ANDRES DAVID CANDELO LOPEZ

TALLER 1 AWK

EXTRACCION

Comprobamos la ruta donde se guardarán los archivos

```
andre@Andres /cygdrive/d
$ pwd
/cygdrive/d
```

Verificamos velocidad de la conexión a internet.

Comandos

time wget -O datag-nonstd-p1.zip "https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EeW-IX84sANCj8WJiWb6ZSEBr-Rg_Qr1JhzfyYXgkgBYRg?e=PZltC3&download=1" && \

time wget -O datag-nonstd-p2.zip "https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EQh7AzMdFCFPsobZz8f5NRwBwldA0c5arCxV0WScl20FZA?e=xyGTyQ&download=1" && \

date && \

time wget -O datag-nonstd-p3.zip https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EfYxbZUWFCRMu4CuqAo7i_QBXrqtZjykKQOmO63Lx1onbw?e=OGSZh6&download=1

```
andre@Andres /cygdrive/d
$ time wget -O datag-nonstd-p1.zip "https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EeW-IX84sANCj8WJiWb6ZSEBr-Rg_Qr1JhzfyYXgkgBYRg?e=PZltC3&download=1" && time wget -O datag-nonstd-p2.z
p "https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EQh7AzMdFCFPsobZz8f5NRwBwldAOc5arCxVOwScl2OFZA?e=xyGTyQ&download=1" && date && time wget -O datag-nonstd-p3.zip "https://icesiedu-my.s
harepoint.com/:u:/g/personal/16282252_icesi_edu_co/EfYxbZUWFCRMu4CuqAo7i_QBXrqtZjykKQOmO63Lx1onbw?e=OGSZh6&download=1"
--2023-11-12 13:03:44--  https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EeW-IX84sANCj8WJiWb6ZSEBr-Rg_Qr1JhzfyYXgkgBYRg?e=PZltC3&download=1
Resolving icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)... 13.107.136.10, 13.107.138.10, 2620:1ec:8f8::10, ...
Connecting to icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)|13.107.136.10|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: /personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p1.zip?ga=1 [following]
--2023-11-12 13:03:45--  https://icesiedu-my.sharepoint.com/personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p1.zip?ga=1
Reusing existing connection to icesiedu-my.sharepoint.com:443.
HTTP request sent, awaiting response... 200 OK
Length: 6443237493 (6.0G) [application/x-zip-compressed]
Saving to: 'datag-nonstd-p1.zip'

datag-nonstd-p1.zip                         100%[=============================================================================================================>]   6.00G  16.9MB/s    in 7m 29s

2023-11-12 13:11:14 (13.7 MB/s) - 'datag-nonstd-p1.zip' saved [6443237493/6443237493]


real    7m30.362s
user    0m13.515s
sys     0m6.671s
--2023-11-12 13:11:14--  https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EQh7AzMdFCFPsobZz8f5NRwBwldAOc5arCxVOwScl2OFZA?e=xyGTyQ&download=1
Resolving icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)... 13.107.136.10, 13.107.138.10, 2620:1ec:8f8::10, ...
Connecting to icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)|13.107.136.10|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: /personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p2.zip?ga=1 [following]
--2023-11-12 13:11:15--  https://icesiedu-my.sharepoint.com/personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p2.zip?ga=1
Reusing existing connection to icesiedu-my.sharepoint.com:443.
HTTP request sent, awaiting response... 200 OK
Length: 6443229928 (6.0G) [application/x-zip-compressed]
Saving to: 'datag-nonstd-p2.zip'

datag-nonstd-p2.zip                         100%[=============================================================================================================>]   6.00G  17.2MB/s    in 6m 21s

2023-11-12 13:17:36 (16.1 MB/s) - 'datag-nonstd-p2.zip' saved [6443229928/6443229928]


real    6m21.874s
user    0m5.328s
sys     0m3.171s
Sun Nov 12 13:17:36 -05 2023
--2023-11-12 13:17:37--  https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EfYxbZUWFCRMu4CuqAo7i_QBXrqtZjykKQOmO63Lx1onbw?e=OGSZh6&download=1
Resolving icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)... 13.107.136.10, 13.107.138.10, 2620:1ec:8f8::10, ...
Connecting to icesiedu-my.sharepoint.com (icesiedu-my.sharepoint.com)|13.107.136.10|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: /personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p3.zip?ga=1 [following]
--2023-11-12 13:17:38--  https://icesiedu-my.sharepoint.com/personal/16282252_icesi_edu_co/Documents/Cursos/MCD-ProcesamientoDistribuidoDatos/StudentKit0-Tools-Data/datagrams/datag-nonstd-p3.zip?ga=1
Reusing existing connection to icesiedu-my.sharepoint.com:443.
HTTP request sent, awaiting response... 200 OK
Length: 5720736254 (5.3G) [application/x-zip-compressed]
Saving to: 'datag-nonstd-p3.zip'

datag-nonstd-p3.zip                         100%[=============================================================================================================>]   5.33G  14.9MB/s    in 5m 1s

2023-11-12 13:22:39 (18.1 MB/s) - 'datag-nonstd-p3.zip' saved [5720736254/5720736254]


real    5m2.641s
user    0m4.109s
sys     0m2.343s
```

Verificación de integridad de los archivos

comandos

time zip -t datag-nonstd-p1.zip

time zip -t datag-nonstd-p2.zip

time zip -t datag-nonstd-p3.zip

```
andre@Andres /cygdrive/d
$ time zip -T datag-nonstd-p2.zip
test of datag-nonstd-p2.zip OK

real    0m29.200s
user    0m5.358s
sys     0m0.639s

andre@Andres /cygdrive/d
$ time zip -T datag-nonstd-p3.zip
test of datag-nonstd-p3.zip OK

real    0m22.627s
user    0m9.125s
sys     0m0.936s

andre@Andres /cygdrive/d
$ time zip -T datag-nonstd-p1.zip
test of datag-nonstd-p1.zip OK

real    0m27.544s
user    0m7.812s
sys     0m0.670s
```

Creamos la ruta donde se va a ejecutar el unzip

```
andre@Andres /cygdrive/d
$ mkdir tarea1_awk
```

Descargamos en la ruta
unzip datag-nonstd-p1.zip -d tarea1_awk

unzip datag-nonstd-p2.zip -d tarea1_awk

unzip datag-nonstd-p3.zip -d tarea1_awk

```
andre@Andres /cygdrive/d
$ unzip datag-nonstd-p1.zip -d tarea1_awk
unzip datag-nonstd-p2.zip -d tarea1_awk
unzip datag-nonstd-p3.zip -d tarea1_awk
Archive:  datag-nonstd-p1.zip
  inflating: tarea1_awk/datagaa
Archive:  datag-nonstd-p2.zip
  inflating: tarea1_awk/datagab
Archive:  datag-nonstd-p3.zip
  inflating: tarea1_awk/datagac
```

Comprobamos que si se hayan descomprimido los archivos en la ruta.

```
andre@Andres /cygdrive/d
$ ls tarea1_awk
datagaa  datagab  datagac
```

Comprimimos los archivos en 1 solo y se cambia de directorio y unificamos los archivos en 1 solo

```
andre@Andres /cygdrive/d
$ cd tarea1_awk

andre@Andres /cygdrive/d/tarea1_awk
$ cat datagaa datagab datagac > datagrams-nonstd.zip

andre@Andres /cygdrive/d/tarea1_awk
$ ls datagrams-nonstd.zip
datagrams-nonstd.zip
```

Verificación de integridad y descomprimimos los archivos

```
andre@Andres /cygdrive/d/tarea1_awk
$ time zip -T datagrams-nonstd.zip
time unzip datagrams-nonstd.zip
test of datagrams-nonstd.zip OK

real    5m37.294s
user    2m2.171s
sys     0m2.718s
Archive:  datagrams-nonstd.zip
  inflating: datagrams-nonstd.csv

real    9m19.352s
user    2m50.390s
sys     0m16.515s
```

TRANSFORMACION

1. Cantidad de filas del dataset nonstd

```
andre@Andres /cygdrive/d/tarea1_awk
$ wc -l datagrams-nonstd.csv
814547049 datagrams-nonstd.csv
```

2. Verificamos consistencia de los datos

   Comandos

   ```
   $ awk -F',' '{ print NR "," NF ", " $0 }' datagrams-nonstd.csv > errores.csv
   ```

   ```
   andre@Andres /cygdrive/d/tarea1_awk
   $ awk -F',' '{ print NR "," NF ", " $0 }' datagrams-nonstd.csv > errores.csv
   ```

   En esta línea agregamos un contador de fila y un contador de columnas por cada fila del set de datos luego comprobamos las filas que presentan inconsistencias en la cantidad de registros por columna con el siguiente comando:

   ```
   $ awk -F',' 'NF != 14 { print }' errores.csv
   ```

   Como resultados obtenemos las filas que tienen mas de 12 campos. Estas se imprimen por numero de fila y cantidad de columnas

   Teniendo encuenta las inconsistencias presentadas se crea un nuevo archivo csv omitiendo estas filas con errores.

   Comando

   ```
   $ awk -F',' 'NF == 12 { print $0 }' datagrams-nonstd.csv > datagrams.csv
   ```

   ```
   andre@Andres /cygdrive/d/tarea1_awk
   $ awk -F',' 'NF == 12 { print $0 }' datagrams-nonstd.csv > datagrams.csv
   ```

- -F',': Especifica que el separador de campos es una coma.
- NF == 12: Verifica si el número de campos en una línea es igual a 12.
- { print $0 }: Si la condición es verdadera, se guarda en un archivo datagrams.

Comprobamos la cantidad de registros en datagrams

```
andre@Andres /cygdrive/d/tarea1_awk
$ wc -l datagrams.csv
814546235 datagrams.csv
```

3. Ahora aplicamos un filtro para obtener las fechas del día 28 de abril del 2019 con el siguiente comando:

```
$ awk -F',' '$11 ~ /28-APR-19/ {print $0}' datagrams.csv > output28-APR-19.csv
```

```
andre@Andres /cygdrive/d/tarea1_awk
$ awk -F',' '$11 ~ /28-APR-19/ {print $0}' datagrams.csv > output28-APR-19.csv
```

Head

```
andre@Andres /cygdrive/d/tarea1_awk
$ head  output28-APR-19.csv
0,29-APR-19,511202,155,33178500,-765365033,1041,3191,99993,6188881314,28-APR-19 09.47.45.000000 PM,1130
0,29-APR-19,514277,-1,34224250,-764709117,40234,3441,99993,6188881315,28-APR-19 09.47.45.000000 PM,1094
0,29-APR-19,516230,338,33977467,-765241767,360,2121,579,6188881316,28-APR-19 09.47.45.000000 PM,1065
0,29-APR-19,501501,5,34489083,-765278317,183,2301,682,6188881317,28-APR-19 09.47.45.000000 PM,153
0,29-APR-19,-1,-1,33585633,-765116517,-1,-1,-1,6188881318,28-APR-19 09.47.46.000000 PM,80
0,29-APR-19,514042,148,34181750,-764660883,74,3411,679,6188881319,28-APR-19 09.47.46.000000 PM,922
0,29-APR-19,514400,411,34340950,-764841917,394,2471,343,6188881320,28-APR-19 09.47.46.000000 PM,161
0,29-APR-19,500110,-1,34893400,-765078183,2189,322,99967,6188881321,28-APR-19 09.47.46.000000 PM,190
0,29-APR-19,516087,31,33961383,-765134217,380,2473,508,6188881322,28-APR-19 09.47.46.000000 PM,1035
0,29-APR-19,518063,16,34199283,-765072717,426,2212,688,6188881323,28-APR-19 09.47.46.000000 PM,1181
```

Tail

```
andre@Andres /cygdrive/d/tarea1_awk
$ tail  output28-APR-19.csv
0,29-APR-19,513119,3,34583700,-765147350,88,2402,484,6189192798,28-APR-19 07.22.30.000000 PM,152
0,29-APR-19,514044,65,34234367,-764636300,313,3411,637,6189192799,28-APR-19 07.22.30.000000 PM,918
0,29-APR-19,511100,23,33684250,-765203883,315,3131,556,6189192800,28-APR-19 07.22.30.000000 PM,700
0,29-APR-19,513101,34,34677000,-765016600,367,2402,481,6189192801,28-APR-19 07.22.30.000000 PM,784
0,29-APR-19,513003,47,34518567,-765370717,156,305,197,6189192802,28-APR-19 07.22.30.000000 PM,1182
0,29-APR-19,502312,2,33666100,-765280450,111,2211,9963,6189192803,28-APR-19 07.22.30.000000 PM,73
0,29-APR-19,519080,5,34465650,-765348800,443,302,602,6189192804,28-APR-19 07.22.30.000000 PM,193
0,29-APR-19,513191,284,34813367,-765093533,31,282,171,6189192805,28-APR-19 07.22.30.000000 PM,145
0,29-APR-19,502101,1,33875467,-765447583,376,441,638,6189192806,28-APR-19 07.22.30.000000 PM,31
0,29-APR-19,514156,465,34432633,-764838833,160,2241,247,6189192807,28-APR-19 07.22.30.000000 PM,820
```

4.Cantidad de filas para el dataset filtrado

```
andre@Andres /cygdrive/d/tarea1_awk
$ wc -l  output28-APR-19.csv
1662470 output28-APR-19.csv
```

Comprobamos consistencia de las variables según el diccionario de datos

Variable	Tipo	Descripción	Rango Min	Rango Max
eventType	Integer	Tipo de Evento	0	10000
registerdate	Date	Fecha y hora en la que se registró el datagram en el log del NetPeerManager	31-may-18	31-may-19
stopId	Integer	Identificador de la última parada por la que pasó el bus en el IVU	-1	26840744
odometer	Integer	La distancia en metros recorridos por el bus desde la última parada hasta la ubicación actual	-1	121600820
latitude	Integer	latitud (y) de la posición del bus en el sistema de coordenadas geográfico del mundo	-1288197850	1934980197
longitude	Integer	longitud (x) de la posición del bus en el sistema de coordenadas geográfico del mundo	-133761367	845963520
taskId	Integer	Identificador de la tarea que tiene asignada el bus en el IVU	-1	98517645
lineId	Integer	Identificador de la línea que tiene asignada el bus en el IVU	-1	4272
tripId	Integer	Identificador del viaje que tiene asignado el bus en el IVU	-1	687964161
unknown1	Integer	? Número que representa el tipo de evento, manteniendo la tipificación registrada en los mensajes.	0	103
datagramDate	Date	Fecha y hora en la que ocurrió el datagram en el bus	31-MAY-18 12.00.00.000000 AM	30-MAY-19 11.59.59.000000 PM
busId		Identificador del bus en el sistema IVU	1	8502

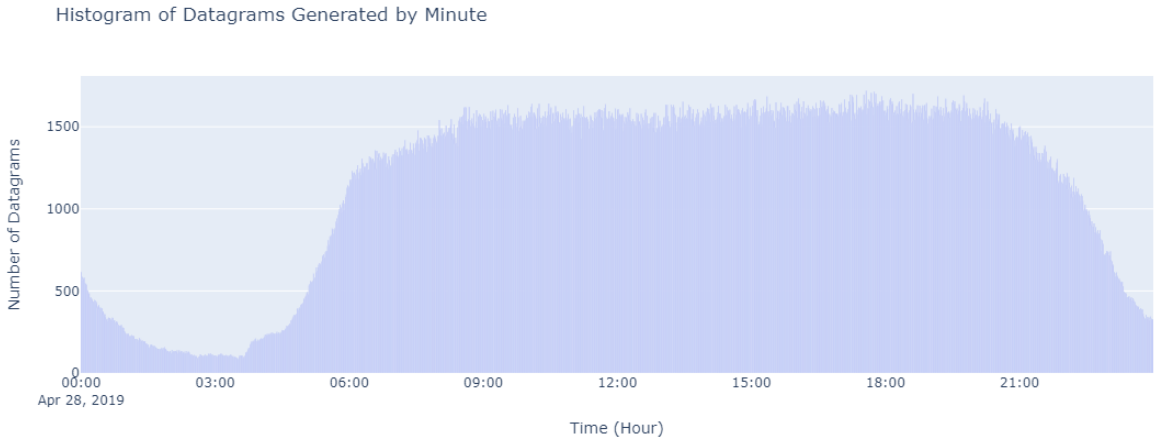
```
andre@Andres /cygdrive/d/tarea1_awk
$ awk -F',' '($1 < 0 || $1 > 10000 || $4 < -1 || $4 > 121600820 || $5 < -1288197850 || $5 > 1934980197
 687964161 || $10 < 0 || $10 > 103 || $12 < 1 || $12 > 8502)'  output28-APR-19.csv | wc -l
1662470
```

Concluimos que los datos están limpios y no presentan errores

SCRIP EN PYTHON

```python
import pandas as pd
import plotly.express as px

# Ruta al archivo CSV
archivo_csv = 'output28-APR-19.csv'

# Lee el archivo CSV en un DataFrame de pandas
df = pd.read_csv(archivo_csv, header=None)

# Convertir la columna 10 a datetime
df[10] = pd.to_datetime(df[10], format='%d-%b-%y %I.%M.%S.%f %p')

# Establecer la columna 10 como el índice
df.set_index(10, inplace=True)

# Resample the data by minute and count the number of occurrences
datagrams_per_minute = df.resample('T').size().reset_index(name='Count')

# Utilizar Plotly Express para crear un gráfico de barras
fig = px.bar(datagrams_per_minute, x=10, y='Count', labels={'10': 'Time (Hour)',
'Count': 'Number of Datagrams'},
             title='Histogram of Datagrams Generated by Minute')

# Mostrar el gráfico
fig.show()
```

HISTOGRAMA POR MINUTO


Histogram of Datagrams Generated by Minute
Number of Datagrams
1500
1000
500
0
00:00
Apr 28, 2019
03:00
06:00
09:00
12:00
15:00
18:00
21:00
Time (Hour)