T2.8: Sqoop

David Fernández Reboredo

Big Data Aplicado

PARTE A

Exercicio a realizar nesta parte:

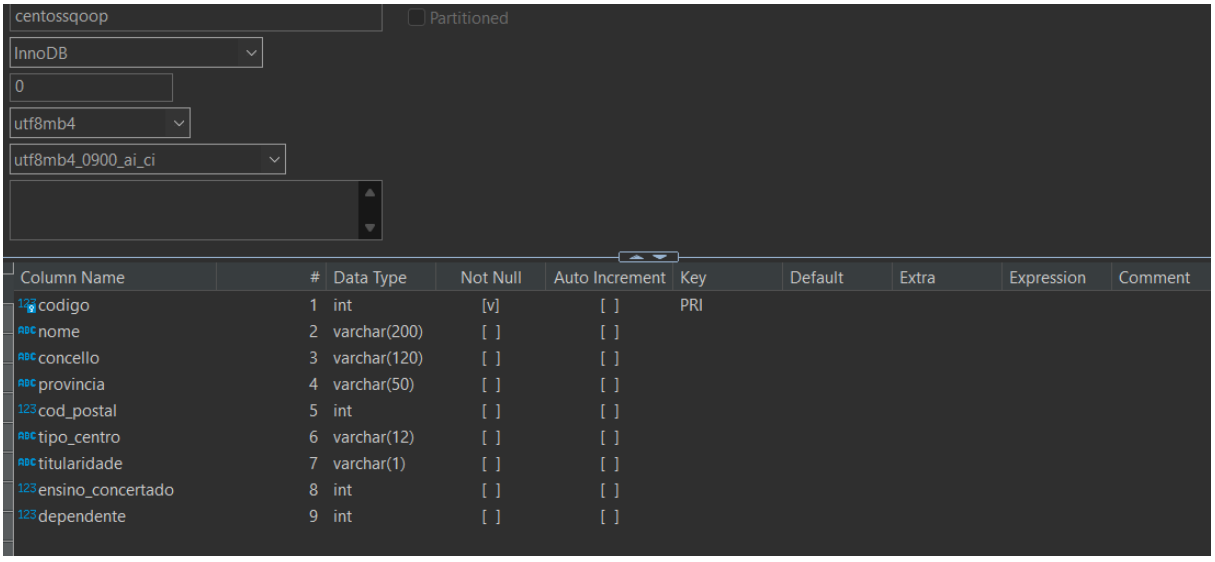
- Fai un import e un export dos centros educativos de Galicia. Indica tódolos pasos (e problemas) que se deron durante a práctica. Indica os comandos en modo texto no documento e a captura co comando e a súa saída.

Para comezar e evitar que sucedan problemáticas vamos a eliminar todos os campos que teñan comas en algún dos seus rexistros:

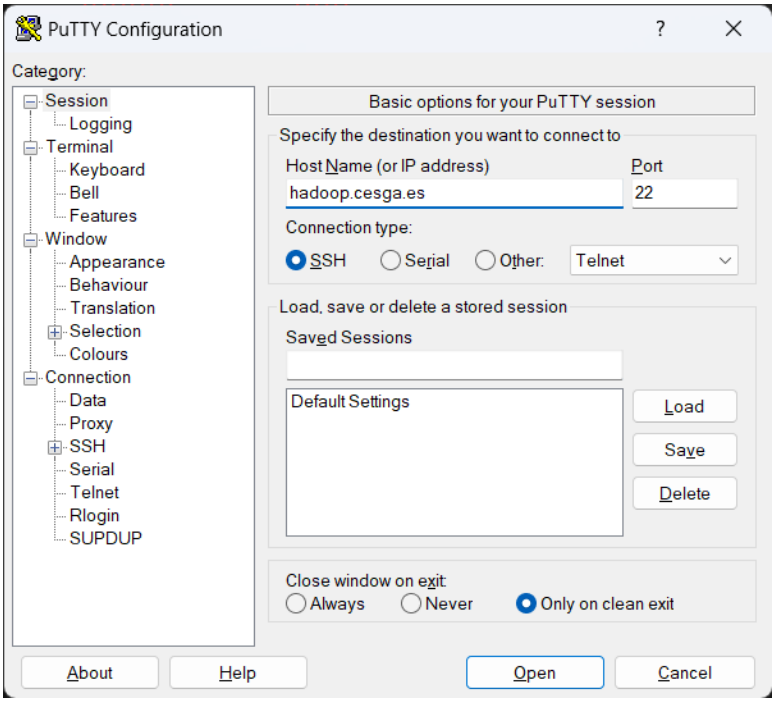
Deixaremos polo tanto os seguintes campos: codigo, nome, concello, provincia, cod_postal, tipo_centro, titularidade, ensino_concertado e dependente.

Debemos coller a practica 2.2 e cambiar a táboa a que se exporta a nosa táboa para o sqoop,

Crearemos a táboa coa información:



Logo disto entramos no Putty e meteremos o noso usuario e contraseña:



Logo disto introduciremos o seguinte código:

```
sqoop import --username xuwira02 --password g5p0h62f4f3whk1  --conn
ect jdbc:mysql://193.144.42.95:9906/xuwira02 --table centossqoop --
target-dir /user/xuwira02/centros_sqoop  --num-mappers 1
```

```
[xuwira02@cdh61-login1 ~]$ sqoop import --username xuwira02 --password g5p0h62f4f3whk1  --con
nect jdbc:mysql://193.144.42.95:9906/xuwira02 --table centossqoop --target-dir /user/xuwira02
/centros_sqoop  --num-mappers 1
Warning: /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/bin/../lib/sqoop/../accumulo do
es not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

```
24/01/14 17:50:34 INFO mapreduce.Job:  map 0% reduce 0%
24/01/14 17:50:38 INFO mapreduce.Job:  map 100% reduce 0%
24/01/14 17:50:38 INFO mapreduce.Job: Job job_1678696618277_12613 completed successfully
24/01/14 17:50:38 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=247161
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=85
                HDFS: Number of bytes written=113368
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2535
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=2535
                Total vcore-milliseconds taken by all map tasks=2535
                Total megabyte-milliseconds taken by all map tasks=2595840
        Map-Reduce Framework
                Map input records=1638
                Map output records=1638
                Input split bytes=85
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=41
                CPU time spent (ms)=1950
                Physical memory (bytes) snapshot=351158272
                Virtual memory (bytes) snapshot=2664894464
                Total committed heap usage (bytes)=600834048
                Peak Map Physical memory (bytes)=351158272
                Peak Map Virtual memory (bytes)=2664894464
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=113368
24/01/14 17:50:38 INFO mapreduce.ImportJobBase: Transferred 110.7109 KB in 25.8611 seconds (4
.281 KB/sec)
24/01/14 17:50:38 INFO mapreduce.ImportJobBase: Retrieved 1638 records.
```

Como podemos ver o proceso foi completado e se realizamos un: hdfs dfs -ls poderemos velo.

```
drwxr-xr-x   - xuwira02 xunta          0 2024-01-14 17:50 centros_sqoop
```

Agora tocará realizar o proceso inverso exportalo, para elo crearemos unha táboa identica en canto aos campos de centrossqoop:

```
CREATE TABLE `centossqoop_a_exportar` (
  `codigo` int NOT NULL,
  `nome` varchar(200) DEFAULT NULL,
  `concello` varchar(120) DEFAULT NULL,
  `provincia` varchar(50) DEFAULT NULL,
  `cod_postal` int DEFAULT NULL,
  `tipo_centro` varchar(12) DEFAULT NULL,
  `titularidade` varchar(1) DEFAULT NULL,
  `ensino_concertado` int DEFAULT NULL,
  `dependente` int DEFAULT NULL
);
```

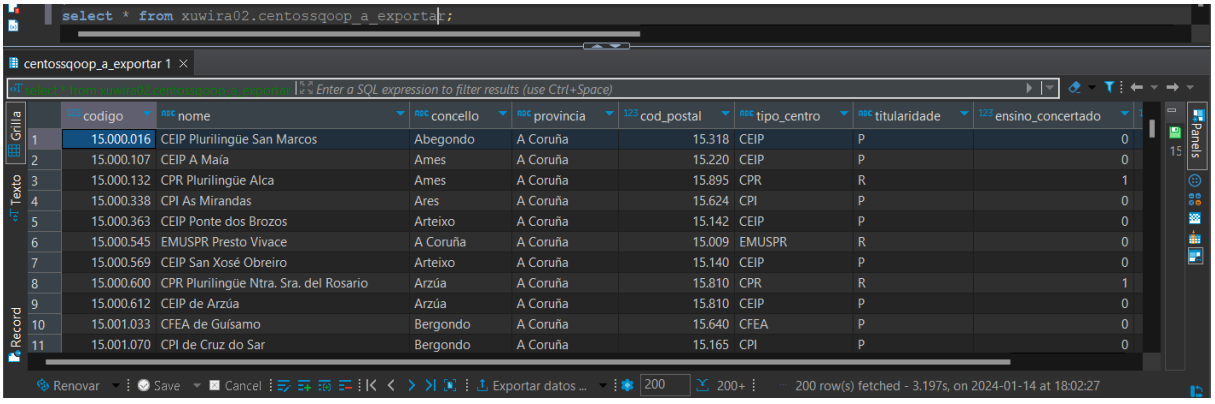
Agora volveremos ao Putty e realizamos o comando de exportación:

```
sqoop export --username xuwira02 --password g5p0h62f4f3whk1 --conne
ct jdbc:mysql://193.144.42.95:9906/xuwira02 --table centossqoop_a_e
xportar --export-dir /user/xuwira02/centros_sqoop --input-fields-te
rminated-by ',' --num-mappers 1
```

```
[xuwira02@cdh61-login1 ~]$ sqoop export --username xuwira02 --password g5p0h62f4f3whk1 --connect jdbc:mysql://193.144.42.95:9906/xuwira02 --table centossqoop_a_exportar --export-dir /user/x
uwira02/centros_sqoop --input-fields-terminated-by ',' --num-mappers 1
Warning: /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
```

```
24/01/14 18:00:11 INFO mapreduce.Job:  map 100% reduce 0%
24/01/14 18:00:11 INFO mapreduce.Job: Job job_1678696618277_12616 completed successfully
24/01/14 18:00:11 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=246891
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=113512
                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)=3093
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=3093
                Total vcore-milliseconds taken by all map tasks=3093
                Total megabyte-milliseconds taken by all map tasks=3167232
        Map-Reduce Framework
                Map input records=1638
                Map output records=1638
                Input split bytes=141
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=46
                CPU time spent (ms)=1960
                Physical memory (bytes) snapshot=343879680
                Virtual memory (bytes) snapshot=2661838848
                Total committed heap usage (bytes)=591921152
                Peak Map Physical memory (bytes)=343879680
                Peak Map Virtual memory (bytes)=2661838848
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
24/01/14 18:00:11 INFO mapreduce.ExportJobBase: Transferred 110.8516 KB in 26.1708 seconds (4.2357 KB/sec)
24/01/14 18:00:11 INFO mapreduce.ExportJobBase: Exported 1638 records.
```

Finalmente solo queda comprobar que a exportación se realizou correctamente:



select * from xuwira02.centossqoop_a_exportar;

centossqoop_a_exportar 1

	codigo	nome	concello	provincia	cod_postal	tipo_centro	titularidade	ensino_concertado
1	15.000.016	CEIP Plurilingüe San Marcos	Abegondo	A Coruña	15.318	CEIP	P	0
2	15.000.107	CEIP A Maía	Ames	A Coruña	15.220	CEIP	P	0
3	15.000.132	CPR Plurilingüe Alca	Ames	A Coruña	15.895	CPR	R	1
4	15.000.338	CPI As Mirandas	Ares	A Coruña	15.624	CPI	P	0
5	15.000.363	CEIP Ponte dos Brozos	Arteixo	A Coruña	15.142	CEIP	P	0
6	15.000.545	EMUSPR Presto Vivace	A Coruña	A Coruña	15.009	EMUSPR	R	0
7	15.000.569	CEIP San Xosé Obreiro	Arteixo	A Coruña	15.140	CEIP	P	0
8	15.000.600	CPR Plurilingüe Ntra. Sra. del Rosario	Arzúa	A Coruña	15.810	CPR	R	1
9	15.000.612	CEIP de Arzúa	Arzúa	A Coruña	15.810	CEIP	P	0
10	15.001.033	CFEA de Guísamo	Bergondo	A Coruña	15.640	CFEA	P	0
11	15.001.070	CPI de Cruz do Sar	Bergondo	A Coruña	15.165	CPI	P	0

Renovar Save Cancel Exportar datos ... 200 200+ 200 row(s) fetched - 3.197s, on 2024-01-14 at 18:02:27

PARTE B

Exercicios a realizar nesta parte:

- En **hadoop.cesga.es**: copia o arquivo: /opt/cesga/cursos/pyspark_2022/datasets/NYC_taxi_trip_records/yellow_tripdata_2018-12.csv ao HDFS
- Con sqoop exporta os datos ao servidor MySQL dado (está na tarefa 2.2 desta aula virtual).
- Fai un checksum do CSV.
- Mira o número de liñas do CSV e un COUNT(*) da táboa exportada.

Facemos un put ao HDFS co csv chamado yellow_tripdata_2018-12.csv.

```
[xuwira02@cdh61-login5]hdfs dfs -put -f /opt/cesga/cursos/pyspark_2
022/datasets/NYC_taxi_trip_records/yellow_tripdata_2018-12.csv
[xuwira02@cdh61-login5]hdfs dfs -ls
```

```
[xuwira02@cdh61-login5 ~]$ hdfs dfs -put -f /opt/cesga/cursos/pyspark_2022/datasets/NYC_tax
i_trip_records/yellow_tripdata_2018-12.csv
[xuwira02@cdh61-login5 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - xuwira02 xunta          0 2024-01-11 22:00 .Trash
drwxr-xr-x   - xuwira02 xunta          0 2023-11-15 22:07 .sparkStaging
drwx------   - xuwira02 xunta          0 2024-01-11 22:07 .staging
drwxr-xr-x   - xuwira02 xunta          0 2024-01-08 21:34 centossqoop
drwxr-xr-x   - xuwira02 xunta          0 2024-01-11 22:06 centrossqoop_export
drwxr-xr-x   - xuwira02 xunta          0 2024-01-11 21:58 centrossqoop_exportar
drwxr-xr-x   - xuwira02 xunta          0 2023-12-11 20:36 compras
drwxr-xr-x   - xuwira02 xunta          0 2023-12-11 21:30 contaje_libros
drwxr-xr-x   - xuwira02 xunta          0 2023-12-11 21:23 libros
drwxr-xr-x   - xuwira02 xunta          0 2023-12-11 20:47 resumo_compras
-rw-r--r--   3 xuwira02 xunta  721522221 2024-01-12 19:32 yellow_tripdata_2018-12.csv
```

Logo faremos un head -n 5 para saber os campos que formarán a tabla no DBeaver e crearemos a tabla agregando a cada un deles a opción de que sean un VARCHAR(200) para evitar posibles erros ao convertir os datos:

```
[xuwira02@cdh61-login5]head -n 5 /opt/cesga/cursos/pyspark_2022/dat
asets/NYC_taxi_trip_r ecords/yellow_tripdata_2018-12.csv
```

```
[xuwira02@cdh61-login5 ~]$ head -n 5 /opt/cesga/cursos/pyspark_2022/datasets/NYC_taxi_trip_
records/yellow_tripdata_2018-12.csv
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeI
D,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount,extra,mta_tax,tip_a
mount,tolls_amount,improvement_surcharge,total_amount

1,2018-12-01 00:28:22,2018-12-01 00:44:07,2,2.50,1,N,148,234,1,12,0.5,0.5,3.95,0,0.3,17.25
1,2018-12-01 00:52:29,2018-12-01 01:11:37,3,2.30,1,N,170,144,1,13,0.5,0.5,2.85,0,0.3,17.15
2,2018-12-01 00:12:52,2018-12-01 00:36:23,1,.00,1,N,113,193,2,2.5,0.5,0.5,0,0,0.3,3.8
[xuwira02@cdh61-login5 ~]$ VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_co
unt,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare
_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount
-bash: VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,Ra
tecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount,extra,mta_ta
x,tip_amount,tolls_amount,improvement_surcharge,total_amount: command not found
```

```sql
create table yellow_tripdata(
    VendorID varchar(100),
    tpep_pickup_datetime varchar(200),
    tpep_dropoff_datetime varchar(200),
    passenger_count varchar(200),
    trip_distance varchar(200),
    RatecodeID varchar(200),
    store_and_fwd_flag varchar(200),
    PULocationID varchar(200),
    DOLocationID varchar(200),
    payment_type varchar(200),
    fare_amount varchar(200),
    extra varchar(200),
    mta_tax varchar(200),
    tip_amount varchar(200),
    tolls_amount varchar(200),
    improvement_surcharge varchar(200),
    total_amount varchar(200)
)
;
```

Logo diso realizaremos o export:

```
sqoop export --username xuwira02 --password g5p0h62f4f3whk1 --conne
ct jdbc:mysql://193.144.42.95:9906/xuwira02 --table yellow_tripdata
--export-dir /user/xuwira02/yellow_tripdata_2018-12.csv --input-fie
lds-terminated-by ',' --num-mappers 1
```

```
[xuwira02@cdh61-login5 ~]$ sqoop export --username xuwira02 --password g5p0h62f4f3whkl --connect jdbc:mysql://193.144
.42.95:9906/xuwira02 --table yellow_tripdata --export-dir /user/xuwira02/yellow_tripdata_2018-12.csv --input-fields-t
erminated-by ',' --num-mappers 1
Warning: /opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/bin/../lib/sqoop/../accumulo does not exist! Accumulo i
mports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/jars/slf4j-log4j12-1.7.25.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.1.1-1.cdh6.1.1.p0.875250/jars/log4j-slf4j-impl-2.8.2.ja
r!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
```

```
24/01/12 20:22:54 INFO mapreduce.Job: Running job: job_1678696618277_12542
24/01/12 20:23:00 INFO mapreduce.Job: Job job_1678696618277_12542 running in uber mode : false
24/01/12 20:23:00 INFO mapreduce.Job:  map 0% reduce 0%
24/01/12 20:23:40 INFO mapreduce.Job:  map 1% reduce 0%
24/01/12 20:24:35 INFO mapreduce.Job:  map 2% reduce 0%
24/01/12 20:25:17 INFO mapreduce.Job:  map 3% reduce 0%
24/01/12 20:26:11 INFO mapreduce.Job:  map 4% reduce 0%
24/01/12 20:27:41 INFO mapreduce.Job:  map 5% reduce 0%
24/01/12 20:29:00 INFO mapreduce.Job:  map 6% reduce 0%
24/01/12 20:30:19 INFO mapreduce.Job:  map 7% reduce 0%
24/01/12 20:31:07 INFO mapreduce.Job:  map 8% reduce 0%


24/01/12 20:32:07 INFO mapreduce.Job:  map 9% reduce 0%
24/01/12 20:32:55 INFO mapreduce.Job:  map 10% reduce 0%

24/01/12 20:34:02 INFO mapreduce.Job:  map 11% reduce 0%

24/01/12 20:35:02 INFO mapreduce.Job:  map 12% reduce 0%
24/01/12 20:36:14 INFO mapreduce.Job:  map 13% reduce 0%
24/01/12 20:37:20 INFO mapreduce.Job:  map 14% reduce 0%



24/01/12 20:38:20 INFO mapreduce.Job:  map 15% reduce 0%
24/01/12 20:39:27 INFO mapreduce.Job:  map 16% reduce 0%
24/01/12 20:40:39 INFO mapreduce.Job:  map 17% reduce 0%
24/01/12 20:41:40 INFO mapreduce.Job:  map 18% reduce 0%
24/01/12 20:43:04 INFO mapreduce.Job:  map 19% reduce 0%
24/01/12 20:44:34 INFO mapreduce.Job:  map 20% reduce 0%
24/01/12 20:46:11 INFO mapreduce.Job:  map 21% reduce 0%

24/01/12 20:47:29 INFO mapreduce.Job:  map 22% reduce 0%
```

```
24/01/14 17:12:14 INFO mapreduce.Job:  map 100% reduce 0%
```

E xa poderíamos ver os datos importados no DBeaver:

Propiedades | Datos | Diagrama ER | 193.144.42.95 | Databases | xuwira02 | Tables | yellow_trip

yellow_tripdata | Enter a SQL expression to filter results (use Ctrl+Space)

Grilla | Texto

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocati
1	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationI
2	[NULL]	[NULL]	[NULL]	[NULL]	[NULL]	[NULL]	[NULL]	[NULL]
3	1	2018-12-01 00:28:22	2018-12-01 00:44:07	2	2.50	1	N	148
4	1	2018-12-01 00:52:29	2018-12-01 01:11:37	3	2.30	1	N	170
5	2	2018-12-01 00:12:52	2018-12-01 00:36:23	1	.00	1	N	113
6	1	2018-12-01 00:35:08	2018-12-01 00:43:11	1	3.90	1	N	95
7	1	2018-12-01 00:21:54	2018-12-01 01:15:13	1	12.80	1	N	163
8	1	2018-12-01 00:00:38	2018-12-01 00:29:26	1	18.80	1	N	132
9	1	2018-12-01 00:59:39	2018-12-01 01:09:07	1	1.00	1	N	246
10	1	2018-12-01 00:19:19	2018-12-01 00:22:19	1	.30	1	N	161
11	1	2018-12-01 00:41:41	2018-12-01 01:09:02	1	3.30	1	N	43
12	1	2018-12-01 00:16:03	2018-12-01 00:52:42	1	5.70	1	N	161
13	1	2018-12-01 00:56:42	2018-12-01 01:22:35	1	17.30	2	N	132
14	1	2018-12-01 00:19:36	2018-12-01 00:24:58	1	.30	1	N	114
15	1	2018-12-01 00:27:51	2018-12-01 00:33:33	1	.80	1	N	79
16	1	2018-12-01 00:44:02	2018-12-01 01:13:33	2	2.60	1	N	79
17	1	2018-12-01 00:30:12	2018-12-01 00:39:09	4	1.80	1	N	261

- CHECKSUM DO CSV

Para ver os datos en formato hash podemos interpretar o comando checksum ao HDFS co fin de que nos mostre os datos relativos ao yellow_tripdata_2018-12.csv:

```
[xuwira02@cdh61-login3 ~]$ hdfs dfs -checksum yellow_tripdata_2018-12.csv
yellow_tripdata_2018-12.csv     MD5-of-262144MD5-of-512CRC32C   0000020000000000
000400005ade031458f865bfa161c3081183015a
```

- COUNT(*)

Debemos contar o número de líneas facendo un cat para mostralo e a posteriori un word count (wc) para saber o número de líneas que contén:

```
cat /opt/cesga/cursos/pyspark_2022/datasets/NYC_taxi_trip_records/y
ellow_tripdata_2018-12.csv | wc -l
```

```
[xuwira02@cdh61-login3 ~]$ cat /opt/cesga/cursos/pyspark_2022/datasets/NYC_taxi_trip_records/yell
ow_tripdata_2018-12.csv | wc -l
8173233
```

Finalmente faremos un:

```
select count(VendedorID) contado from yellow_tripdata;
```

E así concluir que nos mostra o mesmo resultado:

contado
8.173.233