1. How many rows does datagrams-nonstd.csv have?

**RTA**:

## cantidad de registros datagrams-nonstd.csv

wc -l datagrams-nonstd.csv

814547049 datagrams-nonstd.csv

#Tamaño antes de procesar

$ ls -lart | awk "{print \$9 \" -> \" \$5/1024/1024/1024 \" GB\"}" | grep datagrams

datagrams-nonstd.csv -> 74.0533 GB

1. Use awk to fix the problems found in datagrams-nonstd.csv, save the resulting dataset in the file datagrams.csv. How many rows does datagrams.csv have?

**RTA**: Se puede ver que cada 1.000.000 hay un registro (en total hay 814 registros así) que tiene una estructura con las columnas repetidas y con una columna adicional al inicio

Ej.

SUP, 0,27-APR-19,501450,4,34433267,-765261533,137,150,317,6184585259,2019-04-26 20:25:33,645 0,27-APR-19,501450,4,34433267,-765261533,137,150,317,6184585259,2019-04-26 20:25:33,645

Es decir,

SUP, COLUMNAS\_1 COLUMNAS\_1

Para corregir este problema se usó awk (Adjunto) quitando las columnas repetidas manteniendo solo las columnas con la estructura adecuada.



Este script se ejecutó de la siguiente manera

## Procesar registros con errores

date; awk -F\, -f "../99\_scripts/funciones.awk" datagrams-nonstd.csv > datagrams.csv; date;

# Ver la cantidad de registros de la salida después del procesamiento

date; wc -l datagrams.csv; date;

814547049 datagrams.csv

#Tamaño despues de procesar

$ ls -lart | awk "{print \$9 \" -> \" \$5/1024/1024/1024 \" GB\"}" | grep datagrams

datagrams-nonstd.csv -> 74.0533 GB

datagrams.csv -> 74.0532 GB

1. Extract all rows from datagrams.csv that were generated on April 26 2019 (column 11) to output26-APR-19.csv

Transform the format of column 11 (e.g., using substitution): this is a timestamp, but not in a standard format. Change it to the format ‘YYYY-MM-DD HH24:MI’. Eliminating seconds and milliseconds will help you to build the histogram of datagrams by minute.

Example: transform from “30-APR-19 03.41.05.000000” to “2019-04-30 03:41:05”

## Extraer solo las filas de 26 APR 2019

awk -F\, '$11 ~ '/^26-APR-19/' {print $0}' datagrams.csv > temp-26-APR-19.csv

## Ver la cantidad de registros de la salida del borrado

wc -l temp-26-APR-19.csv

2403018 temp-26-APR-19.csv

#Cambiar el formato de las fechas

awk -F\, -f ejercicio\_p2.awk temp-26-APR-19.csv > output-26-APR-19.csv

#contar los datagramas del 26 de abril 2019

wc -l output-26-APR-19.csv

2403018 output-26-APR-19.csv



#revisar filas

head output-26-APR-19.csv

0,27-APR-19,501450,4,34433267,-765261533,137,150,317,6184585259,2019-04-26 20:25:33,645

0,27-APR-19,514413,6,34547417,-764847817,816,2241,579,6184585260,2019-04-26 20:25:33,1030

0,27-APR-19,501251,558,34449317,-765107250,530,140,957,6184585261,2019-04-26 20:25:33,566

0,27-APR-19,513254,253,34872400,-765117100,994,2212,304,6184585262,2019-04-26 20:25:33,28

0,27-APR-19,500165,259,34420833,-764812083,658,2243,68,6184585263,2019-04-26 20:25:33,640

0,27-APR-19,-1,-1,33707767,-765364517,-1,-1,-1,6184585264,2019-04-26 20:25:33,67

0,27-APR-19,500250,12690,34986717,-764995100,-1,999,-1,6184585265,2019-04-26 20:25:33,1175

0,27-APR-19,513188,36,34636217,-764979667,135,3371,280,6184585266,2019-04-26 20:25:33,200

0,27-APR-19,500102,87,34894917,-765074900,343,282,393,6184585267,2019-04-26 20:25:33,808

0,27-APR-19,513188,36,34635883,-764980833,135,3371,280,6184585268,2019-04-26 20:25:33,200

#revisar colas de archivo

tail output-26-APR-19.csv

0,27-APR-19,511501,-1,33659167,-765166150,413,3133,85,6185959721,2019-04-26 09:36:12,116

0,27-APR-19,514081,32,34228483,-764836583,38,2842,600,6185959722,2019-04-26 09:36:12,1107

0,27-APR-19,514075,-1,34152817,-764798967,163,357,44,6185959723,2019-04-26 09:36:12,277

0,27-APR-19,500602,105,34535333,-765317083,857,421,633,6185959724,2019-04-26 09:36:12,637

0,27-APR-19,516233,36,34082933,-765448717,569,2473,66,6185959725,2019-04-26 09:36:12,299

0,27-APR-19,513273,33,34640133,-765311633,399,2301,92,6185959726,2019-04-26 09:36:12,244

0,27-APR-19,511372,14,33777650,-765531117,359,3121,408,6185959727,2019-04-26 09:36:12,118

0,27-APR-19,514056,21,34405617,-764791550,815,3411,449,6185959728,2019-04-26 09:36:12,934

0,27-APR-19,517184,4,34083367,-765707250,19,375,411,6185959729,2019-04-26 09:36:12,1202

0,27-APR-19,502601,39,34187700,-764871550,439,1571,252,6185959730,2019-04-26 09:36:12,21

4.How many datagrams were generated on April 26 2019?

#contar los datagramas del 26 de abril 2019

wc -l output-26-APR-19.csv

2403018 output-26-APR-19.csv

5.Plot a histogram of datagrams generated by minute using Python or R or another tool.

Gráfico, Histograma

Descripción generada automáticamente

Codigo en Python para generarlo



https://github.com/carlosjara/MCD\_PDD/blob/main/procesamiento\_taller\_1.ipynb