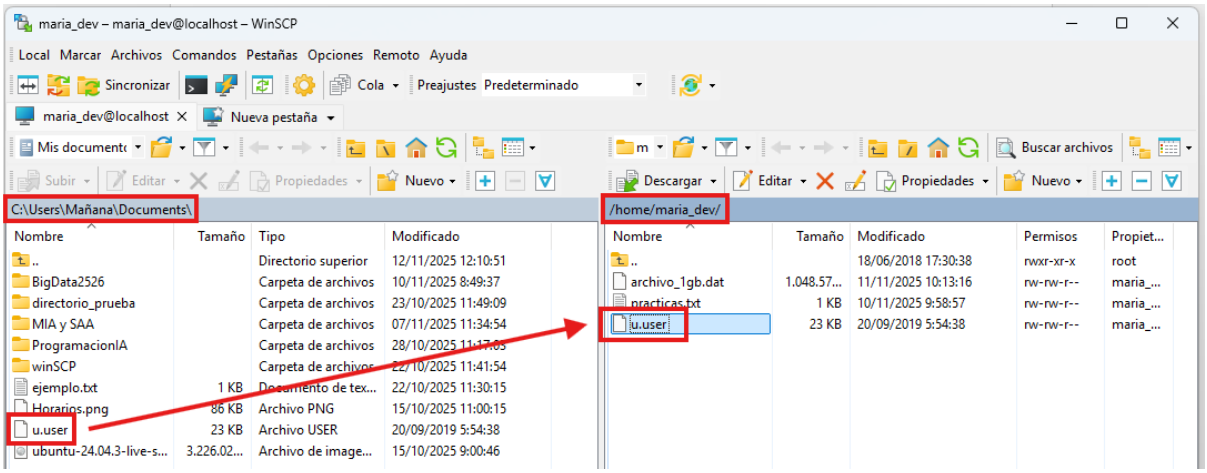
PR_03.1 Dani Gayol Rodríguez

Apartado A

El primer paso va a ser pegar el archivo “u.user” en la maquina virtual y para ello use “WinSCP”



Luego vamos a pasar este archivo de la ruta “/home/maria_dev” en local de Linux, hacia “/user/maria_dev” en HDFS

```
maria_dev@sandbox-hdp:~
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat  practicas.txt  u.user
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 2 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir /user/maria_dev/practica_pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 3 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 08:09 /user/maria_dev/practica_pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/u.user /user/maria_dev/practica_pig/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/practica_pig
Found 1 items
-rw-r--r--   1 maria_dev hdfs      22628 2025-11-13 08:10 /user/maria_dev/practica_pig/u.user
```

1.) Muestra el total de hombres y mujeres que hay en el archivo u.user.

Usamos el comando “pig” para empezar y cargamos los datos

```
maria_dev@sandbox-hdp:~
[maria_dev@sandbox-hdp ~]$ pig
25/11/13 08:20:36 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
25/11/13 08:20:36 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
25/11/13 08:20:36 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
25/11/13 08:20:36 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
25/11/13 08:20:36 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2025-11-13 08:20:36,816 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory
) compiled May 11 2018, 07:56:28
2025-11-13 08:20:36,816 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/maria_dev/pig_1763022036813.
log
2025-11-13 08:20:36,849 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/maria_dev/.pigbootup not
 found
2025-11-13 08:20:37,332 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2025-11-13 08:20:38,076 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-41129240-d556
-4726-bd08-b17697c56358
2025-11-13 08:20:38,343 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service addres
s: http://sandbox-hdp.hortonworks.com:8188/ws/v1/timeline/
2025-11-13 08:20:38,455 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> usuarios = LOAD '/user/maria_dev/practica_pig/u.user' USING PigStorage('|') AS (user_id:int, age:int, gender:char
array, occupation:chararray, zipcode:chararray);
```

Ahora vamos a realizar la consulta

```
grunt> grupo_genero = GROUP usuarios BY gender; conteo_genero = FOREACH grupo_genero GENERATE group AS gender, COUNT(usu
arios) AS total; DUMP conteo_genero;
```

```
(F,273)
(M,670)
```

2.) Mediante instrucciones de PIG encontrar las 10 ocupaciones más frecuentes entre los usuarios.

```
grunt> grupo_ocup = GROUP usuarios BY occupation; conteo_ocup = FOREACH grupo_ocup GENERATE group AS occupation, COUNT(u
suarios) AS total; orden_ocup = ORDER conteo_ocup BY total DESC; top10_ocup = LIMIT orden_ocup 10; DUMP top10_ocup;
```

```
(student,196)
(other,105)
(educator,95)
(administrator,79)
(engineer,67)
(programmer,66)
(librarian,51)
(writer,45)
(executive,32)
(scientist,31)
```

3.) Muestra la edad media por géneros.

```
grunt> grupo_genero2 = GROUP usuarios BY gender; edad_media_genero = FOREACH grupo_genero2 GENERATE group AS gender, AVG
(usuarios.age) AS avg_age; DUMP edad_media_genero;
```

```
(F,33.81318681318681)
(M,34.149253731343286)
```

4.) Muestra la edad media por ocupaciones.

```
grunt> grupo_ocup2 = GROUP usuarios BY occupation; edad_media_ocup = FOREACH grupo_ocup2 GENERATE group AS occupation, A
VG(usuarios.age) AS avg_age; DUMP edad_media_ocup;
```

```
(none,26.555555555555557)
(other,34.523809523809526)
(artist,31.392857142857142)
(doctor,43.57142857142857)
(lawyer,36.75)
(writer,36.31111111111111)
(retired,63.071428571428571)
(student,22.081632653061224)
(educator,42.010526315789048)
(engineer,36.38805970149254)
(salesman,35.666666666666664)
(executive,38.71875)
(homemaker,32.571428571428571)
(librarian,40.0)
(marketing,37.61538461538461)
(scientist,35.54838709677419)
(healthcare,41.5625)
(programmer,33.121212121212125)
(technician,33.1481481481481145)
(administrator,38.746835443037977)
(entertainment,29.222222222222222)
```

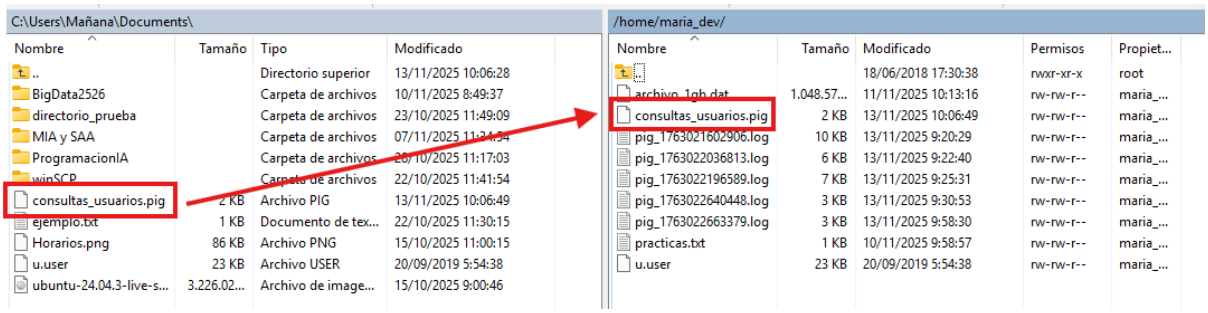
5.) Guarda el resultado de las cuatro consultas anteriores en un script de extensión “.pig”. Ejecútalo. (recuerda, siempre en la carpeta /user/maría_dev)

Como en la maquina no puedo usar el comando “nano” voy a tener que crear el archivo desde Windows y pasarlo a la máquina virtual

consultas_usuarios.pig

Archivo Editar Ver

```
-- Cargar datos
usuarios = LOAD '/user/maria_dev/practica_pig/u.user' USING PigStorage('|') AS (user_id:int, age:int, gender:chararray, occupation:chararray, zipcode:chararray);

-- 1. Total de hombres y mujeres
grupo_genero = GROUP usuarios BY gender;
conteo_genero = FOREACH grupo_genero GENERATE group AS gender, COUNT(usuarios) AS total;
STORE conteo_genero INTO '/user/maria_dev/pig_usuarios/conteo_genero' USING PigStorage(',');

-- 2. 10 ocupaciones más frecuentes
grupo_ocupacion = GROUP usuarios BY occupation;
conteo_ocupacion = FOREACH grupo_ocupacion GENERATE group AS occupation, COUNT(usuarios) AS total;
orden_ocupacion = ORDER conteo_ocupacion BY total DESC;
top10_ocupacion = LIMIT orden_ocupacion 10;
STORE top10_ocupacion INTO '/user/maria_dev/pig_usuarios/top10_ocupacion' USING PigStorage(',');

-- 3. Edad media por género
edad_genero = FOREACH grupo_genero GENERATE group AS gender, AVG(usuarios.age) AS edad_media;
STORE edad_genero INTO '/user/maria_dev/pig_usuarios/edad_genero' USING PigStorage(',');

-- 4. Edad media por ocupación
edad_ocupacion = FOREACH grupo_ocupacion GENERATE group AS occupation, AVG(usuarios.age) AS edad_media;
STORE edad_ocupacion INTO '/user/maria_dev/pig_usuarios/edad_ocupacion' USING PigStorage(',');
```



Ahora vamos a pasarlo a de Linux a HDFS el archivo

```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         pig_1763021602906.log  pig_1763022196589.log  pig_1763022663379.log  u.user
consultas_usuarios.pig  pig_1763022036813.log  pig_1763022640448.log  practicas.txt
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/consultas_usuarios.pig /user/maria_dev/practica_pig/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/practica_pig/
Found 2 items
-rw-r--r--   1 maria_dev hdfs       1282 2025-11-13 09:10 /user/maria_dev/practica_pig/consultas_usuarios.pig
-rw-r--r--   1 maria_dev hdfs      22628 2025-11-13 08:10 /user/maria_dev/practica_pig/u.user
```

Finalmente vamos a ejecutarlo

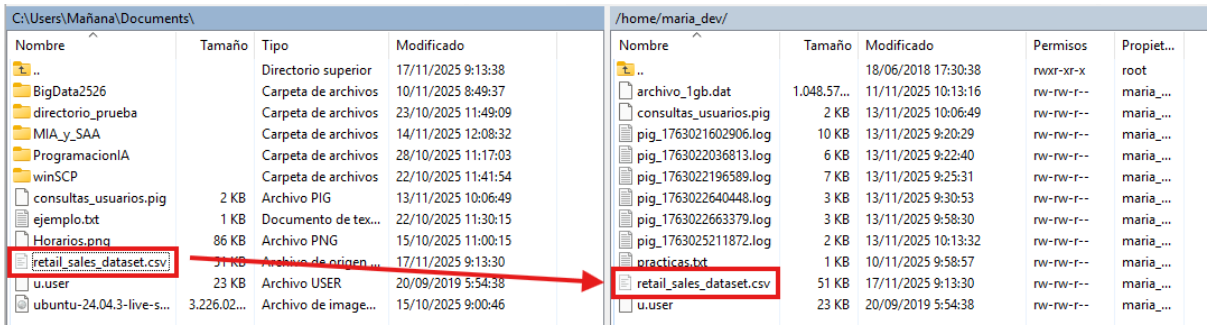
```
2025-11-13 09:15:28,172 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Su
ccess!
2025-11-13 09:15:28,185 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 27 seconds and 52 milliseco
nds (87052 ms)
```

6.) Almacena la salida de las cuatro consultas anteriores en una carpeta de HDFS llamada pig_usuarios.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/pig_usuarios/
Found 4 items
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:14 /user/maria_dev/pig_usuarios/conteo_genero
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:14 /user/maria_dev/pig_usuarios/edad_genero
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:14 /user/maria_dev/pig_usuarios/edad_ocupacion
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:15 /user/maria_dev/pig_usuarios/top10_ocupacion
```

Apartado B

Primero vamos a pasar el archivo “.csv” a la máquina virtual




```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat          pig_1763022036813.log  pig_1763022663379.log  practicas.txt
consultas_usuarios.pig   pig_1763022196589.log  pig_1763025211872.log  retail_sales_dataset.csv
pig_1763021602906.log    pig_1763022640448.log  pig_1763366949543.log  u.user
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 5 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-13 09:15 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:15 /user/maria_dev/pig_usuarios
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:10 /user/maria_dev/practica_pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/retail_sales_dataset.csv /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 6 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-13 09:15 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:15 /user/maria_dev/pig_usuarios
drwxr-xr-x   - maria_dev hdfs          0 2025-11-13 09:10 /user/maria_dev/practica_pig
-rw-r--r--   1 maria_dev hdfs      51673 2025-11-17 08:17 /user/maria_dev/retail_sales_dataset.csv
```

1.) Carga y descripción del dataset

Para cargar el dataset junto a los datos hacemos lo siguiente

```
[maria_dev@sandbox-hdp ~]$ pig
25/11/17 08:19:34 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
25/11/17 08:19:34 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
25/11/17 08:19:34 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
25/11/17 08:19:34 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
25/11/17 08:19:34 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2025-11-17 08:19:34,385 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory
) compiled May 11 2018, 07:56:28
2025-11-17 08:19:34,385 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/maria_dev/pig_1763367574384.
log
2025-11-17 08:19:34,422 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/maria_dev/.pigbootup not
 found
2025-11-17 08:19:35,062 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2025-11-17 08:19:35,586 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-1a8b3d30-cf8f
-4db0-be5b-ce901a525367
2025-11-17 08:19:35,850 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service addres
s: http://sandbox-hdp.hortonworks.com:8188/ws/v1/timeline/
2025-11-17 08:19:35,962 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> ventas = LOAD '/user/maria_dev/retail_sales_dataset.csv' USING PigStorage(',') AS (transaction_id:chararray, date
:chararray, customer_id:chararray, age:int, gender:chararray, product_id:chararray, product_category:chararray, price_pe
r_unit:double, quantity:int, total_amount:double);
```

Y ahora para verlo hacemos esto;

Usando DESCRIBE:

```
grunt> DESCRIBE ventas;
ventas: {transaction_id: chararray,date: chararray,customer_id: chararray,age: int,gender: chararray,product_id: chararr
ay,product_category: chararray,price_per_unit: double,quantity: int,total_amount: double}
```

Usando DUMP y limitándolo a 10 ya que hay demasiados valores;

```
grunt> ventas_sample = LIMIT ventas 10; DUMP ventas_sample;
```

```
(Transaction ID,Date,Customer ID,,Age,Product Category,Quantity,,,)
(1,2023-11-24,CUST001,,34,Beauty,3,50.0,150,)
(2,2023-02-27,CUST002,,26,Clothing,2,500.0,1000,)
(3,2023-01-13,CUST003,,50,Electronics,1,30.0,30,)
(4,2023-05-21,CUST004,,37,Clothing,1,500.0,500,)
(5,2023-05-06,CUST005,,30,Beauty,2,50.0,100,)
(6,2023-04-25,CUST006,,45,Beauty,1,30.0,30,)
(7,2023-03-13,CUST007,,46,Clothing,2,25.0,50,)
(8,2023-02-22,CUST008,,30,Electronics,4,25.0,100,)
(9,2023-12-13,CUST009,,63,Electronics,2,300.0,600,)
```

Finalmente vamos a calcular las transacciones totales

```
grunt> ventas_all = GROUP ventas ALL; total_count = FOREACH ventas_all GENERATE COUNT(ventas) AS total; DUMP total_count
```

```
(1001)
```

2.) Filtrado por rango de edad

Filtrado de clientes mayores de 30 años

```
grunt> clientes_mayores30 = FILTER ventas BY age > 30;
```

Y luego usamos el LIMIT para limitarlo a 10 valores y visualizarlo pero, debido a un error en el formato del dato "age" no puedo visualizar este filtrado

3.) Transformación de campos

Hacemos el UPPER y la operación

```
grunt> transformados = FOREACH ventas GENERATE transaction_id, date, customer_id, age, (UPPER(gender)) AS gender_upper,
product_id,  product_category, price_per_unit, quantity, total_amount, (price_per_unit * quantity * 0.90) AS importe_des
cuento;
```

Resultado:

```
(Transaction ID,Date,Customer ID,,AGE,Product Category,Quantity,,,,)
(1,2023-11-24,CUST001,,34,Beauty,3,50.0,150,,6750.0)
(2,2023-02-27,CUST002,,26,Clothing,2,500.0,1000,,450000.0)
(3,2023-01-13,CUST003,,50,Electronics,1,30.0,30,,810.0)
(4,2023-05-21,CUST004,,37,Clothing,1,500.0,500,,225000.0)
(5,2023-05-06,CUST005,,30,Beauty,2,50.0,100,,4500.0)
(6,2023-04-25,CUST006,,45,Beauty,1,30.0,30,,810.0)
(7,2023-03-13,CUST007,,46,Clothing,2,25.0,50,,1125.0)
(8,2023-02-22,CUST008,,30,Electronics,4,25.0,100,,2250.0)
(9,2023-12-13,CUST009,,63,Electronics,2,300.0,600,,162000.0)
(10,2023-10-07,CUST010,,52,Clothing,4,50.0,200,,9000.0)
(11,2023-02-14,CUST011,,23,Clothing,2,50.0,100,,4500.0)
(12,2023-10-30,CUST012,,35,Beauty,3,25.0,75,,1687.5)
(13,2023-08-05,CUST013,,22,Electronics,3,500.0,1500,,675000.0)
(14,2023-01-17,CUST014,,64,Clothing,4,30.0,120,,3240.0)
(15,2023-01-16,CUST015,,42,Electronics,4,500.0,2000,,900000.0)
(16,2023-02-17,CUST016,,19,Clothing,3,500.0,1500,,675000.0)
(17,2023-04-22,CUST017,,27,Clothing,4,25.0,100,,2250.0)
(18,2023-04-30,CUST018,,47,Electronics,2,25.0,50,,1125.0)
(19,2023-09-16,CUST019,,62,Clothing,2,25.0,50,,1125.0)
```

4.) Agrupación y agregación por categoría de producto

Primero hacemos la agrupación

```
grunt> grp_cat = GROUP ventas BY product_category;
```

A continuación, hacemos la agregación por categoria del producto y lo ordenamos el resultado descendente

```
grunt> agg_cat = FOREACH grp_cat GENERATE group                          AS product_category, COUNT(ventas)
     AS num_transacciones, SUM(ventas.total_amount)       AS suma_total_amount, AVG(ventas.age)                AS edad_p
romedio;
```

```
grunt> agg_cat_sorted = ORDER agg_cat BY suma_total_amount DESC; DUMP agg_cat_sorted;
```

Resultado:

```
(1,253,,)
(2,243,,)
(3,241,,)
(4,263,,)
(Quantity,1,,)
```

5.) Extracción de categorías distintas

Extraemos las categorías de producto distintas

```
grunt> only_categories = FOREACH ventas GENERATE product_category; distinct_categories = DISTINCT only_categories;
```

```
(1)
(2)
(3)
(4)
(Quantity)
```

Y ahora vemos las categorías diferentes que hay

```
grunt> cnt_cat = FOREACH (GROUP distinct_categories ALL) GENERATE COUNT(distinct_categories) AS num_categorias;
```

```
(5)
```

6.) Ordenación y obtención de top-transacciones

Ahora vamos a ordenar las transacciones y limitarlo para extraer solo 5 transacciones

```
grunt> ordered_by_amount = ORDER ventas BY total_amount DESC;
grunt> top5 = LIMIT ordered_by_amount 5;
grunt> top5_select = FOREACH top5 GENERATE transaction_id, customer_id, product_category, total_amount;
```

Ahora teniendo esto ya ordenado, vamos a mostrar el resultado

```
(Transaction ID,Customer ID,Quantity,)
(1,CUST001,3,)
(2,CUST002,2,)
(3,CUST003,1,)
(4,CUST004,1,)
```

7.) Uso de funciones de cadena

Añadimos la nueva columna

```
grunt> cadenas = FOREACH ventas GENERATE transaction_id, product_category, SUBSTRING(product_category, 0, 3) AS categori
a_3char, SIZE(product_category)          AS len_categoria, customer_id, total_amount;
```

Y ahora vamos a mostrar los 15 primeros resultados

```
grunt> cadenas_sample = LIMIT cadenas 15; DUMP cadenas_sample;
```

```
(Transaction ID,Quantity,Qua,8,Customer ID,)
(1,3,3,1,CUST001,)
(2,2,2,1,CUST002,)
(3,1,1,1,CUST003,)
(4,1,1,1,CUST004,)
(5,2,2,1,CUST005,)
(6,1,1,1,CUST006,)
(7,2,2,1,CUST007,)
(8,4,4,1,CUST008,)
(9,2,2,1,CUST009,)
(10,4,4,1,CUST010,)
(11,2,2,1,CUST011,)
(12,3,3,1,CUST012,)
(13,3,3,1,CUST013,)
(14,4,4,1,CUST014,)
```

8.) Filtrado por fecha y condiciones combinadas

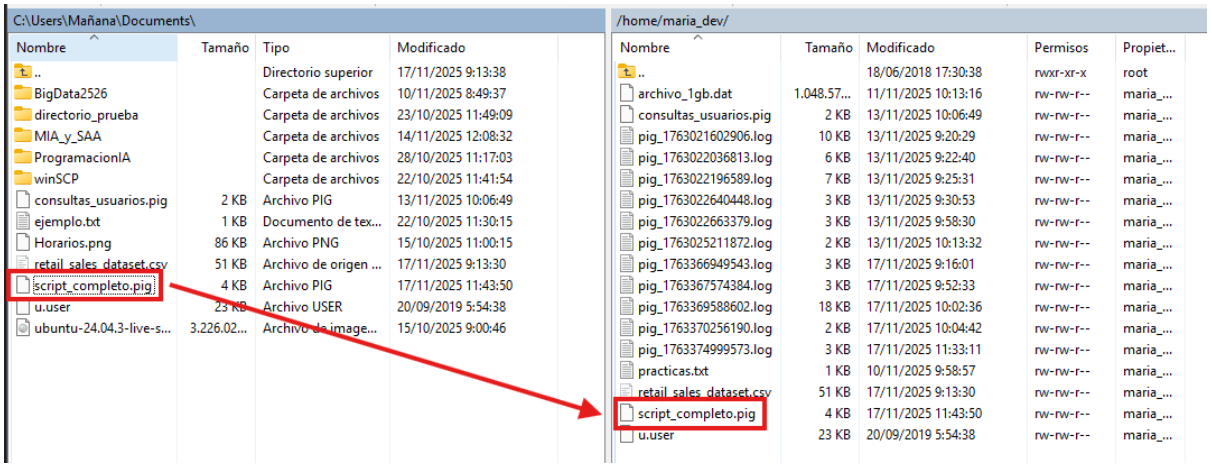
Vamos a intentar filtrar las transacciones que se realizaron antes de una determinada fecha y después de esto, filtrarlas para que sean mayores de 500

```
grunt> antes_fecha = FILTER ventas BY date < '2023-07-01';
grunt> antes_y_mayor500 = FILTER antes_fecha BY total_amount > 500;
```

No se muestran resultados ya que en él “.csv” la fecha viene en un formato diferente y nos da un error

9.) Script completo + almacenamiento

Ahora finalmente vamos a crear el script completo



```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         pig_1763022196589.log  pig_1763366949543.log  pig_1763374999573.log     u.user
consultas_usuarios.pig  pig_1763022640448.log  pig_1763367574384.log  practicas.txt
pig_1763021602906.log   pig_1763022663379.log  pig_1763369588602.log  retail_sales_dataset.csv
pig_1763022036813.log   pig_1763025211872.log  pig_1763370256190.log  script_completo.pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_completo.pig /user/maria_dev/ventas_analisis/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/ventas_analisis/
Found 1 items
-rw-r--r--   1 maria_dev hdfs       3156 2025-11-17 10:48 /user/maria_dev/ventas_analisis/script_completo.pig
```

Una vez puesto el script en el directorio correcto, vamos a ejecutarlo con el siguiente comando;

```
[maria_dev@sandbox-hdp ~]$ pig -x mapreduce script_completo.pig
```

```
Output(s):
Successfully stored 1001 records (56266 bytes) in: "/user/maria_dev/ventas_analisis/transformados"
Successfully stored 5 records (17 bytes) in: "/user/maria_dev/ventas_analisis/distinct_categories"
Successfully stored 5 records (77 bytes) in: "/user/maria_dev/ventas_analisis/top5"
Successfully stored 5 records (45 bytes) in: "/user/maria_dev/ventas_analisis/agg_categoria"

2025-11-17 10:55:31,953 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 44 seconds and 59 milliseco
nds (104059 ms)
```

Ver los resultados:

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/ventas_analisis
Found 5 items
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis/agg_categoria
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:54 /user/maria_dev/ventas_analisis/distinct_categories
-rw-r--r--   1 maria_dev hdfs       3008 2025-11-17 10:53 /user/maria_dev/ventas_analisis/script_completo.pig
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis/top5
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:54 /user/maria_dev/ventas_analisis/transformados
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/ventas_analisis/agg_categoria/part-*
Quantity,1,,
4,263,,
3,241,,
2,243,,
1,253,,
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/ventas_analisis/top5/part-*
996,CUST996,1,
997,CUST997,3,
998,CUST998,4,
999,CUST999,3,
1000,CUST1000,4,
```

Apartado C

1.) Localiza en Internet una versión del Quijote en formato texto. Descárgala y cópiala a en tu sistema HDFS

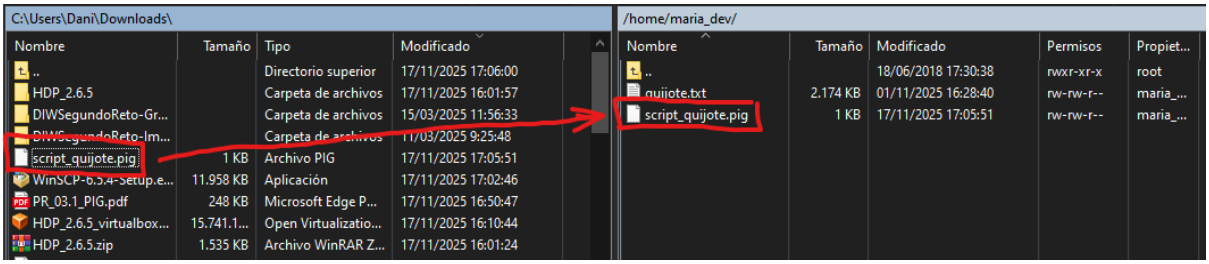
Vamos a descargar el “Quijote”

```
maria_dev@sandbox-hdp:~
[maria_dev@sandbox-hdp ~]$ wget https://www.gutenberg.org/cache/epub/2000/pg2000.txt -O quijote.txt
--2025-11-17 15:55:24--  https://www.gutenberg.org/cache/epub/2000/pg2000.txt
Resolving www.gutenberg.org (www.gutenberg.org)... 152.19.134.47, 2610:28:3090:3000:0:bad:cafe:47
Connecting to www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2225845 (2.1M) [text/plain]
Saving to: ‘quijote.txt’

100%[======================================================================>] 2,225,845   1.64MB/s   in 1.3s

2025-11-17 15:55:26 (1.64 MB/s) - ‘quijote.txt’ saved [2225845/2225845]
```

Ahora una vez descargado vamos a copiarlo en el directorio de “/user/maría_dev”

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put quijote.txt /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 1 items
-rw-r--r--   1 maria_dev hdfs    2225845 2025-11-17 15:56 /user/maria_dev/quijote.txt
```

A continuación, vamos a hacer el script




```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
quijote.txt  script_quijote.pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put script_quijote.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 2 items
-rw-r--r--   1 maria_dev hdfs    2225845 2025-11-17 15:56 /user/maria_dev/quijote.txt
-rw-r--r--   1 maria_dev hdfs        697 2025-11-17 16:08 /user/maria_dev/script_quijote.pig
```

A la hora de ejecutar el script me da un error ya que detecta caracteres en blanco y salta un error

```
[maria_dev@sandbox-hdp ~]$ pig script_quijote.pig
25/11/17 16:31:43 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
25/11/17 16:31:43 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
25/11/17 16:31:43 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
25/11/17 16:31:43 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
25/11/17 16:31:43 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2025-11-17 16:31:43,225 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory
) compiled May 11 2018, 07:56:28
2025-11-17 16:31:43,225 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/maria_dev/pig_1763397103224.
log
2025-11-17 16:31:43,631 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/maria_dev/.pigbootup not
 found
2025-11-17 16:31:43,702 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2025-11-17 16:31:43,988 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-script_quijote.pig-ca
b36059-48be-4282-a7bf-b9c4c3f3fcf2
2025-11-17 16:31:44,199 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service addres
s: http://sandbox-hdp.hortonworks.com:8188/ws/v1/timeline/
2025-11-17 16:31:44,265 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2025-11-17 16:31:44,700 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of siz
e 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2025-11-17 16:31:44,730 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1000: Error during parsing. Lexical error
at line 8, column 0.  Encountered: <EOF> after : ""
Details at logfile: /home/maria_dev/pig_1763397103224.log
2025-11-17 16:31:44,741 [main] INFO  org.apache.pig.Main - Pig script completed in 1 second and 630 milliseconds (1630 m
s)
```