Lenguaje Pig

1) APARTADO A

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

Creamos el script

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > count_gender.pig <<'EOF'
>
>
>
> users = LOAD '/user/maria_dev/ml-100k/u.user'
>
>     USING PigStorage('|')
>
>     AS (user_id:int, age:int, gender:chararray, occupation:chararray, zip:chararray);
>
>
>
> grp_genero = GROUP users BY gender;
>
> total_genero = FOREACH grp_genero GENERATE group AS gender, COUNT(users);
>
>
>
> ordenado = ORDER total_genero BY gender;
>
>
>
> STORE ordenado INTO '/user/maria_dev/resultados/genero' USING PigStorage(',');
>
> EOF
```

Resultado:

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/genero/part-*
F,273
M,670
```

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

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > top10_ocupaciones.pig <<'EOF'
>
> users = LOAD '/user/maria_dev/ml-100k/u.user'
>
>     USING PigStorage('|')
>
>     AS (
>
>         user_id:int,
>
>         age:int,
>
>         gender:chararray,
>
>         occupation:chararray,
>
>         zip:chararray
>
>     );
>
>
>
> grp_ocup = GROUP users BY occupation;
>
> ocup_count = FOREACH grp_ocup GENERATE group AS occupation, COUNT(users) AS total;
>
> ordenado = ORDER ocup_count BY total DESC;
>
> top10 = LIMIT ordenado 10;
>
>
>
> STORE top10 INTO '/user/maria_dev/resultados/top10_ocupaciones'
>
>     USING PigStorage(',');
>
> EOF
```

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/top10_ocupaciones/part-*
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.

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > edad_media_genero.pig <<'EOF'
>
> users = LOAD '/user/maria_dev/ml-100k/u.user'
>
>     USING PigStorage('|')
>
>     AS (
>
>         user_id:int,
>
>         age:int,
>
>         gender:chararray,
>
>         occupation:chararray,
>
>         zip:chararray
>
>     );
>
>
>
> -- Agrupar por género (M, F)
>
> grp_genero = GROUP users BY gender;
>
>
>
> -- Calcular edad media por género
>
> edad_media = FOREACH grp_genero GENERATE
>
>     group AS genero,
>
>     AVG(users.age) AS edad_promedio;
>
>
>
> -- Guardar resultado en HDFS
>
> STORE edad_media INTO '/user/maria_dev/resultados/edad_media_genero'
>
>     USING PigStorage(',');
>
> EOF
```

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/edad_media_genero/part-*
F,33.81318681318681
M,34.149253731343286
```

4) Muestra la edad media por ocupaciones.

```
[maria dev@sandbox-hdp scripts pig]$ cat > edad media ocupaciones.pig <<'EOF'
> users = LOAD '/user/maria dev/ml-100k/u.user'
>     USING PigStorage('|')
>     AS (
>         user id:int,
>         age:int,
>         gender:chararray,
>         occupation:chararray,
>         zip:chararray
>     );
> grp ocup = GROUP users BY occupation;
> edad media ocup = FOREACH grp ocup GENERATE
>     group AS ocupacion,
>     AVG(users.age) AS edad promedio;
> ordenado = ORDER edad media ocup BY edad promedio DESC;
> STORE ordenado INTO '/user/maria dev/resultados/edad media ocupaciones'
>     USING PigStorage(',');
> EOF
```

```
[maria dev@sandbox-hdp scripts pig]$ hdfs dfs -cat /user/maria dev/resultados/edad media ocupaciones/part-*
retired,63.07142857142857
doctor,43.57142857142857
educator,42.01052631578948
healthcare,41.5625
librarian,40.0
administrator,38.74683544303797
executive,38.71875
marketing,37.61538461538461
lawyer,36.75
engineer,36.38805970149254
writer,36.31111111111111
salesman,35.666666666666664
scientist,35.54838709677419
other,34.523809523809526
technician,33.148148148148145
programmer,33.121212121212125
homemaker,32.57142857142857
artist,31.392857142857142
entertainment,29.22222222222222
none,26.555555555555557
student,22.081632653061224
```

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

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > cuatro_consultas.pig <<'EOF'
>
> users = LOAD '/user/maria_dev/ml-100k/u.user'
>
>     USING PigStorage('|')
>
>     AS (
>
>         user_id:int,
>
>         age:int,
>
>         gender:chararray,
>
>         occupation:chararray,
>
>         zip:chararray
>
>     );
>
>
>
> grp_genero = GROUP users BY gender;
>
> total_genero = FOREACH grp_genero GENERATE group, COUNT(users);
>
> STORE total_genero INTO '/user/maria_dev/resultados/consulta1_total_genero'
>
>     USING PigStorage(',');
>
>
>
> grp_ocup = GROUP users BY occupation;
>
> ocup_count = FOREACH grp_ocup GENERATE group, COUNT(users);
>
> orden_ocup = ORDER ocup_count BY $1 DESC;
>
> top10 = LIMIT orden_ocup 10;
>
> STORE top10 INTO '/user/maria_dev/resultados/consulta2_top10_ocupaciones'
>
>     USING PigStorage(',');
>
>
>
> grp_genero2 = GROUP users BY gender;
>
> edad_media_genero = FOREACH grp_genero2 GENERATE group, AVG(users.age);
>
> STORE edad_media_genero INTO '/user/maria_dev/resultados/consulta3_edad_media_genero
>
>     USING PigStorage(',');
>
>
>
> grp_ocup2 = GROUP users BY occupation;
>
> edad_media_ocup = FOREACH grp_ocup2 GENERATE group, AVG(users.age);
>
> orden_edad = ORDER edad_media_ocup BY $1 DESC;
>
> STORE orden_edad INTO '/user/maria_dev/resultados/consulta4_edad_media_ocupaciones'
>
>     USING PigStorage(',');
>
> EOF
```

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

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > cuatro_consultas_pig_usuarios.pig <<'EOF'
>
> users = LOAD '/user/maria_dev/ml-100k/u.user'
>
>     USING PigStorage('|')
>
>     AS (
>
>         user_id:int,
>
>         age:int,
>
>         gender:chararray,
>
>         occupation:chararray,
>
>         zip:chararray
>
>     );
>
>
>
> grp_genero = GROUP users BY gender;
>
> total_genero = FOREACH grp_genero GENERATE group, COUNT(users);
>
> STORE total_genero INTO '/user/maria_dev/pig_usuarios/consulta1_total_genero'
>
>     USING PigStorage(',');
>
>
>
> grp_ocup = GROUP users BY occupation;
>
> ocup_count = FOREACH grp_ocup GENERATE group, COUNT(users);
>
> orden_ocup = ORDER ocup_count BY $1 DESC;
>
> top10 = LIMIT orden_ocup 10;
>
> STORE top10 INTO '/user/maria_dev/pig_usuarios/consulta2_top10_ocupaciones'
>
>     USING PigStorage(',');
>
>
>
> grp_genero2 = GROUP users BY gender;
>
> edad_media_genero = FOREACH grp_genero2 GENERATE group, AVG(users.age);
>
> STORE edad_media_genero INTO '/user/maria_dev/pig_usuarios/consulta3_edad_media_genero'
>
>     USING PigStorage(',');
>
>
>
> grp_ocup2 = GROUP users BY occupation;
>
> edad_media_ocup = FOREACH grp_ocup2 GENERATE group, AVG(users.age);
>
> orden_edad = ORDER edad_media_ocup BY $1 DESC;
>
> STORE orden_edad INTO '/user/maria_dev/pig_usuarios/consulta4_edad_media_ocupaciones'
>
>     USING PigStorage(',');
>
> EOF
```

Tras ejecutarlo comprobamos los directorios

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -ls /user/maria_dev/pig_usuarios
Found 4 items
drwxr-xr-x   - maria_dev hdfs          0 2025-11-14 20:10 /user/maria_dev/pig_usuarios/consulta1_total_genero
drwxr-xr-x   - maria_dev hdfs          0 2025-11-14 20:10 /user/maria_dev/pig_usuarios/consulta2_top10_ocupaciones
drwxr-xr-x   - maria_dev hdfs          0 2025-11-14 20:10 /user/maria_dev/pig_usuarios/consulta3_edad_media_genero
drwxr-xr-x   - maria_dev hdfs          0 2025-11-14 20:10 /user/maria_dev/pig_usuarios/consulta4_edad_media_ocupaciones

[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/pig_usuarios/consulta1_total_genero/part-*
F,273
M,670
```

2) APARTADO B

1) Carga y descripción del dataset
    i. Carga el archivo (retail_sales_dataset.csv) usando PigStorage(',') y define un esquema correctamente para cada tipo de campo.

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > retail_sales_ej1.pig <<'EOF'
>
> sales = LOAD '/user/maria dev/retail sales/retail sales dataset.csv'
>
>     USING PigStorage(',')
>
>     AS (
>
>         transaction_id:int,
>
>         date:chararray,
>
>         customer_id:chararray,
>
>         gender:chararray,
>
>         age:int,
>
>         product category:chararray,
>
>         quantity:int,
>
>         price_per_unit:double,
>
>         total_amount:double
>
>     );
>
>
>
> DESCRIBE sales;
>
>
>
> limite = LIMIT sales 10;
>
> DUMP limite;
>
>
>
> grp_all = GROUP sales ALL;
>
> total transacciones = FOREACH grp all GENERATE COUNT(sales) AS total;
>
>
>
> STORE total_transacciones INTO '/user/maria_dev/resultados/retail_sales/total_transacciones'
>
>     USING PigStorage(',');
>
> EOF
```

ii. Usa DESCRIBE para ver el esquema y DUMP para ver las primeras tuplas.

```
(,Date,Customer ID,Gender,,Product Category,,,)
(1,2023-11-24,CUST001,Male,34,Beauty,3,50.0,150.0)
(2,2023-02-27,CUST002,Female,26,Clothing,2,500.0,1000.0)
(3,2023-01-13,CUST003,Male,50,Electronics,1,30.0,30.0)
(4,2023-05-21,CUST004,Male,37,Clothing,1,500.0,500.0)
(5,2023-05-06,CUST005,Male,30,Beauty,2,50.0,100.0)
(6,2023-04-25,CUST006,Female,45,Beauty,1,30.0,30.0)
(7,2023-03-13,CUST007,Male,46,Clothing,2,25.0,50.0)
(8,2023-02-22,CUST008,Male,30,Electronics,4,25.0,100.0)
(9,2023-12-13,CUST009,Male,63,Electronics,2,300.0,600.0)
```

iii. Calcula cuántas transacciones totales tiene el dataset (COUNT).

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/total_transacciones/part
1000
```

2) Filtrado por rango de edad

i. Filtra los clientes con edad mayor de 30 años y guarda en alias clientes_mayores30.

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > retail_sales_ej2.pig <<'EOF'
>
> sales = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
>
>     USING PigStorage(',')
>
>     AS (
>
>         transaction_id:int,
>
>         date:chararray,
>
>         customer_id:chararray,
>
>         gender:chararray,
>
>         age:int,
>
>         product_category:chararray,
>
>         quantity:int,
>
>         price_per_unit:double,
>
>         total_amount:double
>
>     );
>
>
>
> clientes_mayores30 = FILTER sales BY age > 30;
>
>
>
> limite = LIMIT clientes_mayores30 10;
>
> DUMP limite;
>
>
>
> grp_all = GROUP sales ALL;
>
> total_transacciones = FOREACH grp_all GENERATE COUNT(sales) AS total;
>
>
>
> grp_mayores = GROUP clientes_mayores30 ALL;
>
> total_mayores = FOREACH grp_mayores GENERATE COUNT(clientes_mayores30) AS total_mayores;
>
>
>
> porcentaje = FOREACH (CROSS total_mayores, total_transacciones)
>
>     GENERATE (total_mayores.total_mayores * 100.0) / total_transacciones.total AS porcentaje;
>
>
>
> STORE clientes_mayores30 INTO '/user/maria_dev/resultados/retail_sales/clientes_mayores30'
>
>     USING PigStorage(',');
>
>
>
> STORE porcentaje INTO '/user/maria_dev/resultados/retail_sales/porcentaje_mayores30'
>
>     USING PigStorage(',');
>
> EOF
```

```
(1,2023-11-24,CUST001,Male,34,Beauty,3,50.0,150.0)
(3,2023-01-13,CUST003,Male,50,Electronics,1,30.0,30.0)
(4,2023-05-21,CUST004,Male,37,Clothing,1,500.0,500.0)
(6,2023-04-25,CUST006,Female,45,Beauty,1,30.0,30.0)
(7,2023-03-13,CUST007,Male,46,Clothing,2,25.0,50.0)
(9,2023-12-13,CUST009,Male,63,Electronics,2,300.0,600.0)
(10,2023-10-07,CUST010,Female,52,Clothing,4,50.0,200.0)
(12,2023-10-30,CUST012,Male,35,Beauty,3,25.0,75.0)
(14,2023-01-17,CUST014,Male,64,Clothing,4,30.0,120.0)
(15,2023-01-16,CUST015,Female,42,Electronics,4,500.0,2000.0)
```

ii. Utiliza LIMIT para ver los primeros 10 resultados.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/clientes_mayores30/part-*
1,2023-11-24,CUST001,Male,34,Beauty,3,50.0,150.0
3,2023-01-13,CUST003,Male,50,Electronics,1,30.0,30.0
4,2023-05-21,CUST004,Male,37,Clothing,1,500.0,500.0
6,2023-04-25,CUST006,Female,45,Beauty,1,30.0,30.0
7,2023-03-13,CUST007,Male,46,Clothing,2,25.0,50.0
9,2023-12-13,CUST009,Male,63,Electronics,2,300.0,600.0
10,2023-10-07,CUST010,Female,52,Clothing,4,50.0,200.0
12,2023-10-30,CUST012,Male,35,Beauty,3,25.0,75.0
14,2023-01-17,CUST014,Male,64,Clothing,4,30.0,120.0
15,2023-01-16,CUST015,Female,42,Electronics,4,500.0,2000.0
18,2023-04-30,CUST018,Female,47,Electronics,2,25.0,50.0
19,2023-09-16,CUST019,Female,62,Clothing,2,25.0,50.0
```

iii. ¿Qué porcentaje del total de transacciones corresponde a clientes mayores de 30?

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/porcentaje_mayores30/part-*
72.7
```

3) Transformación de campos

```
[maria_dev@sandbox-hdp scripts_pig]$ cat > retail_sales_ej3_original.pig <<'EOF'
>
> sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
>
>     USING PigStorage(',')
>
>     AS (
>
>         transaction_id_raw:chararray,
>
>         date:chararray,
>
>         customer_id:chararray,
>
>         gender:chararray,
>
>         age:chararray,
>
>         product_category:chararray,
>
>         quantity:chararray,
>
>         price_per_unit:chararray,
>
>         total_amount:chararray
>
>     );
>
>
>
> sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';
>
>
>
> sales_cast = FOREACH sales GENERATE
>
>     (int)transaction_id_raw AS transaction_id,
>
>     date,
>
>     customer_id,
>
>     gender,
>
>     (int)age AS age,
>
>     product_category,
>
>     (int)quantity AS quantity,
>
>     (double)price_per_unit AS price_per_unit,
>
>     (double)total amount AS total amount;

>
>
>
> sales_transform = FOREACH sales_cast GENERATE
>
>     transaction_id,
>
>     date,
>
>     customer_id,
>
>     UPPER(gender) AS gender_mayus,
>
>     age,
>
>     product_category,
>
>     quantity,
>
>     price_per_unit,
>
>     total_amount,
>
>     (price_per_unit * quantity * 0.90) AS importe_descuento;
>
>
>
> primeros20 = LIMIT sales_transform 20;
>
>
>
> STORE sales_transform INTO '/user/maria_dev/resultados/retail_sales/sales_transform_original'
>
>     USING PigStorage(',');
>
>
>
> DUMP primeros20;
>
> EOF
```

i. A partir del conjunto original, crea un alias donde generes:
    1. el género en mayúsculas (UPPER(gender)),

2. una nueva columna importe_descuento que calcule, por ejemplo, price_per_unit * quantity * 0.90 (aplicando un 10% de “descuento ficticio”).

```
2025-11-14 21:48:19,275 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,2023-11-24,CUST001,MALE,34,Beauty,3,50.0,150.0,135.0)
(2,2023-02-27,CUST002,FEMALE,26,Clothing,2,500.0,1000.0,900.0)
(3,2023-01-13,CUST003,MALE,50,Electronics,1,30.0,30.0,27.0)
(4,2023-05-21,CUST004,MALE,37,Clothing,1,500.0,500.0,450.0)
(5,2023-05-06,CUST005,MALE,30,Beauty,2,50.0,100.0,90.0)
(6,2023-04-25,CUST006,FEMALE,45,Beauty,1,30.0,30.0,27.0)
(7,2023-03-13,CUST007,MALE,46,Clothing,2,25.0,50.0,45.0)
(8,2023-02-22,CUST008,MALE,30,Electronics,4,25.0,100.0,90.0)
(9,2023-12-13,CUST009,MALE,63,Electronics,2,300.0,600.0,540.0)
(10,2023-10-07,CUST010,FEMALE,52,Clothing,4,50.0,200.0,180.0)
(11,2023-02-14,CUST011,MALE,23,Clothing,2,50.0,100.0,90.0)
(12,2023-10-30,CUST012,MALE,35,Beauty,3,25.0,75.0,67.5)
(13,2023-08-05,CUST013,MALE,22,Electronics,3,500.0,1500.0,1350.0)
(14,2023-01-17,CUST014,MALE,64,Clothing,4,30.0,120.0,108.0)
(15,2023-01-16,CUST015,FEMALE,42,Electronics,4,500.0,2000.0,1800.0)
(16,2023-02-17,CUST016,MALE,19,Clothing,3,500.0,1500.0,1350.0)
(17,2023-04-22,CUST017,FEMALE,27,Clothing,4,25.0,100.0,90.0)
(18,2023-04-30,CUST018,FEMALE,47,Electronics,2,25.0,50.0,45.0)
(19,2023-09-16,CUST019,FEMALE,62,Clothing,2,25.0,50.0,45.0)
(20,2023-11-05,CUST020,MALE,22,Clothing,3,300.0,900.0,810.0)
```

ii. Muestra los primeros 20 registros resultantes.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/sales_transform_original/part-* | head -n 20
1,2023-11-24,CUST001,MALE,34,Beauty,3,50.0,150.0,135.0
2,2023-02-27,CUST002,FEMALE,26,Clothing,2,500.0,1000.0,900.0
3,2023-01-13,CUST003,MALE,50,Electronics,1,30.0,30.0,27.0
4,2023-05-21,CUST004,MALE,37,Clothing,1,500.0,500.0,450.0
5,2023-05-06,CUST005,MALE,30,Beauty,2,50.0,100.0,90.0
6,2023-04-25,CUST006,FEMALE,45,Beauty,1,30.0,30.0,27.0
7,2023-03-13,CUST007,MALE,46,Clothing,2,25.0,50.0,45.0
8,2023-02-22,CUST008,MALE,30,Electronics,4,25.0,100.0,90.0
9,2023-12-13,CUST009,MALE,63,Electronics,2,300.0,600.0,540.0
10,2023-10-07,CUST010,FEMALE,52,Clothing,4,50.0,200.0,180.0
11,2023-02-14,CUST011,MALE,23,Clothing,2,50.0,100.0,90.0
12,2023-10-30,CUST012,MALE,35,Beauty,3,25.0,75.0,67.5
13,2023-08-05,CUST013,MALE,22,Electronics,3,500.0,1500.0,1350.0
14,2023-01-17,CUST014,MALE,64,Clothing,4,30.0,120.0,108.0
15,2023-01-16,CUST015,FEMALE,42,Electronics,4,500.0,2000.0,1800.0
16,2023-02-17,CUST016,MALE,19,Clothing,3,500.0,1500.0,1350.0
17,2023-04-22,CUST017,FEMALE,27,Clothing,4,25.0,100.0,90.0
18,2023-04-30,CUST018,FEMALE,47,Electronics,2,25.0,50.0,45.0
19,2023-09-16,CUST019,FEMALE,62,Clothing,2,25.0,50.0,45.0
20,2023-11-05,CUST020,MALE,22,Clothing,3,300.0,900.0,810.0
```

4) Agrupación y agregación por categoría de producto

```
cat > retail_sales_ej4.pig <<'EOF'
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

sales_cast = FOREACH sales GENERATE
    (int)transaction_id_raw AS transaction_id,
    date,
    customer_id,
    gender,
    (int)age AS age,
    product_category,
    (int)quantity AS quantity,
    (double)price_per_unit AS price_per_unit,
    (double)total_amount AS total_amount;

grp_categoria = GROUP sales_cast BY product_category;

categoria_stats = FOREACH grp_categoria GENERATE
    group AS product_category,
    COUNT(sales_cast) AS num_transacciones,
    SUM(sales_cast.total_amount) AS total_ventas,
    AVG(sales_cast.age) AS edad_promedio;

ordenado = ORDER categoria_stats BY total_ventas DESC;

STORE ordenado INTO '/user/maria_dev/resultados/retail_sales/categoria_stats'
    USING PigStorage(',');

DUMP ordenado;
EOF
```

i. Agrupa por product_category.
ii. Para cada categoría calcula: número de transacciones (COUNT), suma de total_amount (SUM), edad promedio de cliente (AVG(age)).
iii. Ordena el resultado por la suma de total_amount descendente.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/categoria_stats/part-*
Electronics,342,156905.0,41.73684210526316
Clothing,351,155580.0,41.94871794871795
Beauty,307,143515.0,40.37133550488599
```

5) Extracción de categorías distintas

```
cat > retail_sales_ej5.pig <<'EOF'
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

categorias_distintas = DISTINCT (FOREACH sales GENERATE product_category);

num_categorias = FOREACH (GROUP categorias_distintas ALL) GENERATE COUNT(categorias_distintas) AS total_categorias;

STORE categorias_distintas INTO '/user/maria_dev/resultados/retail_sales/categorias_distintas'
    USING PigStorage(',');

STORE num_categorias INTO '/user/maria_dev/resultados/retail_sales/num_categorias'
    USING PigStorage(',');

DUMP categorias_distintas;
DUMP num_categorias;
EOF
```

i. En este dataset extrae las categorías de producto distintas (DISTINCT product_category).

```
[maria dev@sandbox-hdp scripts pig]$ hdfs dfs -cat /user/maria dev/resultados/retail sales/categorias distintas/part-*
Beauty
Clothing
Electronics
```

ii. Pregunta: ¿Cuántas categorías diferentes hay?

```
[maria dev@sandbox-hdp scripts pig]$ hdfs dfs -cat /user/maria dev/resultados/retail sales/num categorias/part-*
3
```

6) Ordenación y obtención de top-transacciones

```
cat > retail_sales_ej6.pig <<'EOF'
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

sales_cast = FOREACH sales GENERATE
    (int)transaction_id_raw AS transaction_id,
    customer_id,
    product_category,
    (double)total_amount AS total_amount;

ordenado = ORDER sales_cast BY total_amount DESC;

top5 = LIMIT ordenado 5;

STORE top5 INTO '/user/maria_dev/resultados/retail_sales/top5_transacciones'
    USING PigStorage(',');

DUMP top5;
EOF
```

i. Ordena todas las transacciones por total_amount descendente.
ii. Usa LIMIT para extraer, por ejemplo, las 5 transacciones con mayor total_amount.
iii. Muestra: transaction_id, customer_id, product_category, total_amount.

```
[maria_dev@sandbox-hdp scripts pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/top5_transacciones/part-*
592,CUST592,Beauty,2000.0
447,CUST447,Beauty,2000.0
927,CUST927,Electronics,2000.0
74,CUST074,Beauty,2000.0
72,CUST072,Electronics,2000.0
```

7) Uso de funciones de cadena

```
cat > retail_sales_ej7.pig <<'EOF'
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

sales_cast = FOREACH sales GENERATE
    (int)transaction_id_raw AS transaction_id,
    date,
    customer_id,
    gender,
    (int)age AS age,
    product_category,
    (int)quantity AS quantity,
    (double)price_per_unit AS price_per_unit,
    (double)total_amount AS total_amount;
```

```
sales_string = FOREACH sales_cast GENERATE
    transaction_id,
    date,
    customer_id,
    gender,
    age,
    product_category,
    SUBSTRING(product_category,0,3) AS product_cat_short,
    SIZE(product_category) AS product_cat_len,
    quantity,
    price_per_unit,
    total_amount;

primeros15 = LIMIT sales_string 15;

STORE sales_string INTO '/user/maria_dev/resultados/retail_sales/sales_string'
    USING PigStorage(',');

DUMP primeros15;
EOF
```

i. Añade una nueva columna al alias original donde el product_category se recorte a los primeros 3 caracteres (SUBSTRING(product_category,0,3)) y otra que sea la longitud del product_category (SIZE(product_category)).
ii. Muestra los primeros 15 registros resultantes

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/sales_string/part-* | head -n 15
1,2023-11-24,CUST001,Male,34,Beauty,Bea,6,3,50.0,150.0
2,2023-02-27,CUST002,Female,26,Clothing,Clo,8,2,500.0,1000.0
3,2023-01-13,CUST003,Male,50,Electronics,Ele,11,1,30.0,30.0
4,2023-05-21,CUST004,Male,37,Clothing,Clo,8,1,500.0,500.0
5,2023-05-06,CUST005,Male,30,Beauty,Bea,6,2,50.0,100.0
6,2023-04-25,CUST006,Female,45,Beauty,Bea,6,1,30.0,30.0
7,2023-03-13,CUST007,Male,46,Clothing,Clo,8,2,25.0,50.0
8,2023-02-22,CUST008,Male,30,Electronics,Ele,11,4,25.0,100.0
9,2023-12-13,CUST009,Male,63,Electronics,Ele,11,2,300.0,600.0
10,2023-10-07,CUST010,Female,52,Clothing,Clo,8,4,50.0,200.0
11,2023-02-14,CUST011,Male,23,Clothing,Clo,8,2,50.0,100.0
12,2023-10-30,CUST012,Male,35,Beauty,Bea,6,3,25.0,75.0
13,2023-08-05,CUST013,Male,22,Electronics,Ele,11,3,500.0,1500.0
14,2023-01-17,CUST014,Male,64,Clothing,Clo,8,4,30.0,120.0
15,2023-01-16,CUST015,Female,42,Electronics,Ele,11,4,500.0,2000.0
```

8) Filtrado por fecha y condiciones combinadas

```
cat > retail_sales_ej8.pig <<'EOF'
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

sales_cast = FOREACH sales GENERATE
    (int)transaction_id_raw AS transaction_id,
    date,
    customer_id,
    gender,
    (int)age AS age,
    product_category,
    (int)quantity AS quantity,
    (double)price_per_unit AS price_per_unit,
    (double)total_amount AS total_amount;
```

```
ventas_filtradas_fecha = FILTER sales_cast BY date < '2023-07-01';

ventas_filtradas = FILTER ventas_filtradas_fecha BY total_amount > 500;

grp_edad = GROUP ventas_filtradas ALL;

edad_promedio = FOREACH grp_edad GENERATE AVG(ventas_filtradas.age) AS edad_promedio;

STORE ventas_filtradas INTO '/user/maria_dev/resultados/retail_sales/ventas_filtradas' USING PigStorage(',');

STORE edad_promedio INTO '/user/maria_dev/resultados/retail_sales/edad_promedio_filtradas' USING PigStorage(',');

DUMP ventas_filtradas;
DUMP edad_promedio;
EOF
```

i. Filtra primero las transacciones que se han hecho antes de una determinada fecha, por ejemplo, date < '2023-07-01'. (Suponiendo que el campo date es tipo chararray con formato 'YYYY-MM-DD').

ii. De ese conjunto, filtra adicionalmente las transacciones con total_amount > 500.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/ventas_filtradas/part-*
2,2023-02-27,CUST002,Female,26,Clothing,2,500.0,1000.0
15,2023-01-16,CUST015,Female,42,Electronics,4,500.0,2000.0
16,2023-02-17,CUST016,Male,19,Clothing,3,500.0,1500.0
31,2023-05-23,CUST031,Male,44,Electronics,4,300.0,1200.0
36,2023-06-24,CUST036,Male,52,Beauty,3,300.0,900.0
42,2023-02-17,CUST042,Male,22,Clothing,3,300.0,900.0
46,2023-06-26,CUST046,Female,20,Electronics,4,300.0,1200.0
48,2023-05-16,CUST048,Male,54,Electronics,3,300.0,900.0
49,2023-01-23,CUST049,Female,54,Electronics,2,500.0,1000.0
54,2023-02-10,CUST054,Female,38,Electronics,3,500.0,1500.0
56,2023-05-31,CUST056,Female,26,Clothing,3,300.0,900.0
67,2023-05-29,CUST067,Female,48,Beauty,4,300.0,1200.0
72,2023-05-23,CUST072,Female,20,Electronics,4,500.0,2000.0
94,2023-05-19,CUST094,Female,47,Beauty,2,500.0,1000.0
101,2023-01-29,CUST101,Male,32,Clothing,2,300.0,600.0
104,2023-06-11,CUST104,Female,34,Beauty,2,500.0,1000.0
107,2023-02-03,CUST107,Female,21,Clothing,4,300.0,1200.0
110,2023-06-11,CUST110,Male,27,Clothing,3,300.0,900.0
111,2023-04-19,CUST111,Female,34,Electronics,3,500.0,1500.0
117,2023-03-15,CUST117,Male,19,Electronics,2,500.0,1000.0
118,2023-05-16,CUST118,Female,30,Electronics,4,500.0,2000.0
129,2023-04-23,CUST129,Female,21,Beauty,2,300.0,600.0
133,2023-02-16,CUST133,Male,20,Electronics,3,300.0,900.0
136,2023-03-20,CUST136,Male,44,Electronics,2,300.0,600.0
```

iii. Muestra el resultado, y calcula la edad promedio (AVG(age)) de los clientes que cumplen estas condiciones.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/resultados/retail_sales/edad_promedio_filtradas/part-*
39.33116883116883
```

9) Script completo + almacenamiento

i. Crea un script .pig que contenga los pasos: carga, filtrado, transformación, agrupación, ordenación, y finalmente almacenamiento (STORE) del resultado final en un directorio (por ejemplo /usr/maria_dev/ventas_analisis). Debes crear tú los filtros, transforamciones, etc. que desees.

```
cat > retail_sales_analisis.pig <<'EOF'
-- Carga del dataset original
sales_raw = LOAD '/user/maria_dev/retail_sales/retail_sales_dataset.csv'
    USING PigStorage(',')
    AS (
        transaction_id_raw:chararray,
        date:chararray,
        customer_id:chararray,
        gender:chararray,
        age:chararray,
        product_category:chararray,
        quantity:chararray,
        price_per_unit:chararray,
        total_amount:chararray
    );

-- Filtrado para quitar la cabecera
sales = FILTER sales_raw BY transaction_id_raw != 'Transaction ID';

-- Conversión de tipos
sales_cast = FOREACH sales GENERATE
    (int)transaction_id_raw AS transaction_id,
    date,
    customer_id,
    UPPER(gender) AS gender_mayus,
    (int)age AS age,
    product_category,
    (int)quantity AS quantity,
    (double)price_per_unit AS price_per_unit,
    (double)total_amount AS total_amount;

-- Filtrado adicional: ventas mayores a 200
ventas_filtradas = FILTER sales_cast BY total_amount > 200;

-- Agregación por categoría de producto
grp_categoria = GROUP ventas_filtradas BY product_category;

categoria_stats = FOREACH grp_categoria GENERATE
    group AS product_category,
    COUNT(ventas_filtradas) AS num_transacciones,
    SUM(ventas_filtradas.total_amount) AS total_ventas,
    AVG(ventas_filtradas.age) AS edad_promedio;

-- Ordenación por ventas totales descendente
ordenado = ORDER categoria_stats BY total_ventas DESC;

-- Almacenamiento del resultado final en HDFS
STORE ordenado INTO '/user/maria_dev/ventas_analisis'
    USING PigStorage(',');

DUMP ordenado;
EOF
```

ii. Asegúrate de comentar la operación de cada bloque del script con -- comentario.

iii. Ejecuta el script en modo MapReduce estándar (pig script.pig).

```
[maria_dev@sandbox-hdp scripts_pig]$ pig -x mapreduce -f retail_sales_analisis.pig
25/11/14 22:17:41 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
25/11/14 22:17:41 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
25/11/14 22:17:41 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2025-11-14 22:17:41,264 [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
2025-11-14 22:17:41,265 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/maria_dev/scripts_pig/pig_1763158661258.log
2025-11-14 22:17:45,260 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/maria_dev/.pigbootup not found
2025-11-14 22:17:45,603 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://sandbox-hdp
s.com:8020
2025-11-14 22:17:49,115 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-retail_sales_analisis.pig-74c476f4-2497-49fc-ab80-4babb67d
2025-11-14 22:17:51,542 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox-hdp.hortonworks.com
/timeline/
2025-11-14 22:17:52,649 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2025-11-14 22:17:58,476 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2025-11-14 22:17:58,892 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FILTER
2025-11-14 22:17:58,959 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
```

iv. Verifica los archivos de salida y comprueba que los resultados tienen sentido.

```
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -ls /user/maria_dev/ventas_analisis
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2025-11-14 22:22 /user/maria_dev/ventas_analisis/_SUCCESS
-rw-r--r--   1 maria_dev hdfs        122 2025-11-14 22:22 /user/maria_dev/ventas_analisis/part-r-00000
[maria_dev@sandbox-hdp scripts_pig]$ hdfs dfs -cat /user/maria_dev/ventas_analisis/part-*
Electronics,139,139400.0,42.15827338129496
Clothing,136,136400.0,40.455882352941174
Beauty,121,127100.0,39.30578512396694
```

3) APARTADO C

1) Localiza en Internet una versión del Quijote en formato texto. Descárgala y cópiala a en tu sistema HDFS. Implementa un contador de palabras (cuantas veces aparece cada palabra en un texto).

   i. Implementa en PIG el script necesario para hacer dicha operación.

```
cat > contador_palabras_quijote.pig <<'EOF'
quijote = LOAD '/user/maria_dev/dataset_quijote/el_quijote.txt' USING TextLoader AS
(linea:chararray);

palabras = FOREACH quijote GENERATE FLATTEN(TOKENIZE(linea)) AS palabra;

palabras_lower = FOREACH palabras GENERATE LOWER(palabra) AS palabra;

grupo_palabras = GROUP palabras_lower BY palabra;

conteo_palabras = FOREACH grupo_palabras GENERATE group AS palabra, COUNT(palabras_lower) AS
cantidad;

ordenado = ORDER conteo_palabras BY cantidad DESC;

STORE ordenado INTO '/user/maria_dev/pig_quijote' USING PigStorage(',');

DUMP LIMIT ordenado 20;
EOF
```

   ii. Muestra un ejemplo de ejecución sobre El Quijote en pantalla.

```
[maria_dev@sandbox-hdp ~]$ pig -f contador_palabras_quijote.pig
25/11/14 22:48:00 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
25/11/14 22:48:00 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
25/11/14 22:48:00 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
25/11/14 22:48:00 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
25/11/14 22:48:00 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2025-11-14 22:48:00,329 [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
2025-11-14 22:48:00,329 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/maria_dev/pig_1763160480326.log
2025-11-14 22:48:04,089 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/maria_dev/.pigbootup not found
2025-11-14 22:48:04,519 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://sandbox-hdp.hortonwork
s.com:8020
2025-11-14 22:48:07,042 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-contador_palabras_quijote.pig-2029d8c1-f3eb-40ab-806e-3ede14cc9f01
2025-11-14 22:48:09,088 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox-hdp.hortonworks.com:8188/ws/v1
/timeline/
2025-11-14 22:48:09,451 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2025-11-14 22:48:12,733 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold
 = 489580128, usageThreshold = 489580128
2025-11-14 22:48:13,527 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY
2025-11-14 22:48:13,598 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2025-11-14 22:48:13,729 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatte
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2025-11-14 22:48:14,609 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/maria_dev/staging and resources direc
tory is /tmp/temp41326311
2025-11-14 22:48:15,038 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshold: 100 optimistic? false
2025-11-14 22:48:15,419 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to move algebraic foreach to combiner
```

iii. Almacena la salida en una carpeta de HDFS llamada /usr/maria_dev/pig_quijote.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/pig_quijote/part-* | head -n 20
que,10725
de,9030
y,8638
la,5009
a,4807
en,4031
el,3854
no,2977
se,2382
los,2148
con,2079
por,1911
su,1861
lo,1803
le,1802
las,1488
me,1155
como,1149
del,1127
don,1070
```