14 Redimientos formato de Archivos

June 18, 2025

Creado por:

Isabel Maniega

1 Ordenador con el que se ha hecho el experimento

Características:

- Intel® CoreTM i9-12900H CPU @ $2.40 \mathrm{GHz} \times 4$
- 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:

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

```
3
                           0.3
                                       34.80
                                                                              NaN
                                                                NaN
     4
                                       16.55
                           0.3
                                                                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
            8.760687e+06
                                              8760687
                                                                            8760687
     count
            1.560978e+00
                           2018-01-17 05:33:04.130081
                                                        2018-01-17 05:48:43.847028
    mean
    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
                                  2018-07-27 04:06:37
                                                               2018-07-27 04:46:57
     max
            2.000000e+00
            4.962678e-01
                                                                                NaN
     std
                                               RatecodeID PULocationID
                              trip_distance
            passenger_count
               8.760687e+06
                               8.760687e+06
                                             8.760687e+06
                                                           8.760687e+06
     count
                               2.804022e+00
               1.606807e+00
                                             1.039545e+00
                                                            1.644579e+02
     mean
                                                            1.000000e+00
    min
               0.000000e+00
                               0.000000e+00
                                             1.000000e+00
     25%
                                             1.000000e+00
                                                            1.160000e+02
               1.000000e+00
                               9.100000e-01
     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
                                             9.900000e+01
                                                            2.650000e+02
    max
               9.000000e+00
                               1.894838e+05
     std
               1.258420e+00
                               6.412050e+01
                                             4.450619e-01
                                                            6.635990e+01
            DOLocationID
                                          fare_amount
                           payment_type
                                                               extra
                                                                            mta_tax
            8.760687e+06
                           8.760687e+06
                                         8.760687e+06 8.760687e+06
                                                                      8.760687e+06
     count
            1.627270e+02
                          1.310613e+00
                                         1.224443e+01 3.246882e-01
                                                                      4.975066e-01
     mean
```

```
1.000000e+00
                      1.000000e+00 -4.500000e+02 -4.469000e+01 -5.000000e-01
min
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
        2.650000e+02
                      4.000000e+00
                                     8.016000e+03
                                                   6.000000e+01
                                                                  4.549000e+01
max
std
       7.031145e+01
                      4.817808e-01
                                     1.168321e+01
                                                   4.502555e-01
                                                                  4.333281e-02
         tip_amount
                      tolls_amount
                                     improvement_surcharge
                                                             total amount
       8.760687e+06
                      8.760687e+06
                                              8.760687e+06
                                                            8.760687e+06
count
        1.818759e+00
                      3.026157e-01
                                              2.996307e-01
                                                             1.549109e+01
mean
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
       4.417100e+02
                      9.507000e+02
                                              1.000000e+00
                                                             8.016800e+03
max
std
        2.486375e+00
                      1.738184e+00
                                              1.442748e-02
                                                             1.419546e+01
        congestion_surcharge
                              airport_fee
                        12.0
                                      12.0
count
                         2.5
                                       0.0
mean
                         2.5
                                       0.0
min
25%
                         2.5
                                       0.0
50%
                         2.5
                                       0.0
75%
                         2.5
                                       0.0
                         2.5
                                       0.0
max
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
                         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
                      2 2018-01-31 23:24:40
                                                 2018-01-31 23:25:28
      8760685
                                                                                     1
                       2 2018-01-31 23:28:16
                                                 2018-01-31 23:28:38
      8760686
                               RatecodeID store_and_fwd_flag
                                                               PULocationID
               trip_distance
                         2.80
      8760682
                                        1
                                                                         158
      8760683
                         0.60
                                        1
                                                            N
                                                                         163
      8760684
                         2.95
                                        1
                                                            N
                                                                          74
```

fare_amount

12.0

N

N

0.5

7

7

2.65

extra mta tax tip amount \

0.5

1

1

1

payment_type

8760685

8760686

8760682

0.00

0.00

163

DOLocationID

8760683	162		1		4.5	0.5	0.5	1.15
8760684	69		2	1	0.5	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	improve	ment_surch	arge	tota:	l_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_su	rcharge	airport_f	ee				
8760682		NaN	N	aN				
8760683		NaN	N	aN				
8760684		NaN	N	aN				
8760685		NaN	N	aN				

NaN

Tiempos:

8760686

• Con 1TB HDD sobre Windows: 0 nanosegundos

NaN

- 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 (1^a forma)

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.

CPU times: user 10.7 s, sys: 3.08 s, total: 13.8 s

Wall time: 12.1 s

[nn].	ш	II		Vd	TD +				
[23]:					_	p_pickup_dat		DatasadaTD	
				_		trip_dis LocationID		RatecodeID	
	store_and_ fare amoun		-	Locatio		mount tol			
	_			_		congestion_	_		- f
	0	.t_sur	Charge	1		8-01-01 00:2	_	_	
	1	U	0.5	1	1	0-01-01 00:2 N	1:05	2010-01-01	41
	24		2		4.5	0.5	0.5	0.0	41
	0.0		0.3		4.5	5.8	nan	0.0	
	nan		0.3			5.0	IIaII		
	1	1		1	201	8-01-01 00:4	1.55	2018-01-01	01 • 03 • 05
	1	1	2.7	1	1	N 01 01 00.4	1.00	2010 01 01	239
	140		2		14.0	0.5	0.5	0.0	200
	0.0		0.3		11.0	15.3	nan	0.0	
	nan		0.0			10.0	11011		
	2	2		1	201	8-01-01 00:0	8:26	2018-01-01	00:14:21
	2	_	0.8	_	1	N	0.120		262
	141		1		6.0	0.5	0.5	1.0	
	0.0		0.3			8.3	nan		
	nan								
	3	3		1	201	8-01-01 00:2	0:22	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	201	8-01-01 00:0	9:18	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	8760		1		8-01-31 23:2	1:35	2018-01-31	
	2		2.8		1	N	2 =	2 2-	158
	163		1		12.0	0.5	0.5	2.65	
	0.0		0.3			15.95	nan		

```
nan
8,760,683 8760683
                                        2018-01-31 23:35:51
                                                                  2018-01-31 23:38:57
                    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
                                        2018-01-31 23:28:00
                                                                  2018-01-31 23:37:09
1
                    2.95
                                       1
                                                      N
                                                                              74
                 2
                                                                        0.0
69
                                   10.5
                                                   0.5
                                                            0.5
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
                    0.0
                                                      N
                                                                              7
193
                 2
                                   0.0
                                                   0.0
                                                            0.0
                                                                        0.0
0.0
                 0.0
                                            0.0
                                                             nan
nan
                                        2018-01-31 23:28:16
                                                                  2018-01-31 23:28:38
8,760,686
           8760686
                    0.0
                 2
193
                                   0.0
                                                   0.0
                                                             0.0
                                                                        0.0
0.0
                 0.0
                                            0.0
                                                             nan
nan
```

- 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

tpep_dropoff_datetime

extra

store_and_fwd_flag

fare_amount

```
[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
```

passenger_count

PULocationID

mta_tax

tip_amount

DOLocationID

trip_distance

tolls_amount

RatecodeID

payment_type

improvemen	t_surcharge	total_amoun	t congestion_surchar	ge airport_fee
0	0		2018-01-01 00:21:05	-
1	0.5		1 N	41
24	2	4.5	0.5 0.5	0.0
0.0	0.3		5.8 na	n
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	0.3		15.3 na	
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	
0.0	0.3		8.3 na	
nan			270	_
3	3	1	2018-01-01 00:20:22	2018-01-01 00:52:51
1	10.2			140
257	2	33.5		
0.0	0.3	33.3	34.8 na	
nan	0.0		01.0	••
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		
0.0	0.3	12.0	16.55 na	
nan	0.0		10.00	••
		•••		
8 760 682				
	8760682	1	2018-01-31 23:21:35	2018-01-31 23:34:20
	8760682 2.8		2018-01-31 23:21:35 1 N	
2	2.8		1 N	158
2 163	2.8		1 N 0.5 0.5	158 2.65
2 163 0.0	2.8		1 N	158 2.65
2 163 0.0 nan	2.8 1 0.3	12.0	1 N 0.5 0.5 15.95 na	158 2.65 n
2 163 0.0 nan 8,760,683	2.8 1 0.3 8760683	12.0	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51	158 2.65 n 2018-01-31 23:38:57
2 163 0.0 nan 8,760,683	2.8 1 0.3 8760683 0.6	12.0	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N	2.65 n 2018-01-31 23:38:57 163
2 163 0.0 nan 8,760,683 1 162	2.8 1 0.3 8760683 0.6 1	12.0	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5	158 2.65 n 2018-01-31 23:38:57 163 1.15
2 163 0.0 nan 8,760,683 1 162 0.0	2.8 1 0.3 8760683 0.6	12.0	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N	158 2.65 n 2018-01-31 23:38:57 163 1.15
2 163 0.0 nan 8,760,683 1 162 0.0 nan	2.8 1 0.3 8760683 0.6 1 0.3	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na	158 2.65 n 2018-01-31 23:38:57 163 1.15
2 163 0.0 nan 8,760,683 1 162 0.0 nan 8,760,684	2.8 1 0.3 8760683 0.6 1 0.3	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na	158 2.65 n 2018-01-31 23:38:57 163 1.15 n 2018-01-31 23:37:09
2 163 0.0 nan 8,760,683 1 162 0.0 nan 8,760,684 1	2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na 2018-01-31 23:28:00 1 N	158 2.65 n 2018-01-31 23:38:57 163 1.15 n 2018-01-31 23:37:09 74
2 163 0.0 nan 8,760,683 1 162 0.0 nan 8,760,684 1 69	2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na 2018-01-31 23:28:00 1 N 0.5 0.5	158 2.65 n 2018-01-31 23:38:57 163 1.15 n 2018-01-31 23:37:09 74 0.0
2 163 0.0 nan 8,760,683 1 162 0.0 nan 8,760,684 1 69 0.0	2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na 2018-01-31 23:28:00 1 N	158 2.65 n 2018-01-31 23:38:57 163 1.15 n 2018-01-31 23:37:09 74 0.0
2 163 0.0 nan 8,760,683 1 162 0.0 nan 8,760,684 1 69	2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95 2 0.3	12.0 1 4.5	1 N 0.5 0.5 15.95 na 2018-01-31 23:35:51 1 N 0.5 0.5 6.95 na 2018-01-31 23:28:00 1 N 0.5 0.5	158 2.65 n 2018-01-31 23:38:57 163 1.15 n 2018-01-31 23:37:09 74 0.0 n

```
0.0
1
                                        1
                                                       N
                                                                                7
                  2
                                   0.0
193
                                                    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
```

- 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	${\tt PULocationID}$	\
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
                                                               fare_amount
                                        payment_type
data_type
                          int64
                                                int64
                                                                   float64
count
                        8760687
                                             8760687
                                                                   8760687
NA
                              0
                                                    0
                                                                          0
            162.72696718876043
                                 1.3106125124662027
                                                       12.244425999924395
mean
                      70.311441
                                            0.481781
std
                                                                 11.683206
min
                              1
                                                    1
                                                                    -450.0
                            265
                                                    4
                                                                    8016.0
max
                          extra
                                              mta_tax
                                                                 tip_amount
                        float64
                                              float64
                                                                    float64
data_type
count
                        8760687
                                              8760687
                                                                    8760687
NA
                              0
                                                     0
                                                                           0
            0.3246882464811264
                                 0.49750661106828725
                                                        1.8187593050633293
mean
std
                       0.450255
                                             0.043333
                                                                   2.486375
                         -44.69
                                                  -0.5
                                                                      -88.8
min
                           60.0
                                                 45.49
                                                                     441.71
max
                  tolls_amount improvement_surcharge
                                                                total_amount
data_type
                        float64
                                               float64
                                                                     float64
count
                        8760687
                                               8760687
                                                                     8760687
NA
                              0
                                                                            0
            0.3026157183791013
                                   0.29963067965078694
                                                         15.491088333535643
mean
                       1.738184
                                              0.014427
                                                                   14.195456
std
min
                          -15.0
                                                   -0.3
                                                                      -450.3
max
                          950.7
                                                    1.0
                                                                      8016.8
           congestion_surcharge airport_fee
                         float64
                                      float64
data_type
count
                              12
                                           12
                         8760675
                                      8760675
NA
                             2.5
                                          0.0
mean
std
                             0.0
                                          0.0
                             2.5
                                          0.0
min
                             2.5
                                          0.0
max
```

- 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
```

- 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 (2^{a} 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
                                                   trip_distance
                                                                     RatecodeID
                                passenger_count
      store_and_fwd_flag
                            PULocationID
                                             DOLocationID
                                                              payment_type
                                                        tolls_amount
      fare_amount
                     extra
                               mta_tax
                                          tip_amount
      improvement_surcharge
                                total_amount
                                                congestion_surcharge
                                                                         airport_fee
                                                                     2018-01-01 00:24:23
                                            2018-01-01 00:21:05
      0
                 0
                                1
```

24	1	0.5		1	N			41
nam 1 2.7 1 2018-01-01 00:44:55 2018-01-01 01:03:05 239 140 2.7 1 N 239 140 2.0 0.3 1 14.0 0.5 0.5 0.0 0.0 0.0 0.3 1 2018-01-01 00:08:26 2018-01-01 00:14:21 2 2018-01-01 00:14:21 2 2018-01-01 00:14:21 2 2018-01-01 00:14:21 2 2018-01-01 00:14:21 2 262 1 1 2018-01-01 00:00:12:2 2018-01-01 00:01:25 2018-01-01 00:01:25 2018-01-01 00:05:25 1 0 0.5 0.5 0.5 0.0 0						0.5	0.0	
1 1 2.7 1 2.7 1 1 0.141.5	0.0	0.3		5	.8	nan		
140					<u>.</u>			
140			1			5	2018-01-01	
Name						۸ ۶	0.0	239
nam 2 2 1 2018-01-01 00:08:26 2018-01-01 00:14:21 2018-01-01 00:14:21 2018-01-01 00:14:21 2018-01-01 00:14:21 2018-01-01 00:14:21 2018-01-01 00:14:21 2018-01-01 00:20:22 2018-01-01 00:52:51 1 0 0 5 1 0 0 5 1 0 <								
2 2 0.8 1 2018-01-01 00:08:26 2018-01-01 00:14:21 262 141 1 6.0 0.5 0.5 1.0 262 141 1 6.0 0.5 0.5 1.0 0 0.0 0.3 1 2018-01-01 00:20:22 2018-01-01 00:52:51 1 140 257 2 1 1 0.5 0.5 0.0		0.3		13	5.5	IIaII		
2 0.8 1 N 262 141 1 6.0 0.5 0.5 1.0 1 0.0 0.3 8.3 nan 1 2018-01-01 00:20:22 2018-01-01 00:52:51 1 1 10.2 1 2018-01-01 00:02:22 2018-01-01 00:52:51 140 257 2 33.5 0.5 0.5 0.0 0.0 0.3 34.8 nan 140 257 2 33.5 0.5 0.5 0.0 0.0 0.3 34.8 nan 2018-01-01 00:27:06 246 239 1 12.5 0.5 0.5 2.75 246 239 1 12.5 0.5 0.5 2.75 246 239 1 12.5 0.5 0.5 2.75 2.75 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 2.018-01-01 <td< td=""><td></td><td>2</td><td>1</td><td>2018-0</td><td>01-01 00:08:2</td><td>96</td><td>2018-01-01</td><td>00.14.21</td></td<>		2	1	2018-0	01-01 00:08:2	96	2018-01-01	00.14.21
141			-			.0	2010 01 01	
0.0 0.3 8.3 nam 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 0.0 0.0 0.3 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 1 12.5 0.5 0.5 2.75 0.0 0.3 1 12.5 0.5 0.5 2.75 0.0 0.3 1 12.5 0.5 0.5 2.75 0.0 0.3 1 2018-01-31 23:21:35 nam nam 8,760,682 1 2018-01-31 23:21:35 2018-01-31 23:33:35:1 2018-01-31 23:33:35:1 2018-01-31 23:33:35:1 2018-01-31 23:33:35:1 2018-01-31 23:33:35:1 2018-01-31 23:33:35:1 2018-01-31 23:33:37:0 163 1 1 N <						0.5	1.0	
3 3 1 2018-01-01 00:20:22 2018-01-01 00:52:51 140 257 2 33.5 0.5 0.5 0.0 0.0 0.0 0.3 34.8 nan nan 1								
1 10.2 13.5 140 14	nan							
257 2 33.5 0.5 0.5 0.0 0.0 0.0 0.3 34.8 nan nan nan nan nan nan 1 2018-01-01 00:09:18 2018-01-01 00:27:06 2018-01-01 00:27:06 226 230 21 0.0 0.5 0.5 2.75 2.75 2.75 0.0 0.0 0.3 1 12.5 0.5 0.5 2.75 0.5 2.75 0.5 2.75 0.0 0.0 0.3 16.55 nan 1 0.0 <th< td=""><td>3</td><td>3</td><td>1</td><td>2018-0</td><td>01-01 00:20:2</td><td>22</td><td>2018-01-01</td><td>00:52:51</td></th<>	3	3	1	2018-0	01-01 00:20:2	22	2018-01-01	00:52:51
0.0 0.3 34.8 nam nam 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 1 16.55 nam 1 1 1 16.55 nam 1 1 1 1 16.55 nam 1 <	1	10.2		1	N			140
nan 4 4 1 2018-01-01 00:09:18 2018-01-01 00:09:18 2018-01-01 00:27:06 239 1 12.5 0.5 0.5 0.5 2.75 0.0 0.0 0.3 1 12.5 0.5 0.5 0.5 2.75 0.0 nan 1 12.5 0.5 0.5 0.5 2.75 0.0 0.0 0.3 0.5	257	2		33.5	0.5	0.5	0.0	
4 4 1 2018-01-01 00:09:18 2018-01-01 00:27:06 239 1 12.5 0.5 0.5 2.75 0.0 0.3 12.5 0.5 0.5 2.75 0.0 0.3 12.5 0.5 0.5 2.75 0.0 0.3 0.5 0.5 0.5 2.75 0.0 0.3 0.5 0.5 0.5 0.5 0.0 <td>0.0</td> <td>0.3</td> <td></td> <td>34</td> <td>1.8</td> <td>nan</td> <td></td> <td></td>	0.0	0.3		34	1.8	nan		
2 2.5 1 N 246 239 1 12.5 0.5 0.5 2.75 0.0 0.3 16.55 nan 16.55 nan nan """"""""""""""""""""""""""""""""""""								
239			1			.8	2018-01-01	
0.0 0.3 16.55 nan nan								246
nam								
		0.3		16	5.55	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 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 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 0.5 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								
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 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 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 0.5 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	IIaII							
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 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 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 0.5 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								
2 2.8 1 N 158 163 1 12.0 0.5 0.5 2.65 0.0 0.3 15.95 nan 0.0 8,760,683 8760683 1 2018-01-31 23:35:51 2018-01-31 23:38:57 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 0.0 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 nan 8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7								
2 2.8 1 N 158 163 1 12.0 0.5 0.5 2.65 0.0 0.3 15.95 nan 0.0 8,760,683 8760683 1 2018-01-31 23:35:51 2018-01-31 23:38:57 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 0.0 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 nan 8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7		 						
163 1 12.0 0.5 0.5 2.65 0.0 0.0 0.3 15.95 nan 15.95 nan 8,760,683 8760683 1 2018-01-31 23:35:51 2018-01-31 23:38:57 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 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 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7		 		 	 	•••		
0.0 0.3 15.95 nam 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 0.5 1.15 0.0 0.3 6.95 nan 74 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 0.5 0.0 0.0 0.3 11.8 nan 11.8 nan 8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7		 8760682		 2018-0	 01-31 23:21:3	 35	 2018-01-31	23:34:20
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	 8,760,682		1			 35	2018-01-31	
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 0.5 1.15 0.0 0.3 6.95 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 0.5 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	 8,760,682	2.8	1	1	N			
1 0.6 1 N 163 162 1 4.5 0.5 0.5 1.15 0.0 0.3 6.95 nan nan 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	 8,760,682 2 163	2.8	1	1 12.0	N 0.5	0.5		
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 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7	 8,760,682 2 163 0.0	2.8 1 0.3		1 12.0	N 0.5 5.95	0.5 nan	2.65	158
0.0 0.3 6.95 nam nan 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 0.5 0.0 0.0 0.3 11.8 nan	 8,760,682 2 163 0.0 nan 8,760,683	2.8 1 0.3 8760683		1 12.0 1! 2018-0	N 0.5 5.95 01-31 23:35:5	0.5 nan	2.65	158 23:38:57
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 1 1 1 8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 2 2 1 0.0 1 N 7 7	 8,760,682 2 163 0.0 nan 8,760,683	2.8 1 0.3 8760683 0.6		1 12.0 19 2018-0	N 0.5 5.95 01-31 23:35:5 N	0.5 nan	2.65 2018-01-31	158 23:38:57
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 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7	 8,760,682 2 163 0.0 nan 8,760,683 1 162	2.8 1 0.3 8760683 0.6		1 12.0 18 2018-0 1 4.5	N 0.5 5.95 01-31 23:35:5 N 0.5	0.5 nan 51 0.5	2.65 2018-01-31 1.15	158 23:38:57
1 2.95 1 N 74 69 2 10.5 0.5 0.5 0.0 0.0 0.3 11.8 nan	 8,760,682 2 163 0.0 nan 8,760,683 1 162 0.0	2.8 1 0.3 8760683 0.6		1 12.0 18 2018-0 1 4.5	N 0.5 5.95 01-31 23:35:5 N 0.5	0.5 nan 51 0.5	2.65 2018-01-31 1.15	158 23:38:57
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	 8,760,682 2 163 0.0 nan 8,760,683 1 162 0.0 nan	2.8 1 0.3 8760683 0.6 1 0.3	1	1 12.0 19 2018-0 1 4.5	N 0.5 5.95 01-31 23:35:5 N 0.5	0.5 nan 51 0.5 nan	2.65 2018-01-31 1.15	158 23:38:57 163
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		2.8 1 0.3 8760683 0.6 1 0.3	1 2	1 12.0 18 2018-0 1 4.5 6	N 0.5 5.95 01-31 23:35:5 N 0.5 .95	0.5 nan 51 0.5 nan	2.65 2018-01-31 1.15	158 23:38:57 163 23:37:09
nan 8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7		2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95	1 2	1 12.0 18 2018-0 1 4.5 6	N 0.5 5.95 01-31 23:35:5 N 0.5 .95	0.5 nan 61 0.5 nan	2.65 2018-01-31 1.15 2018-01-31	158 23:38:57 163 23:37:09
8,760,685 8760685 2 2018-01-31 23:24:40 2018-01-31 23:25:28 1 0.0 1 N 7		2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95 2	1 2	1 12.0 18 2018-0 1 4.5 6 2018-0	N 0.5 5.95 01-31 23:35:5 N 0.5 .95 01-31 23:28:0 N 0.5	0.5 nan 61 0.5 nan 00	2.65 2018-01-31 1.15 2018-01-31 0.0	158 23:38:57 163 23:37:09
1 0.0 1 N 7		2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95 2	1 2	1 12.0 18 2018-0 1 4.5 6 2018-0	N 0.5 5.95 01-31 23:35:5 N 0.5 .95 01-31 23:28:0 N 0.5	0.5 nan 61 0.5 nan 00	2.65 2018-01-31 1.15 2018-01-31 0.0	158 23:38:57 163 23:37:09
193 2 0.0 0.0 0.0 0.0		2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95 2 0.3	1	1 12.0 18 2018-0 1 4.5 6 2018-0 1 10.5	N 0.5 5.95 01-31 23:35:5 N 0.5 .95 01-31 23:28:0 N 0.5	0.5 nan 61 0.5 nan 00 0.5 nan	2.65 2018-01-31 1.15 2018-01-31 0.0	158 23:38:57 163 23:37:09 74
		2.8 1 0.3 8760683 0.6 1 0.3 8760684 2.95 2 0.3	1	1 12.0 18 2018-0 1 4.5 6 2018-0 1 10.5	N 0.5 5.95 01-31 23:35:5 N 0.5 .95 01-31 23:28:0 N 0.5	0.5 nan 61 0.5 nan 00 0.5 nan	2.65 2018-01-31 1.15 2018-01-31 0.0	158 23:38:57 163 23:37:09 74 23:25:28

```
0.0
                       0.0
                                                  0.0
                                                                   nan
      nan
      8,760,686
                  8760686
                                              2018-01-31 23:28:16
                                                                        2018-01-31 23:28:38
                                             1
                          0.0
      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
                                                                   8760687
      count
                        8760687
                                             8760687
      NA
                                                                          0
                      4380343.0
                                  1.560978037452999
      mean
      std
                  2528992.49887
                                            0.496268
      min
                               0
                                                   1
                                                   2
                        8760686
      max
                 tpep_dropoff_datetime
                                             passenger_count
                                                                    trip_distance \
                                                        int64
                                                                           float64
      data_type
                                 string
      count
                                8760687
                                                     8760687
                                                                           8760687
      NA
                                      0
                                                            0
      mean
                                          1.6068074341658365
                                                               2.8040221012347804
                                                     1.25842
                                                                         64.120494
      std
      min
                                                            0
                                                                               0.0
                                                            9
                                                                         189483.84
      max
                          RatecodeID store_and_fwd_flag
                                                                  PULocationID
                                int64
                                                   string
                                                                          int64
      data_type
                              8760687
                                                  8760687
                                                                        8760687
      count
```

0

0

1

265

66.359897

164.45790906580729

0

1

99

0.445062

1.0395453004998352

NΑ

mean

std min

max

```
DOLocationID
                                                               fare_amount
                                        payment_type
                          int64
                                                                   float64
data_type
                                                int64
count
                       8760687
                                             8760687
                                                                   8760687
NA
                                                    0
            162.72696718876043
                                  1.3106125124662027
                                                       12.244425999924394
mean
std
                      70.311441
                                            0.481781
                                                                 11.683206
                                                                    -450.0
min
                                                    1
                                                    4
                            265
                                                                    8016.0
max
                                                                 tip_amount
                          extra
                                              mta_tax
                                                                    float64
data_type
                        float64
                                               float64
count
                        8760687
                                               8760687
                                                                    8760687
NA
mean
            0.3246882464811265
                                 0.49750661106828725
                                                        1.8187593050633293
                       0.450255
                                             0.043333
std
                                                                   2.486375
min
                         -44.69
                                                  -0.5
                                                                      -88.8
                           60.0
                                                 45.49
                                                                     441.71
max
                  tolls_amount improvement_surcharge
                                                               total_amount
                        float64
                                                float64
                                                                    float64
data_type
                        8760687
                                                8760687
                                                                    8760687
count
NΑ
                              0
                                                      0
                                                                           0
            0.3026157183791013
                                     0.299630679650787
                                                          15.49108833353564
mean
                                               0.014427
                                                                  14.195456
std
                       1.738184
                                                   -0.3
                                                                     -450.3
min
                          -15.0
max
                          950.7
                                                    1.0
                                                                     8016.8
           congestion_surcharge airport_fee
data_type
                         float64
                                      float64
count
                              12
                                           12
                         8760675
                                      8760675
NΑ
                             2.5
                                          0.0
mean
                             0.0
                                          0.0
std
min
                             2.5
                                          0.0
                             2.5
                                          0.0
max
```

[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

```
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
                               \mathtt{mta}\_\mathtt{tax}
                                          tip_amount
                                                       tolls_amount
      improvement_surcharge
                                total amount
                                                congestion_surcharge
                                                                          airport_fee
                 0
                                            2018-01-01 00:21:05
                                                                     2018-01-01 00:24:23
      1
                          0.5
                                                                                 41
                                                          N
      24
                       2
                                       4.5
                                                       0.5
                                                                0.5
                                                                            0.0
      0.0
                       0.3
                                                 5.8
                                                                 nan
      nan
      1
                                            2018-01-01 00:44:55
                                                                     2018-01-01 01:03:05
                 1
                                1
                          2.7
                                           1
                                                          N
                                                                                 239
                       2
      140
                                       14.0
                                                       0.5
                                                                0.5
                                                                            0.0
      0.0
                       0.3
                                                 15.3
                                                                 nan
      nan
                 2
                                            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
                                                 8.3
      0.0
                       0.3
                                                                 nan
      nan
      3
                 3
                                            2018-01-01 00:20:22
                                1
                                                                      2018-01-01 00:52:51
                          10.2
      1
                                           1
                                                                                 140
      257
                       2
                                       33.5
                                                       0.5
                                                                0.5
                                                                            0.0
      0.0
                       0.3
                                                 34.8
                                                                 nan
      nan
                                            2018-01-01 00:09:18
                                                                     2018-01-01 00:27:06
                                1
      2
                          2.5
                                           1
                                                                                 246
                                                          N
      239
                       1
                                       12.5
                                                       0.5
                                                                0.5
                                                                            2.75
      0.0
                       0.3
                                                 16.55
      nan
      •••
                                             2018-01-31 23:21:35
                                                                     2018-01-31 23:34:20
      8,760,682
                 8760682
                                           1
                          2.8
                                                                                 158
      163
                                       12.0
                                                       0.5
                                                                0.5
                                                                            2.65
                       1
      0.0
                                                 15.95
                       0.3
                                                                 nan
      nan
      8,760,683 8760683
                                1
                                            2018-01-31 23:35:51
                                                                     2018-01-31 23:38:57
```

[33]: %%time

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	8-01-3	1 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	8-01-3	1 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	8-01-3	1 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								

- 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 $6.91~\rm s$ $6.31~\rm ms$ $8.2~\rm s$ Describe $2.38~\rm s$ $9.22~\rm s$ $9.29~\rm s$ Creado por:

Isabel Maniega