14_Redimientos formato de Archivos

June 18, 2025

Creado por:

Isabel Maniega

1 Ordenador con el que se ha hecho el experimento

Características:

- Intel® Core™ i9-12900H CPU @ 2.40GHz × 4
- 32 GiB RAM DDR4
- 1T SSD m.2
- Sistema Operativo: Ubuntu
- Empleamos el archivo mencionado en el manual. (DATOS DEL 2018)

Ya podemos intuir que no siendo SSD no se obtengan buenos tiempos.

NOTA IMPORTANTE:

- AMD (up to 2.4 GHz)
- 16 GB DDR3
- 1000 GB HDD con 600 GB libres
- Sistema Operativo: Windows

2 Crear un entorno virtual y usarlo en Jupyter

En el caso de crear un entorno virtual que use una versión concreta de Python y el entorno virtual en Jupyter notebook realizaremos los siguientes pasos:

- Crear el entorno virtual con la versión de python que se quiere: virtualenv venv –python=python3.8
- Activa el entorno
- Instala jupyter: pip install notebook

3 Librerías Vaex y Dask documentación

- Vaex

https://vaex.io/

https://vaex.io/docs/index.html

- Dataset taxis de Nueva York:

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

4 Pandas

[1]:
```
%%time
import pandas as pd
```

```
CPU times: user 636 ms, sys: 1.24 s, total: 1.87 s
Wall time: 265 ms
```

---

[2]:
```
# read_parquet AÑADIDO PARA LA NUEVA VERSIÓN:
```

[3]:
```
# pip install fastparquet
```

[4]:
```
%%time
df_pandas = pd.read_parquet("./files/yellow_tripdata_2018-01.parquet")
df_pandas.head()
```

```
CPU times: user 2.7 s, sys: 1.72 s, total: 4.42 s
Wall time: 557 ms
```

[4]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge	airport_fee
0	1	2018-01-01 00:21:05	2018-01-01 00:24:23	1	0.5	1	N	41	24	2	4.5	0.5	0.5	0.00	0.0	0.3	5.80	NaN	NaN
1	1	2018-01-01 00:44:55	2018-01-01 01:03:05	1	2.7	1	N	239	140	2	14.0	0.5	0.5	0.00	0.0	0.3	15.30	NaN	NaN
2	1	2018-01-01 00:08:26	2018-01-01 00:14:21	2	0.8	1	N	262	141	1	6.0	0.5	0.5	1.00	0.0	0.3	8.30	NaN	NaN
3	1	2018-01-01 00:20:22	2018-01-01 00:52:51	1	10.2	1	N	140	257	2	33.5	0.5	0.5	0.00	0.0	0.3	34.80	NaN	NaN
4	1	2018-01-01 00:09:18	2018-01-01 00:27:06	2	2.5	1	N	246	239	1	12.5	0.5	0.5	2.75	0.0	0.3	16.55	NaN	NaN

```
3                    0.3         34.80                    NaN          NaN
4                    0.3         16.55                    NaN          NaN
```

[5]:
```
# Transformar el dataframe para poder trabajar con él.
df_pandas.to_csv("./files/yellow_tripdata_2.csv")
```

Tiempos:

- Con 1TB HDD sobre Windows: 1 minuto 6 segundos (a veces algo más)
- Con 1TB SSD sobre Linux: 56.6 segundos
- Con 250 SSD sobre linux: 21.6 segundos
- Con 1TB SSD m.2 Linux: 2.64 s

[6]:
```
print('Number of Rows: ',len(df_pandas.index))
print("Number of Columns: " + str(len(df_pandas.axes[1])))
```

```
Number of Rows:  8760687
Number of Columns: 19
```

[7]:
```
%%time
df_pandas.describe()
```

```
CPU times: user 2.44 s, sys: 457 ms, total: 2.9 s
Wall time: 2.9 s
```

[7]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime
count	8.760687e+06	8760687	8760687
mean	1.560978e+00	2018-01-17 05:33:04.130081	2018-01-17 05:48:43.847028
min	1.000000e+00	2001-01-05 11:45:23	2001-01-05 11:52:05
25%	1.000000e+00	2018-01-09 23:36:44.500000	2018-01-09 23:52:34
50%	2.000000e+00	2018-01-17 12:56:08	2018-01-17 13:11:38
75%	2.000000e+00	2018-01-24 20:20:37	2018-01-24 20:34:16
max	2.000000e+00	2018-07-27 04:06:37	2018-07-27 04:46:57
std	4.962678e-01	NaN	NaN

	passenger_count	trip_distance	RatecodeID	PULocationID
count	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06
mean	1.606807e+00	2.804022e+00	1.039545e+00	1.644579e+02
min	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
25%	1.000000e+00	9.100000e-01	1.000000e+00	1.160000e+02
50%	1.000000e+00	1.550000e+00	1.000000e+00	1.620000e+02
75%	2.000000e+00	2.840000e+00	1.000000e+00	2.340000e+02
max	9.000000e+00	1.894838e+05	9.900000e+01	2.650000e+02
std	1.258420e+00	6.412050e+01	4.450619e-01	6.635990e+01

	DOLocationID	payment_type	fare_amount	extra	mta_tax
count	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06	8.760687e+06
mean	1.627270e+02	1.310613e+00	1.224443e+01	3.246882e-01	4.975066e-01

```
min    1.000000e+00  1.000000e+00 -4.500000e+02 -4.469000e+01 -5.000000e-01
25%    1.130000e+02  1.000000e+00  6.000000e+00  0.000000e+00  5.000000e-01
50%    1.620000e+02  1.000000e+00  9.000000e+00  0.000000e+00  5.000000e-01
75%    2.340000e+02  2.000000e+00  1.350000e+01  5.000000e-01  5.000000e-01
max    2.650000e+02  4.000000e+00  8.016000e+03  6.000000e+01  4.549000e+01
std    7.031145e+01  4.817808e-01  1.168321e+01  4.502555e-01  4.333281e-02

         tip_amount  tolls_amount  improvement_surcharge  total_amount  \
count  8.760687e+06  8.760687e+06           8.760687e+06  8.760687e+06
mean   1.818759e+00  3.026157e-01           2.996307e-01  1.549109e+01
min   -8.880000e+01 -1.500000e+01          -3.000000e-01 -4.503000e+02
25%    0.000000e+00  0.000000e+00           3.000000e-01  8.300000e+00
50%    1.360000e+00  0.000000e+00           3.000000e-01  1.130000e+01
75%    2.350000e+00  0.000000e+00           3.000000e-01  1.662000e+01
max    4.417100e+02  9.507000e+02           1.000000e+00  8.016800e+03
std    2.486375e+00  1.738184e+00           1.442748e-02  1.419546e+01

       congestion_surcharge  airport_fee
count                  12.0         12.0
mean                    2.5          0.0
min                     2.5          0.0
25%                     2.5          0.0
50%                     2.5          0.0
75%                     2.5          0.0
max                     2.5          0.0
std                     0.0          0.0
```

Tiempos:

- Con 1TB HDD sobre Windows: 11.7 segundos
- Con 1TB SSD sobre Linux: 7.76 segundos
- Con 250 SSD sobre linux: 4.81 segundos
- Con 1TB SSD m.2 Linux: 2.3 s

[8]:
```
%%time
df_pandas.fare_amount.value_counts()
```

```
CPU times: user 42.8 ms, sys: 3.98 ms, total: 46.7 ms
Wall time: 46.3 ms
```

[8]:
```
fare_amount
6.00       473270
5.50       465207
6.50       461959
7.00       446414
5.00       433292
            …
30.60           1
2409.00         1
```

```
168.88          1
201.50          1
33.96           1
Name: count, Length: 1714, dtype: int64
```

Tiempos:

- Con 1TB HDD sobre Windows: 555 milisegundos
- Con 1TB SSD sobre Linux: 313 milisegundos
- Con 250 SSD sobre linux: 213 milisegundos
- Con 1TB SSD m.2 Linux: 44 ms

[9]:
```
%%time
# casi 9 millones de filas en este caso
len(df_pandas), df_pandas.shape
```

```
CPU times: user 15 µs, sys: 1 µs, total: 16 µs
Wall time: 18.1 µs
```

[9]:
```
(8760687, (8760687, 19))
```

Tiempos:

- Con 1TB HDD sobre Windows: 0 nanosegundos
- Con 1TB SSD sobre Linux: 46.7 microsegundos
- Con 250 SSD sobre linux: 21 microsegundos
- Con 1TB SSD m.2 Linux: 18 microsegundos

[10]:
```
%%time
df_pandas.tail()
```

```
CPU times: user 126 µs, sys: 12 µs, total: 138 µs
Wall time: 134 µs
```

[10]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	\
8760682	1	2018-01-31 23:21:35	2018-01-31 23:34:20	2	
8760683	1	2018-01-31 23:35:51	2018-01-31 23:38:57	1	
8760684	2	2018-01-31 23:28:00	2018-01-31 23:37:09	1	
8760685	2	2018-01-31 23:24:40	2018-01-31 23:25:28	1	
8760686	2	2018-01-31 23:28:16	2018-01-31 23:28:38	1	

	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	\
8760682	2.80	1	N	158	
8760683	0.60	1	N	163	
8760684	2.95	1	N	74	
8760685	0.00	1	N	7	
8760686	0.00	1	N	7	

	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	\
8760682	163	1	12.0	0.5	0.5	2.65	

```
8760683           162             1          4.5    0.5      0.5        1.15
8760684            69             2         10.5    0.5      0.5        0.00
8760685           193             2          0.0    0.0      0.0        0.00
8760686           193             2          0.0    0.0      0.0        0.00

         tolls_amount  improvement_surcharge  total_amount  \
8760682           0.0                    0.3         15.95
8760683           0.0                    0.3          6.95
8760684           0.0                    0.3         11.80
8760685           0.0                    0.0          0.00
8760686           0.0                    0.0          0.00

         congestion_surcharge  airport_fee
8760682                   NaN          NaN
8760683                   NaN          NaN
8760684                   NaN          NaN
8760685                   NaN          NaN
8760686                   NaN          NaN
```

Tiempos:

- Con 1TB HDD sobre Windows: 0 nanosegundos
- Con 1TB SSD sobre Linux: 215 microsegundos
- Con 250 SSD sobre linux: 327 microsegundos
- Con 1TB SSD m.2 Linux: 753 microsegundos

5 VAEX desde un CSV (1ª forma)

[11]:
```
# pip install vaex
```

[12]:
```
%%time
# https://pypi.org/project/vaex/

import vaex
```

```
CPU times: user 270 ms, sys: 19.9 ms, total: 290 ms
Wall time: 299 ms
```

Tiempos:

- Con 1TB HDD sobre Windows: 7.64 segundos (en otro intento fueron 14.7 segundos)
- Con 1TB SSD sobre Linux: 5.14 segundos
- Con 250 SSD sobre linux: 1.85 segundos
- Con 1TB SSD m.2 Linux: 365 milisegundos

[23]:
```
%%time
# Necesito añadir convert=True para que me convierta .csv en .HDF5
# Default chunk_size for converting is 5 million rows,
# which corresponds to around 1Gb memory on an example of NYC Taxi dataset.
```

```
df_vaex = vaex.from_csv("./files/yellow_tripdata_2.csv",
                        convert=True)#, chunk_size = 5_000_000)

df_vaex
```

```
CPU times: user 10.7 s, sys: 3.08 s, total: 13.8 s
Wall time: 12.1 s
```

[23]:
```
#          Unnamed: 0    VendorID    tpep_pickup_datetime
tpep_dropoff_datetime    passenger_count    trip_distance    RatecodeID
store_and_fwd_flag    PULocationID    DOLocationID    payment_type
fare_amount    extra    mta_tax    tip_amount    tolls_amount
improvement_surcharge    total_amount    congestion_surcharge    airport_fee
0          0             1           2018-01-01 00:21:05     2018-01-01 00:24:23
1                  0.5             1             N                       41
24              2               4.5            0.5      0.5        0.0
0.0             0.3                      5.8             nan
nan
1          1             1           2018-01-01 00:44:55     2018-01-01 01:03:05
1                  2.7             1             N                       239
140             2               14.0           0.5      0.5        0.0
0.0             0.3                      15.3            nan
nan
2          2             1           2018-01-01 00:08:26     2018-01-01 00:14:21
2                  0.8             1             N                       262
141             1               6.0            0.5      0.5        1.0
0.0             0.3                      8.3             nan
nan
3          3             1           2018-01-01 00:20:22     2018-01-01 00:52:51
1                  10.2            1             N                       140
257             2               33.5           0.5      0.5        0.0
0.0             0.3                      34.8            nan
nan
4          4             1           2018-01-01 00:09:18     2018-01-01 00:27:06
2                  2.5             1             N                       246
239             1               12.5           0.5      0.5        2.75
0.0             0.3                      16.55           nan
nan
…          …             …           …                       …
…                  …               …              …                       …
…              …               …              …        …         …
…              …                          …              …
…
8,760,682  8760682       1           2018-01-31 23:21:35     2018-01-31 23:34:20
2                  2.8             1             N                       158
163             1               12.0           0.5      0.5        2.65
0.0             0.3                      15.95           nan
```

```
nan
8,760,683  8760683       1           2018-01-31 23:35:51     2018-01-31 23:38:57
1                  0.6             1             N                      163
162             1               4.5            0.5      0.5        1.15
0.0             0.3                     6.95            nan
nan
8,760,684  8760684       2           2018-01-31 23:28:00     2018-01-31 23:37:09
1                  2.95            1             N                      74
69              2               10.5           0.5      0.5        0.0
0.0             0.3                     11.8            nan
nan
8,760,685  8760685       2           2018-01-31 23:24:40     2018-01-31 23:25:28
1                  0.0             1             N                      7
193             2               0.0            0.0      0.0        0.0
0.0             0.0                     0.0             nan
nan
8,760,686  8760686       2           2018-01-31 23:28:16     2018-01-31 23:28:38
1                  0.0             1             N                      7
193             2               0.0            0.0      0.0        0.0
0.0             0.0                     0.0             nan
nan
```

Tiempos:

- Con 1TB HDD sobre Windows: 3 minutos 24 segundos (alguna vez mucho menos)
- Con 1TB SSD sobre Linux: 1 minuto 10 segundos
- Con 250 SSD sobre linux: 1 minuto 8 segundos

6 VAEX desde HDF5

[24]:
```
%%time
# https://pypi.org/project/vaex/
# pip install vaex
import vaex
```

```
CPU times: user 5 µs, sys: 0 ns, total: 5 µs
Wall time: 7.63 µs
```

[26]:
```
%%time
vaex_hdf5 = vaex.open("./files/yellow_tripdata_2.csv.hdf5")
vaex_hdf5
```

```
CPU times: user 5.19 ms, sys: 2.03 ms, total: 7.21 ms
Wall time: 6.9 ms
```

[26]:
```
#          Unnamed: 0    VendorID    tpep_pickup_datetime
tpep_dropoff_datetime    passenger_count    trip_distance    RatecodeID
store_and_fwd_flag    PULocationID    DOLocationID    payment_type
fare_amount    extra    mta_tax    tip_amount    tolls_amount
```

```
improvement_surcharge    total_amount    congestion_surcharge    airport_fee
0          0             1           2018-01-01 00:21:05     2018-01-01 00:24:23
1                  0.5             1             N                       41
24             2               4.5            0.5      0.5        0.0
0.0            0.3                      5.8             nan
nan
1          1             1           2018-01-01 00:44:55     2018-01-01 01:03:05
1                  2.7             1             N                       239
140            2               14.0           0.5      0.5        0.0
0.0            0.3                      15.3            nan
nan
2          2             1           2018-01-01 00:08:26     2018-01-01 00:14:21
2                  0.8             1             N                       262
141            1               6.0            0.5      0.5        1.0
0.0            0.3                      8.3             nan
nan
3          3             1           2018-01-01 00:20:22     2018-01-01 00:52:51
1                  10.2            1             N                       140
257            2               33.5           0.5      0.5        0.0
0.0            0.3                      34.8            nan
nan
4          4             1           2018-01-01 00:09:18     2018-01-01 00:27:06
2                  2.5             1             N                       246
239            1               12.5           0.5      0.5        2.75
0.0            0.3                      16.55           nan
nan
…        …            …         …                       …
…                 …              …           …                    …
…             …             …            …       …        …
…             …                      …             …
…
8,760,682  8760682       1           2018-01-31 23:21:35     2018-01-31 23:34:20
2                  2.8             1             N                       158
163            1               12.0           0.5      0.5        2.65
0.0            0.3                      15.95           nan
nan
8,760,683  8760683       1           2018-01-31 23:35:51     2018-01-31 23:38:57
1                  0.6             1             N                       163
162            1               4.5            0.5      0.5        1.15
0.0            0.3                      6.95            nan
nan
8,760,684  8760684       2           2018-01-31 23:28:00     2018-01-31 23:37:09
1                  2.95            1             N                       74
69             2               10.5           0.5      0.5        0.0
0.0            0.3                      11.8            nan
nan
8,760,685  8760685       2           2018-01-31 23:24:40     2018-01-31 23:25:28
```

```
1                  0.0              1             N                     7
193             2               0.0            0.0      0.0        0.0
0.0             0.0                      0.0             nan
nan
8,760,686  8760686       2           2018-01-31 23:28:16     2018-01-31 23:28:38
1                  0.0              1             N                     7
193             2               0.0            0.0      0.0        0.0
0.0             0.0                      0.0             nan
nan
```

Tiempos:

- Con 1TB HDD sobre Windows: 5.11 segundos (alguna ocasión fueron 57 milisegundos)
- Con 1TB SSD sobre Linux: 40.1 milisegundos
- Con 250 SSD sobre linux: 26.5 milisegundos
- Con 1TB SSD m.2 Linux: 9.53 milisegundos

[27]:
```
%%time
vaex_hdf5.describe()
```

```
CPU times: user 9.59 s, sys: 70 ms, total: 9.66 s
Wall time: 1.02 s
```

[27]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	\
data_type	int64	int64	string	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	4380343.0	1.560978037452999	--	
std	2528992.49887	0.496268	--	
min	0	1	--	
max	8760686	2	--	

	tpep_dropoff_datetime	passenger_count	trip_distance	\
data_type	string	int64	float64	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	--	1.6068074341658365	2.8040221012347795	
std	--	1.25842	64.120494	
min	--	0	0.0	
max	--	9	189483.84	

	RatecodeID	store_and_fwd_flag	PULocationID	\
data_type	int64	string	int64	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	1.0395453004998352	--	164.45790906580729	
std	0.445062	--	66.359897	
min	1	--	1	
max	99	--	265	

```
                DOLocationID        payment_type         fare_amount  \
data_type              int64               int64             float64
count                8760687             8760687             8760687
NA                         0                   0                   0
mean      162.72696718876043  1.3106125124662027  12.244425999924395
std                70.311441            0.481781           11.683206
min                        1                   1              -450.0
max                      265                   4              8016.0

                       extra              mta_tax          tip_amount  \
data_type            float64              float64             float64
count                8760687              8760687             8760687
NA                         0                    0                   0
mean      0.3246882464811264  0.49750661106828725  1.8187593050633293
std                0.450255             0.043333            2.486375
min                  -44.69                 -0.5               -88.8
max                    60.0                45.49              441.71

                tolls_amount improvement_surcharge        total_amount  \
data_type            float64              float64             float64
count                8760687              8760687             8760687
NA                         0                    0                   0
mean      0.3026157183791013   0.29963067965078694  15.491088333535643
std                1.738184             0.014427           14.195456
min                  -15.0                 -0.3              -450.3
max                  950.7                  1.0              8016.8

          congestion_surcharge airport_fee
data_type              float64     float64
count                       12          12
NA                     8760675     8760675
mean                       2.5         0.0
std                        0.0         0.0
min                        2.5         0.0
max                        2.5         0.0
```

Tiempos:

- Con 1TB HDD sobre Windows: 9.56 segundos (a veces bastante más)
- Con 1TB SSD sobre Linux: 4.56 segundos
- Con 250 SSD sobre linux: 4.29 segundos
- Con 1TB SSD m.2 Linux: 8.1 segundos

[28]:
```
%%time
vaex_hdf5.fare_amount.value_counts()
```

```
CPU times: user 235 ms, sys: 0 ns, total: 235 ms
Wall time: 20.9 ms
```

[28]:
```
6.00      473270
5.50      465207
6.50      461959
7.00      446414
5.00      433292
           …
228.09         1
63.70          1
268.50         1
23.70          1
125.25         1
Length: 1714, dtype: int64
```

Tiempos:

- Con 1TB HDD sobre Windows: 1.62 segundos (No ejecuta correctamente alguna vez)
- Con 1TB SSD sobre Linux: 114 milisegundos
- Con 250 SSD sobre linux: 336 milisegundos
- Con 1TB SSD m.2 Linux: 373 milisegundos

7 VAEX (2ª forma)

[29]:
```
%%time
# https://pypi.org/project/vaex/
# pip install vaex
import vaex
```

```
CPU times: user 4 µs, sys: 1 µs, total: 5 µs
Wall time: 6.91 µs
```

Tiempos:

- Con 1TB HDD sobre Windows: 7.28 segundos
- Con 1TB SSD sobre Linux: 19.6 microsegundos
- Con 250 SSD sobre linux: 13.1 microsegundos

[30]:
```
%%time
df_vaex_3 = vaex.from_csv("./files/yellow_tripdata_2.csv")
df_vaex_3
```

```
CPU times: user 7.89 s, sys: 930 ms, total: 8.82 s
Wall time: 8.06 s
```

[30]:
```
#          Unnamed: 0    VendorID    tpep_pickup_datetime
tpep_dropoff_datetime    passenger_count    trip_distance    RatecodeID
store_and_fwd_flag    PULocationID    DOLocationID    payment_type
fare_amount    extra    mta_tax    tip_amount    tolls_amount
improvement_surcharge    total_amount    congestion_surcharge    airport_fee
0          0             1           2018-01-01 00:21:05     2018-01-01 00:24:23
```

```
1                 0.5             1             N                     41
24              2               4.5           0.5      0.5        0.0
0.0             0.3                     5.8            nan
nan
1          1            1          2018-01-01 00:44:55     2018-01-01 01:03:05
1                 2.7             1             N                     239
140             2               14.0          0.5      0.5        0.0
0.0             0.3                     15.3           nan
nan
2          2            1          2018-01-01 00:08:26     2018-01-01 00:14:21
2                 0.8             1             N                     262
141             1               6.0           0.5      0.5        1.0
0.0             0.3                     8.3            nan
nan
3          3            1          2018-01-01 00:20:22     2018-01-01 00:52:51
1                 10.2            1             N                     140
257             2               33.5          0.5      0.5        0.0
0.0             0.3                     34.8           nan
nan
4          4            1          2018-01-01 00:09:18     2018-01-01 00:27:06
2                 2.5             1             N                     246
239             1               12.5          0.5      0.5        2.75
0.0             0.3                     16.55          nan
nan
…         …            …          …                       …
…                 …               …             …                   …
…              …               …             …       …         …
…              …                       …              …
…
8,760,682  8760682       1          2018-01-31 23:21:35     2018-01-31 23:34:20
2                 2.8             1             N                     158
163             1               12.0          0.5      0.5        2.65
0.0             0.3                     15.95          nan
nan
8,760,683  8760683       1          2018-01-31 23:35:51     2018-01-31 23:38:57
1                 0.6             1             N                     163
162             1               4.5           0.5      0.5        1.15
0.0             0.3                     6.95           nan
nan
8,760,684  8760684       2          2018-01-31 23:28:00     2018-01-31 23:37:09
1                 2.95            1             N                     74
69              2               10.5          0.5      0.5        0.0
0.0             0.3                     11.8           nan
nan
8,760,685  8760685       2          2018-01-31 23:24:40     2018-01-31 23:25:28
1                 0.0             1             N                     7
193             2               0.0           0.0      0.0        0.0
```

```
0.0             0.0                      0.0             nan
nan
8,760,686  8760686       2           2018-01-31 23:28:16     2018-01-31 23:28:38
1                  0.0             1             N                      7
193             2               0.0            0.0      0.0        0.0
0.0             0.0                      0.0             nan
nan
```

Tiempos:

- Con 1TB HDD sobre Windows: 1 minuto 23 segundos
- Con 1TB SSD sobre Linux: 54.6 segundos
- Con 250 SSD sobre linux: 47.9 segundos
- Con 1TB SSD m.2 Linux: 7.92 segundos

[31]:
```
%%time

df_vaex_3.describe()
```

```
CPU times: user 9.98 s, sys: 116 ms, total: 10.1 s
Wall time: 1.59 s
```

[31]:

	Unnamed: 0	VendorID	tpep_pickup_datetime	\
data_type	int64	int64	string	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	4380343.0	1.560978037452999	--	
std	2528992.49887	0.496268	--	
min	0	1	--	
max	8760686	2	--	

	tpep_dropoff_datetime	passenger_count	trip_distance	\
data_type	string	int64	float64	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	--	1.6068074341658365	2.8040221012347804	
std	--	1.25842	64.120494	
min	--	0	0.0	
max	--	9	189483.84	

	RatecodeID	store_and_fwd_flag	PULocationID	\
data_type	int64	string	int64	
count	8760687	8760687	8760687	
NA	0	0	0	
mean	1.0395453004998352	--	164.45790906580729	
std	0.445062	--	66.359897	
min	1	--	1	
max	99	--	265	

```
                DOLocationID        payment_type         fare_amount  \
data_type              int64               int64             float64
count                8760687             8760687             8760687
NA                         0                   0                   0
mean      162.72696718876043  1.3106125124662027  12.244425999924394
std                70.311441            0.481781           11.683206
min                        1                   1              -450.0
max                      265                   4              8016.0

                       extra              mta_tax          tip_amount  \
data_type            float64              float64             float64
count                8760687              8760687             8760687
NA                         0                    0                   0
mean      0.3246882464811265  0.49750661106828725  1.8187593050633293
std                 0.450255             0.043333            2.486375
min                   -44.69                 -0.5               -88.8
max                     60.0                45.49              441.71

                tolls_amount improvement_surcharge       total_amount  \
data_type            float64               float64            float64
count                8760687               8760687            8760687
NA                         0                     0                  0
mean      0.3026157183791013     0.299630679650787  15.49108833353564
std                 1.738184              0.014427          14.195456
min                   -15.0                  -0.3              -450.3
max                   950.7                   1.0              8016.8

          congestion_surcharge airport_fee
data_type              float64     float64
count                       12          12
NA                     8760675     8760675
mean                       2.5         0.0
std                        0.0         0.0
min                        2.5         0.0
max                        2.5         0.0
```

[32]:
```
%%time
df_vaex_3.export('yellow_tripdata_2018-01.csv.hdf5')
```

```
CPU times: user 1.8 s, sys: 861 ms, total: 2.66 s
Wall time: 2.44 s
```

Tiempos:

- Con 1TB HDD sobre Windows: 1 minuto 19 segundos (53.8 segundos en otro intento)
- Con 1TB SSD sobre Linux: 9.73 segundos
- Con 250 SSD sobre linux: 7.13 segundos
- Con 1TB SSD m.2 Linux: 1.36 segundos

[33]:
```
%%time
df_vaex_4 = vaex.open('yellow_tripdata_2018-01.csv.hdf5')
df_vaex_4
```

```
CPU times: user 4.99 ms, sys: 1.89 ms, total: 6.88 ms
Wall time: 6.72 ms
```

[33]:
```
#          Unnamed: 0    VendorID    tpep_pickup_datetime
tpep_dropoff_datetime    passenger_count    trip_distance    RatecodeID
store_and_fwd_flag    PULocationID    DOLocationID    payment_type
fare_amount    extra    mta_tax    tip_amount    tolls_amount
improvement_surcharge    total_amount    congestion_surcharge    airport_fee
0          0             1           2018-01-01 00:21:05     2018-01-01 00:24:23
1                  0.5             1             N                      41
24             2               4.5            0.5      0.5        0.0
0.0            0.3                     5.8             nan
nan
1          1             1           2018-01-01 00:44:55     2018-01-01 01:03:05
1                  2.7             1             N                      239
140            2               14.0           0.5      0.5        0.0
0.0            0.3                     15.3            nan
nan
2          2             1           2018-01-01 00:08:26     2018-01-01 00:14:21
2                  0.8             1             N                      262
141            1               6.0            0.5      0.5        1.0
0.0            0.3                     8.3             nan
nan
3          3             1           2018-01-01 00:20:22     2018-01-01 00:52:51
1                  10.2            1             N                      140
257            2               33.5           0.5      0.5        0.0
0.0            0.3                     34.8            nan
nan
4          4             1           2018-01-01 00:09:18     2018-01-01 00:27:06
2                  2.5             1             N                      246
239            1               12.5           0.5      0.5        2.75
0.0            0.3                     16.55           nan
nan
…          …             …           …                       …
…                  …               …              …                       …
…              …               …              …        …         …
…              …                       …              …
…
8,760,682  8760682       1           2018-01-31 23:21:35     2018-01-31 23:34:20
2                  2.8             1             N                      158
163            1               12.0           0.5      0.5        2.65
0.0            0.3                     15.95           nan
nan
8,760,683  8760683       1           2018-01-31 23:35:51     2018-01-31 23:38:57
```

```
1                   0.6              1               N                       163
162             1               4.5             0.5       0.5         1.15
0.0             0.3                      6.95             nan
nan
8,760,684  8760684       2           2018-01-31 23:28:00     2018-01-31 23:37:09
1                   2.95             1               N                       74
69              2               10.5            0.5       0.5         0.0
0.0             0.3                      11.8             nan
nan
8,760,685  8760685       2           2018-01-31 23:24:40     2018-01-31 23:25:28
1                   0.0              1               N                       7
193             2               0.0             0.0       0.0         0.0
0.0             0.0                      0.0              nan
nan
8,760,686  8760686       2           2018-01-31 23:28:16     2018-01-31 23:28:38
1                   0.0              1               N                       7
193             2               0.0             0.0       0.0         0.0
0.0             0.0                      0.0              nan
nan
```

Tiempos:

- Con 1TB HDD sobre Windows: 9.28 segundos (5.83 segundos en otro intento, 84 milisegundos en otro)
- Con 1TB SSD sobre Linux: 61.1 milisegundos
- Con 250 SSD sobre linux: 142 milisegundos
- Con 1TB SSD m.2 Linux: 6.31 milisegundos

CONCLUSIÓN

Cuando se genera con Vaex el hdf5 lo ejecuta mas rápido que sin él, sino tarda más que el resto. PANDAS VAEX Tiempo CSV HDF5 CSV Read 6.91 s 6.31 ms 8.2 s Describe 2.38 s 9.22 s 9.29 s *Creado por:*

Isabel Maniega