num-mappers 1



Como podemos apreciar las tablas de world se han importando correctamente

Ahora vamos a exportar las tablas de world en Postgres, la ip de conexión de la instancia del cesga puede variar dependiendo de si fue reiniciada.

Antes de exportar las tablas tenemos que asegurarnos de que la base de datos a la que vamos a exportar tenga una estructura que acepte los datos.

La primera tabla que camos a exportar es city

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/world --username postgres --password 1234 --table city --export-dir /user/xuedua095/world/city --input-fields-terminated-by '\000' --num-mappers 1

```
25/01/24 16:52:12 INFO mapreduce.Job: Job job_1708445019134_13710 running in uber mode : false
25/01/24 16:52:12 INFO mapreduce.Job:  map 0% reduce 0%
25/01/24 16:52:17 INFO mapreduce.Job:  map 100% reduce 0%
25/01/24 16:52:17 INFO mapreduce.Job: Job job_1708445019134_13710 completed successfully
25/01/24 16:52:17 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=246873
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=18403
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2742
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2742
                Total vcore-milliseconds taken by all map tasks=2742
                Total megabyte-milliseconds taken by all map tasks=2807808
        Map-Reduce Framework
                Map input records=984
                Map output records=984
                Input split bytes=150
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=41
                CPU time spent (ms)=1530
                Physical memory (bytes) snapshot=342843392
                Virtual memory (bytes) snapshot=2662432768
                Total committed heap usage (bytes)=607649792
                Peak Map Physical memory (bytes)=342843392
                Peak Map Virtual memory (bytes)=2662432768
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
25/01/24 16:52:17 INFO mapreduce.ExportJobBase: Transferred 17.9717 KB in 13.256 seconds (1.3557 KB/sec)
25/01/24 16:52:17 INFO mapreduce.ExportJobBase: Exported 984 records.
```

Parece que exportó correctamente

Ahora vamos a exportar country

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/world --username postgres --password 1234 --table country --export-dir /user/xuedua095/world/country --input-fields-terminated-by '\000' –num-mappers 1

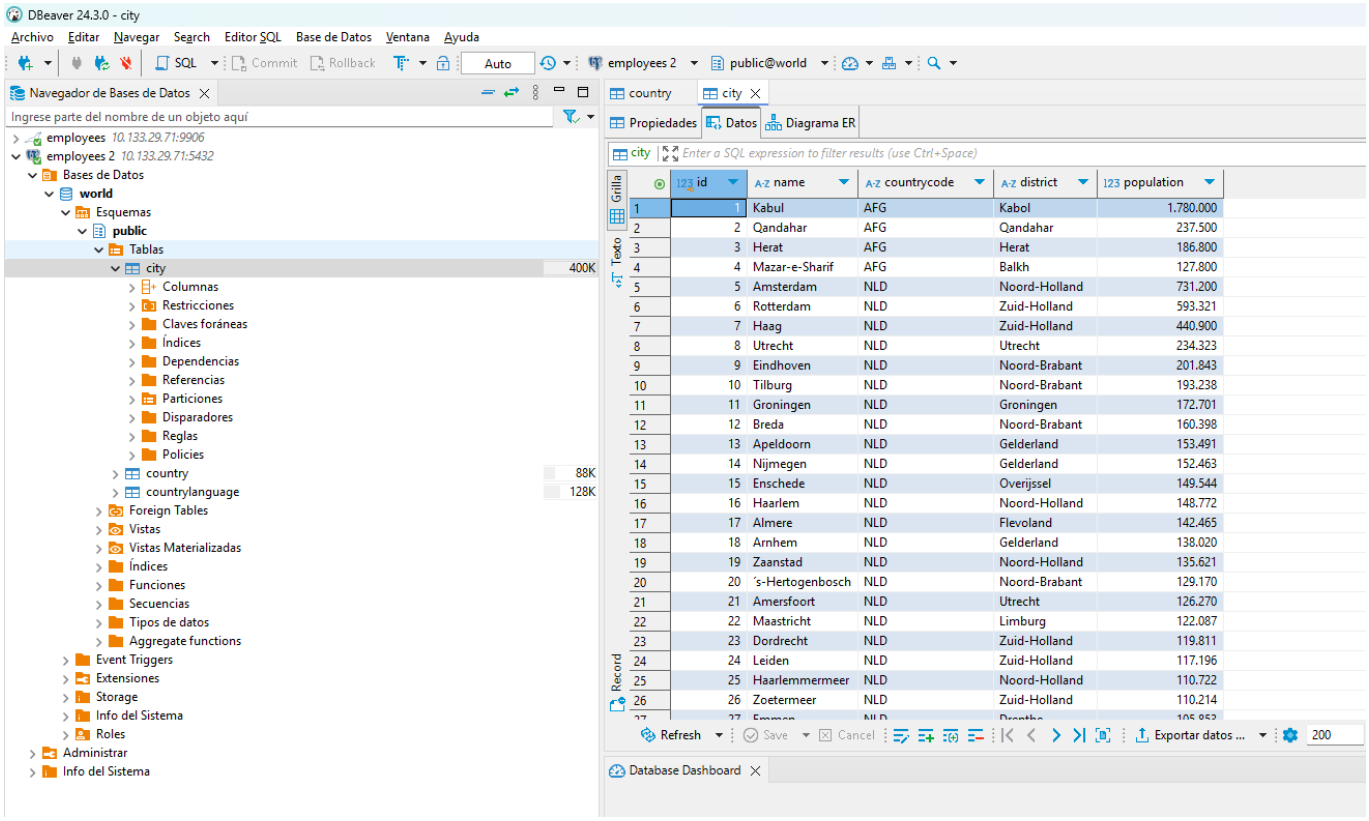
```
25/01/24 17:17:35 INFO mapreduce.Job: Job job_1708445019134_13738 running in uber mode : false
25/01/24 17:17:35 INFO mapreduce.Job:  map 0% reduce 0%
25/01/24 17:17:40 INFO mapreduce.Job:  map 100% reduce 0%
25/01/24 17:17:40 INFO mapreduce.Job: Job job_1708445019134_13738 completed successfully
25/01/24 17:17:40 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=246941
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=32286
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2762
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2762
                Total vcore-milliseconds taken by all map tasks=2762
                Total megabyte-milliseconds taken by all map tasks=2828288
        Map-Reduce Framework
                Map input records=239
                Map output records=239
                Input split bytes=142
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=48
                CPU time spent (ms)=1290
                Physical memory (bytes) snapshot=349691904
                Virtual memory (bytes) snapshot=2662420480
                Total committed heap usage (bytes)=586678272
                Peak Map Physical memory (bytes)=349691904
                Peak Map Virtual memory (bytes)=2662420480
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
25/01/24 17:17:40 INFO mapreduce.ExportJobBase: Transferred 31.5293 KB in 13.2672 seconds (2.3765 KB/sec)
25/01/24 17:17:40 INFO mapreduce.ExportJobBase: Exported 239 records.
```

Y por último vamos a exportar countrylanguage

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/world --username postgres --password 1234 --table countrylanguage --export-dir /user/xuedua095/world/countrylanguage --input-fields-terminated-by '\000' --num-mappers 1

```
25/01/24 16:47:03 INFO mapreduce.Job: Job job_1708445019134_13706 running in uber mode : false
25/01/24 16:47:03 INFO mapreduce.Job:  map 0% reduce 0%
25/01/24 16:47:08 INFO mapreduce.Job:  map 100% reduce 0%
25/01/24 16:47:08 INFO mapreduce.Job: Job job_1708445019134_13706 completed successfully
25/01/24 16:47:08 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=246806
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=144627
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2860
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2860
                Total vcore-milliseconds taken by all map tasks=2860
                Total megabyte-milliseconds taken by all map tasks=2928640
        Map-Reduce Framework
                Map input records=4079
                Map output records=4079
                Input split bytes=139
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=45
                CPU time spent (ms)=1610
                Physical memory (bytes) snapshot=347791360
                Virtual memory (bytes) snapshot=2664488960
                Total committed heap usage (bytes)=590348288
                Peak Map Physical memory (bytes)=347791360
                Peak Map Virtual memory (bytes)=2664488960
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
25/01/24 16:47:08 INFO mapreduce.ExportJobBase: Transferred 141.2373 KB in 13.3509 seconds (10.5789 KB/sec)
25/01/24 16:47:08 INFO mapreduce.ExportJobBase: Exported 4079 records.
```

Parece que todas las exportaciones se han hecho sin problemas, vamos a comprobar con dbeaver si en la conexión de postgres hay tablas con datos en world



Mirando en dbeaver parece que funcionó y los datos se exportaron correctamente

2.2.2(employees)

Importamos las tablas de la base de datos de employees al hadoop.

sqoop import-all-tables --connect jdbc:mysql://10.133.29.71:9906/employees --username root --password 'iR*-`ySoKm8bq"g[W;mb`:w-rW3DR%$V' --warehouse-dir /user/xuedua095/employees --num-mappers 1

```
[xuedua095@cdh61-login3 ~]$ hdfs dfs -ls
Found 5 items
drwx------   - xuedua095 xunta          0 2025-01-23 22:00 .Trash
drwx------   - xuedua095 xunta          0 2025-01-24 18:34 .staging
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:33 employees
-rw-r--r--   3 xuedua095 xunta         12 2025-01-10 17:45 holamundo.txt
drwxr-xr-x   - xuedua095 xunta          0 2025-01-22 21:44 world
^[[A[xuedua095@cdh61-login3 ~]$ hdfs dfs -ls employees
Found 8 items
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:33 employees/current_dept_emp
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:32 employees/departments
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:33 employees/dept_emp
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:32 employees/dept_emp_latest_date
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:32 employees/dept_manager
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:33 employees/employees
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:33 employees/salaries
drwxr-xr-x   - xuedua095 xunta          0 2025-01-24 18:32 employees/titles
[xuedua095@cdh61-login3 ~]$ hdfs dfs -ls employees/salaries
Found 2 items
-rw-r--r--   3 xuedua095 xunta          0 2025-01-24 18:33 employees/salaries/_SUCCESS
-rw-r--r--   3 xuedua095 xunta   98781181 2025-01-24 18:33 employees/salaries/part-m-00000
[xuedua095@cdh61-login3 ~]$
```

Como podemos apreciar, las tablas se importaron correctamente

Vamos a exportar las tablas:

salaries

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/employees --username postgres --password 1234 --table salaries --export-dir /user/xuedua095/employees/salaries --input-fields-terminated-by '\000' –num- mappers 1

employees

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/employees --username postgres --password 1234 --table employees --export-dir /user/xuedua095/employees/employees --input-fields-terminated-by '\000' --num-mappers 1

titles

sqoop export --connect jdbc:postgresql://10.133.29.71:5432/employees --username postgres --password 1234 --table titles --export-dir /user/xuedua095/employees/titles --input-fields-terminated-by '\000' –num-mappers 1

dept_manager

sqoop export --connect jdbc:postgresql://10.133.28.76:5432/employees --username postgres --password 1234 --table dept_manager --export-dir /user/xuedua095/employees/dept_manager --input-fields-terminated-by '\000' --num-mappers 1

dept_emp

sqoop export --connect jdbc:postgresql://10.133.28.76:5432/employees --username postgres --password 1234 --table dept_emp --export-dir /user/xuedua095/employees/dept_emp --input-fields-terminated-by '\000' --num-mappers 1

departments

sqoop export --connect jdbc:postgresql://10.133.28.76:5432/employees --username postgres --password 1234 --table departments --export-dir /user/xuedua095/employees/departments --input-fields-terminated-by '\000' --num-mappers 1

2.2.3(Centros)

Importar centros a mariadb

sqoop import-all-tables --connect jdbc:mysql://10.133.28.76:9906/Centros --username root --password 'iR*-`ySoKm8bq"g[W;mb`:w-rW3DR%$V' --warehouse-dir /user/xuedua095/Centros --num-mappers 1

```
[xuedua095@cdh61-login2 ~]$ hdfs dfs -ls Centros
Found 1 items
drwxr-xr-x   - xuedua095 xunta          0 2025-01-27 20:17 Centros/ProvJuntas
[xuedua095@cdh61-login2 ~]$
```

Exportar centros a postgres

sqoop export --connect jdbc:postgresql://10.133.28.76:5432/centros --username postgres --password 1234 --table provjuntas --export-dir /user/xuedua095/Centros/ProvJuntas --input-fields-terminated-by '\000' –num-mappers 1

3. Un CSV ‘pesado’

Importar csv a hdfs

```
[xuedua095@cdh61-login2 ~]$ hdfs dfs -put /opt/cesga/cursos/pyspark_2022/datasets/NYC_taxi_trip_records/yellow_tripdata_2018-12.csv /user/xuedua095/
```

Exportar a dbeaver

sqoop export --connect jdbc:mysql://10.133.28.76:9906/employees --username root --password 'iR*-`ySoKm8bq"g[W;mb`:w-rW3DR%$V' --table taxi_trips --export-dir /user/xuedua095/yellow_tripdata_2018-12.csv –num-mappers 1

--Checksum del CSV

```
[xuedua095@cdh61-login2 ~]$ ls
ProvJuntas.java  country.java          current_dept_emp.java  dept_emp.java
city.java        countrylanguage.java  departments.java       dept_emp_late
[xuedua095@cdh61-login2 ~]$ md5sum yellow_tripdata_2018-12.csv
93f56ae952cebe9e44dbcdc553884063  yellow_tripdata_2018-12.csv
[xuedua095@cdh61-login2 ~]$
```

--Número de lineas del CSV

```
[xuedua095@cdh61-login2 ~]$ wc -l yellow_tripdata_2018-12.csv
8173233 yellow_tripdata_2018-12.csv
[xuedua095@cdh61-login2 ~]$
```

--Count de la tabla

```
cesgaxuser@xuedua095-alejandro-regueiro:~$ mariadb -h 172.17.0.2 -u root -p -e "SELECT COUNT(*) FROM employees.taxi_trips;"
Enter password:
+----------+
| COUNT(*) |
+----------+
|  8173233 |
+----------+
cesgaxuser@xuedua095-alejandro-regueiro:~$
```