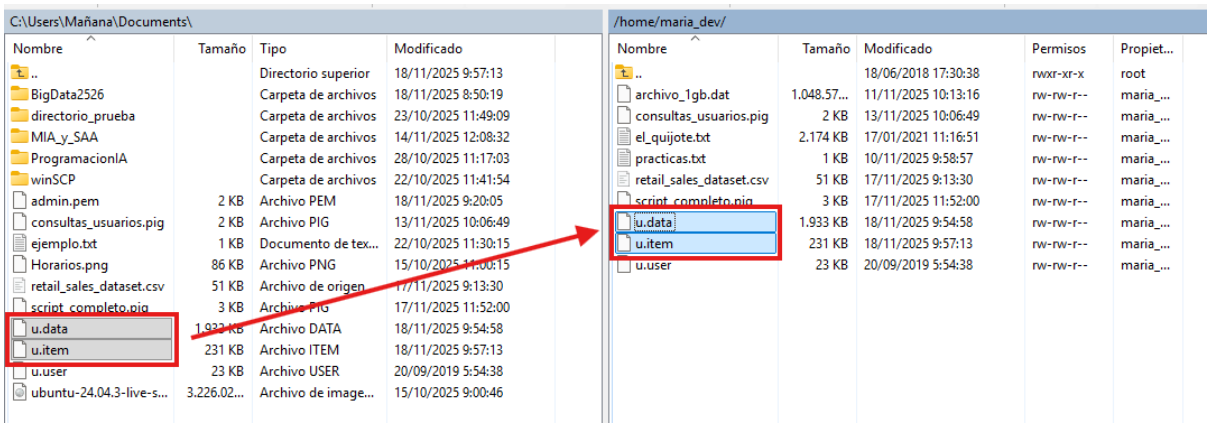
PR_03.2 Dani Gayol Rodríguez

PR_03.2 Dani Gayol Rodríguez 1
Apartado A 1
1.) Describe informalmente los pasos que darás para llegar a la solución 2
2.) Implementa en PIG el script necesario para obtener la información deseada. 3
3.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS. 5
Apartado B 5
1.) Describe informalmente los pasos que darás para llegar a la solución. 6
2.) Implementa en PIG el script necesario para obtener la información deseada. 6
3.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS. 8
Apartado C 9
1.) Implementa en PIG el script necesario para obtener la información deseada. 10
2.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS 11
Apartado D 12
1.) Implementa en PIG el script necesario para obtener la información deseada. 13
2.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS. 15
Apartado E 16
1.) Implementa en PIG el script necesario para obtener la información deseada. 17
2.) Abrir el archive en Excel y generar un gráfico de barras con los datos. 18

Apartado A

1.) Describe informalmente los pasos que darás para llegar a la solución

Primero tenemos que pasar los archivos “u.data” y “u.item” a nuestra máquina virtual y para ello vamos a usar “WinSCP”



Una vez pasados al directorio “/home/maria_dev” de nuestra máquina virtual, vamos a copiarlos al servicio HDFS en el directorio “/user/maria_dev”

```
C:\Users\Mañana>ssh maria_dev@localhost -p 2222
maria_dev@localhost's password:
Last login: Mon Nov 17 10:23:15 2025 from 172.18.0.3
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         el_quijote.txt  retail_sales_dataset.csv  u.data  u.user
consultas_usuarios.pig  practicas.txt   script_completo.pig       u.item
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/u.data /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/u.item /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 10 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
-rw-r--r--   1 maria_dev hdfs    2226045 2025-11-17 11:14 /user/maria_dev/el_quijote.txt
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
-rw-r--r--   1 maria_dev hdfs    1979173 2025-11-18 09:03 /user/maria_dev/u.data
-rw-r--r--   1 maria_dev hdfs     236344 2025-11-18 09:03 /user/maria_dev/u.item
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis
```

Ahora vamos a cargar los datos de “u.data” y “u.item”

```
grunt> data = LOAD '/user/maria_dev/u.data' USING PigStorage('\t') AS (user_id:int, movie_id:int, rating:int, timestamp:
chararray);
grunt> items = LOAD '/user/maria_dev/u.item' USING PigStorage('|') AS (movie_id:int, title:chararray, gen1:chararray, ge
n2:chararray, gen3:chararray, gen4:chararray, gen5:chararray, gen6:chararray, gen7:chararray, gen8:chararray, gen9:chara
rray, gen10:chararray, gen11:chararray, gen12:chararray, gen13:chararray, gen14:chararray, gen15:chararray, gen16:charar
ray, gen17:chararray, gen18:chararray, gen19:chararray);
```

Después de cargar los datos, vamos a contar los votos

```
grunt> grouped = GROUP data BY movie_id; counts = FOREACH grouped GENERATE group AS movie_id,
COUNT(data) AS num_votes;
```

Una vez contados, vamos a unirlos, seleccionar solo los datos que nos piden y finalmente ordenarlos

```
grunt> joined = JOIN counts BY movie_id, items BY movie_id;
grunt> final = FOREACH joined GENERATE counts::movie_id AS movie_id, items::title AS title, counts::num_votes AS votes;
```

```
grunt> ordered = ORDER final BY votes DESC;
```

Una vez ordenados, nos quedamos solo con las cinco primeras

```
grunt> top5 = LIMIT ordered 5;
```

Finalmente mostramos los resultados usando el comando DUMP

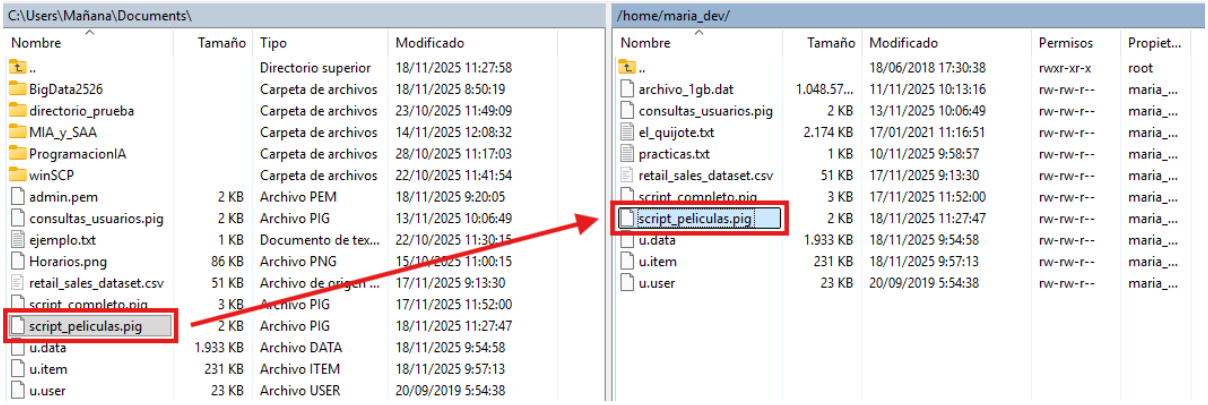
```
(50,Star Wars (1977),583)
(258,Contact (1997),509)
(100,Fargo (1996),508)
(181,Return of the Jedi (1983),507)
(294,Liar Liar (1997),485)
```

2.) Implementa en PIG el script necesario para obtener la información deseada.

Primero creamos el script, en mi caso lo voy a hacer en un bloc de notas y sustituir el “.txt” por un “.pig”

```
-- Cargar u.data
data = LOAD '/user/maria_dev/u.data'
       USING PigStorage('\t')
       AS (user_id:int, movie_id:int, rating:int, timestamp:chararray);

-- Cargar u.item
items = LOAD '/user/maria_dev/u.item'
        USING PigStorage('|')
        AS (movie_id:int, title:chararray, gen1:chararray, gen2:chararray,
            gen3:chararray, gen4:chararray, gen5:chararray, gen6:chararray,
            gen7:chararray, gen8:chararray, gen9:chararray, gen10:chararray,
            gen11:chararray, gen12:chararray, gen13:chararray, gen14:chararray,
            gen15:chararray, gen16:chararray, gen17:chararray, gen18:chararray,
            gen19:chararray);

-- Contar votos por película
grouped = GROUP data BY movie_id;
counts = FOREACH grouped GENERATE
            group AS movie_id,
            COUNT(data) AS num_votes;

-- Unir con títulos
joined = JOIN counts BY movie_id, items BY movie_id;

-- Seleccionar solo los campos necesarios
final = FOREACH joined GENERATE
            counts::movie_id AS movie_id,
            items::title AS title,
            counts::num_votes AS votes;

-- Ordenar DESC y quedarnos con las 5 más votadas
ordered = ORDER final BY votes DESC;
top5 = LIMIT ordered 5;

-- Mostrar resultado por pantalla
DUMP top5;

-- Guardar resultado en HDFS
STORE top5 INTO '/user/maria_dev/top5_peliculas'
     USING PigStorage(',');
```

Ahora usando “WinSCP” lo vamos a copiar a la máquina virtual



```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         el_quijote.txt  retail_sales_dataset.csv  script_peliculas.pig  u.item
consultas_usuarios.pig  practicas.txt   script_completo.pig       u.data                u.user
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_peliculas.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 12 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
-rw-r--r--   1 maria_dev hdfs    2226045 2025-11-17 11:14 /user/maria_dev/el_quijote.txt
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:11 /user/maria_dev/maria_dev
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
-rw-r--r--   1 maria_dev hdfs       1401 2025-11-19 10:13 /user/maria_dev/script_peliculas.pig
-rw-r--r--   1 maria_dev hdfs    1979173 2025-11-18 09:03 /user/maria_dev/u.data
-rw-r--r--   1 maria_dev hdfs     236344 2025-11-18 09:03 /user/maria_dev/u.item
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis
```

Ahora ya podemos ejecutar el script

```
[maria_dev@sandbox-hdp ~]$ pig script_peliculas.pig
```

```
(50,Star Wars (1977),583)
(258,Contact (1997),509)
(100,Fargo (1996),508)
(181,Return of the Jedi (1983),507)
(294,Liar Liar (1997),485)
```

3.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS.

Una vez ejercitado el script nos muestra toda la información y el resultado por pantalla fue este;

```
(50,Star Wars (1977),583)
(258,Contact (1997),509)
(100,Fargo (1996),508)
(181,Return of the Jedi (1983),507)
(294,Liar Liar (1997),485)
```

Ahora vamos a ver los resultados en los archivos de HDFS

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/top5_peliculas/
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2025-11-19 10:21 /user/maria_dev/top5_peliculas/_SUCCESS
-rw-r--r--   1 maria_dev hdfs        127 2025-11-19 10:21 /user/maria_dev/top5_peliculas/part-v007-o000-r-00000
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/top5_peliculas/part*
50,Star Wars (1977),583
258,Contact (1997),509
100,Fargo (1996),508
181,Return of the Jedi (1983),507
294,Liar Liar (1997),485
```

Apartado B

1.) Describe informalmente los pasos que darás para llegar a la solución.

Como ya hice en el apartado anterior, ya subí los archivos necesarios a la máquina virtual, por lo tanto, solo tengo que cargarlos en "pig"

```
grunt> udata = LOAD '/user/maria_dev/u.data' USING PigStorage('\t') AS (user_id:int, movie_id:int, rating:int, timestamp
:long);
grunt> uitem = LOAD '/user/maria_dev/u.item' USING PigStorage('|') AS (movie_id:int, title:chararray, info:chararray);
```

Una vez cargada los datos, vamos a agruparlas y calcular la media de puntación

```
grunt> grouped = GROUP udata BY movie_id;
grunt> medias = FOREACH grouped GENERATE group AS movie_id, AVG(udata.rating) AS media_rating;
```

Ahora vamos a seleccionar únicamente los campos necesarios

```
grunt> joined = JOIN medias BY movie_id, uitem BY movie_id;
grunt> final = FOREACH joined GENERATE medias::movie_id AS movie_id, uitem::title AS title, medias::media_rating AS medi
a_rating;
```

Después, las ordenamos y nos quedamos con las 10 primeras

```
grunt> ordenado = ORDER final BY media_rating DESC;
grunt> top10 = LIMIT ordenado 10;
```

Finalmente las vamos a mostrar por pantalla usando el comando "DUMP"

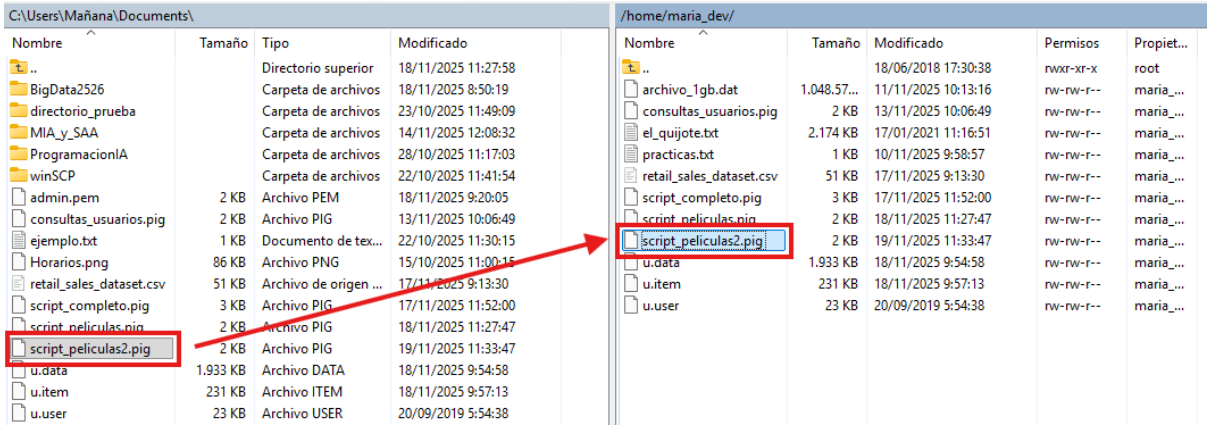
```
(1653,Entertaining Angels: The Dorothy Day Story (1996),5.0)
(1293,Star Kid (1997),5.0)
(1467,Saint of Fort Washington, The (1993),5.0)
(814,Great Day in Harlem, A (1994),5.0)
(1500,Santa with Muscles (1996),5.0)
(1201,Marlene Dietrich: Shadow and Light (1996) ,5.0)
(1122,They Made Me a Criminal (1939),5.0)
(1189,Prefontaine (1997),5.0)
(1599,Someone Else's America (1995),5.0)
(1536,Aiqing wansui (1994),5.0)
```

2.) Implementa en PIG el script necesario para obtener la información deseada.

Primero creamos el script, en mi caso lo voy a hacer en un bloc de notas y sustituir el ".txt" por un ".pig"

```
-- Cargar u.data: user, movie, rating, timestamp
udata = LOAD '/user/maria_dev/u.data'
        USING PigStorage('\t')
        AS (user_id:int, movie_id:int, rating:int, timestamp:long);

-- Cargar u.item: movie_id | title | ...
uitem = LOAD '/user/maria_dev/u.item'
        USING PigStorage('|')
        AS (movie_id:int, title:chararray, info:chararray);

-- Agrupar todas las puntuaciones por película
grouped = GROUP udata BY movie_id;

-- Calcular la media de puntuación
medias = FOREACH grouped GENERATE
            group AS movie_id,
            AVG(udata.rating) AS media_rating;

-- Unir con títulos
joined = JOIN medias BY movie_id, uitem BY movie_id;

-- Seleccionar campos finales
final = FOREACH joined GENERATE
            medias::movie_id AS movie_id,
            uitem::title AS title,
            medias::media_rating AS media_rating;

-- Ordenar por media descendente
ordenado = ORDER final BY media_rating DESC;

-- Obtener las 10 mejores valoradas
top10 = LIMIT ordenado 10;

-- Mostrar en pantalla
DUMP top10;

-- Guardar en HDFS
STORE top10 INTO '/user/maria_dev/top10_valoradas'
      USING PigStorage('\t');
```

Ahora usando “WinSCP” lo vamos a copiar a la máquina virtual



```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev/
archivo_1gb.dat         el_quijote.txt  retail_sales_dataset.csv  script_peliculas2.pig  u.data  u.user
consultas_usuarios.pig  practicas.txt   script_completo.pig       script_peliculas.pig   u.item
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_peliculas2.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 15 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
-rw-r--r--   1 maria_dev hdfs    2226045 2025-11-17 11:14 /user/maria_dev/el_quijote.txt
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:11 /user/maria_dev/maria_dev
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
-rw-r--r--   1 maria_dev hdfs       1401 2025-11-19 10:13 /user/maria_dev/script_peliculas.pig
-rw-r--r--   1 maria_dev hdfs       1173 2025-11-19 10:46 /user/maria_dev/script_peliculas2.pig
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:43 /user/maria_dev/top10_valoradas
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:21 /user/maria_dev/top5_peliculas
-rw-r--r--   1 maria_dev hdfs    1979173 2025-11-18 09:03 /user/maria_dev/u.data
-rw-r--r--   1 maria_dev hdfs     236344 2025-11-18 09:03 /user/maria_dev/u.item
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis
```

Ahora ya podemos ejecutar el script

```
[maria_dev@sandbox-hdp ~]$ pig script_peliculas2.pig
```

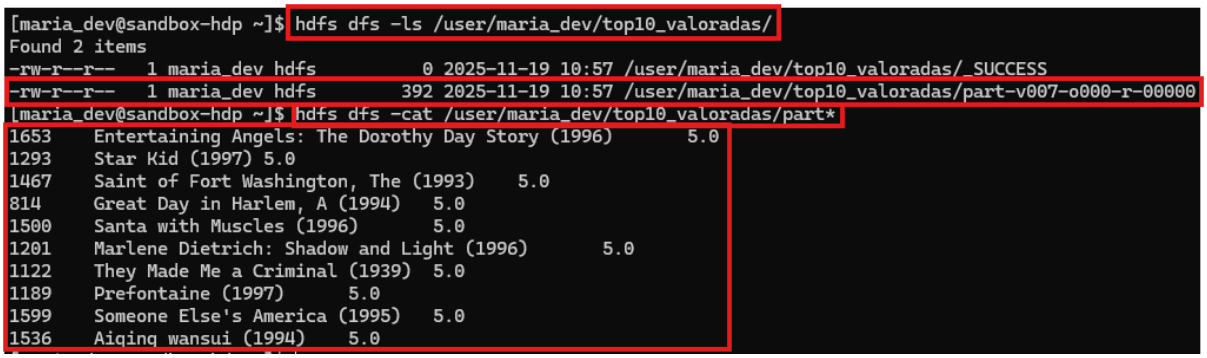
```
(1653,Entertaining Angels: The Dorothy Day Story (1996),5.0)
(1293,Star Kid (1997),5.0)
(1467,Saint of Fort Washington, The (1993),5.0)
(814,Great Day in Harlem, A (1994),5.0)
(1500,Santa with Muscles (1996),5.0)
(1201,Marlene Dietrich: Shadow and Light (1996) ,5.0)
(1122,They Made Me a Criminal (1939),5.0)
(1189,Prefontaine (1997),5.0)
(1599,Someone Else's America (1995),5.0)
(1536,Aiqing wansui (1994),5.0)
```

3.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS.

Una vez ejercitado el script nos muestra toda la información y el resultado por pantalla fue este;

```
(1653,Entertaining Angels: The Dorothy Day Story (1996),5.0)
(1293,Star Kid (1997),5.0)
(1467,Saint of Fort Washington, The (1993),5.0)
(814,Great Day in Harlem, A (1994),5.0)
(1500,Santa with Muscles (1996),5.0)
(1201,Marlene Dietrich: Shadow and Light (1996) ,5.0)
(1122,They Made Me a Criminal (1939),5.0)
(1189,Prefontaine (1997),5.0)
(1599,Someone Else's America (1995),5.0)
(1536,Aiqing wansui (1994),5.0)
```

Ahora vamos a ver los resultados en los archivos de HDFS



Apartado C

1.) Implementa en PIG el script necesario para obtener la información deseada.

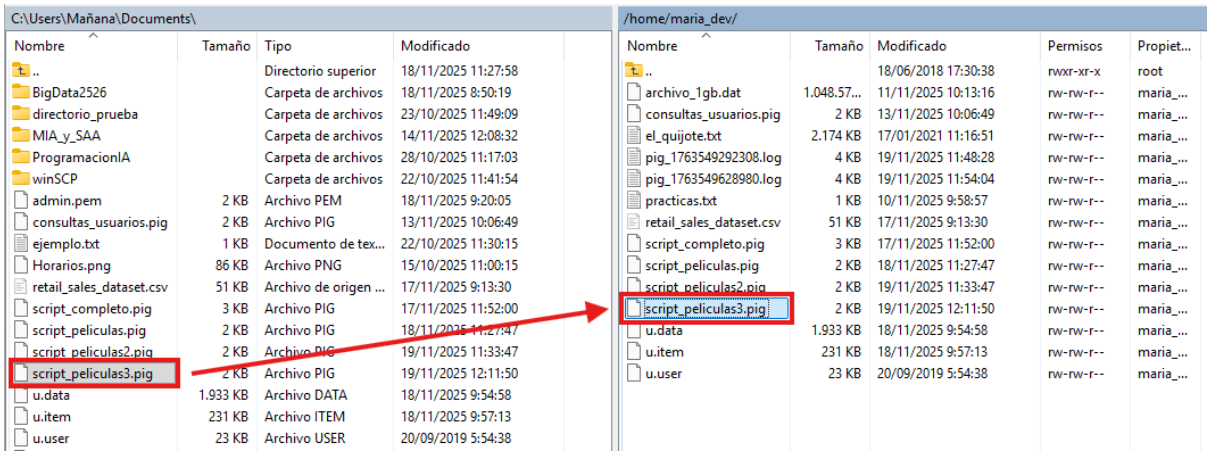
Para saber el año nos tenemos que fijar en el "titulo" del archivo "u.item", ahi aparece el título de la película junto con el año, entonces lo que tenemos que hacer es extraer el año de ahí

```
grunt> peliculas = FOREACH uitem GENERATE movie_id, title, (int) REGEX_EXTRACT(title, '.*\\((\\d{4})\\)$', 1) AS year;
```

Ahora creamos el script, en mi caso lo voy a hacer en un bloc de notas y sustituir el ".txt" por un ".pig"

```
-- Cargar u.data: user, movie, rating, timestamp
udata = LOAD '/user/maria_dev/u.data'
        USING PigStorage('\t')
        AS (user_id:int, movie_id:int, rating:int, timestamp:long);

-- Cargar u.item: movie_id | title | otros campos...
uitem = LOAD '/user/maria_dev/u.item'
        USING PigStorage('|')
        AS (movie_id:int, title:chararray, info:chararray);

-- Extraer año de las películas (del título)
peliculas = FOREACH uitem GENERATE
                movie_id,
                title,
                (int) REGEX_EXTRACT(title, '.*\\((\\d{4})\\)$', 1) AS year;

-- Agrupar valoraciones por película
grouped = GROUP udata BY movie_id;

-- Calcular media de puntuación
medias = FOREACH grouped GENERATE
            group AS movie_id,
            AVG(udata.rating) AS media_rating;

-- Unir medias con títulos y año
joined = JOIN medias BY movie_id, peliculas BY movie_id;

-- Filtrar películas con media > 4
filtradas = FILTER joined BY medias::media_rating > 4.0;

-- Seleccionar campos finales
final = FOREACH filtradas GENERATE
            peliculas::movie_id AS movie_id,
            peliculas::title AS title,
            peliculas::year AS year,
            medias::media_rating AS media_rating;

-- Ordenar por año ascendente (más antiguas primero)
ordenadas = ORDER final BY year ASC;

-- Tomar las 5 más antiguas
top5_antiguas = LIMIT ordenadas 5;

-- Mostrar en pantalla
DUMP top5_antiguas;

-- Guardar en HDFS
STORE top5_antiguas INTO '/user/maria_dev/top5_antiguas_valoradas'
      USING PigStorage('\t');
```

Ahora usando “WinSCP” lo vamos a copiar a la máquina virtual




```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         pig_1763549292308.log  retail_sales_dataset.csv  script_peliculas3.pig  u.item
consultas_usuarios.pig  pig_1763549628980.log  script_completo.pig       script_peliculas.pig   u.user
el_quijote.txt          practicas.txt          script_peliculas2.pig     u.data
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_peliculas3.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 16 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
-rw-r--r--   1 maria_dev hdfs    2226045 2025-11-17 11:14 /user/maria_dev/el_quijote.txt
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:11 /user/maria_dev/maria_dev
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
-rw-r--r--   1 maria_dev hdfs       1401 2025-11-19 10:13 /user/maria_dev/script_peliculas.pig
-rw-r--r--   1 maria_dev hdfs       1173 2025-11-19 10:46 /user/maria_dev/script_peliculas2.pig
-rw-r--r--   1 maria_dev hdfs       1589 2025-11-19 11:16 /user/maria_dev/script_peliculas3.pig
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:57 /user/maria_dev/top10_valoradas
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:21 /user/maria_dev/top5_peliculas
-rw-r--r--   1 maria_dev hdfs    1979173 2025-11-18 09:03 /user/maria_dev/u.data
-rw-r--r--   1 maria_dev hdfs     236344 2025-11-18 09:03 /user/maria_dev/u.item
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis
```

Ahora ya podemos ejecutar el script

```
[maria_dev@sandbox-hdp ~]$ pig script_peliculas3.pig
```

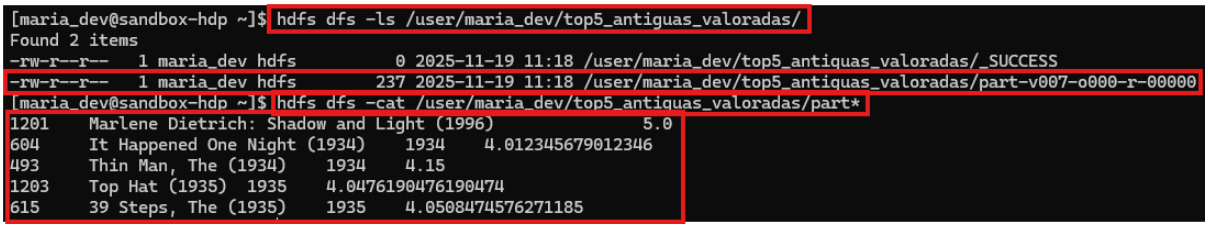
```
(1201,Marlene Dietrich: Shadow and Light (1996) ,,5.0)
(604,It Happened One Night (1934),1934,4.0123456790123456)
(493,Thin Man, The (1934),1934,4.15)
(1203,Top Hat (1935),1935,4.047619047619047)
(615,39 Steps, The (1935),1935,4.050847457627119)
```

2.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS

Una vez ejercitado el script nos muestra toda la información y el resultado por pantalla fue este;

```
(1201,Marlene Dietrich: Shadow and Light (1996) ,,5.0)
(604,It Happened One Night (1934),1934,4.012345679012346)
(493,Thin Man, The (1934),1934,4.15)
(1203,Top Hat (1935),1935,4.0476190476190474)
(615,39 Steps, The (1935),1935,4.0508474576271185)
```

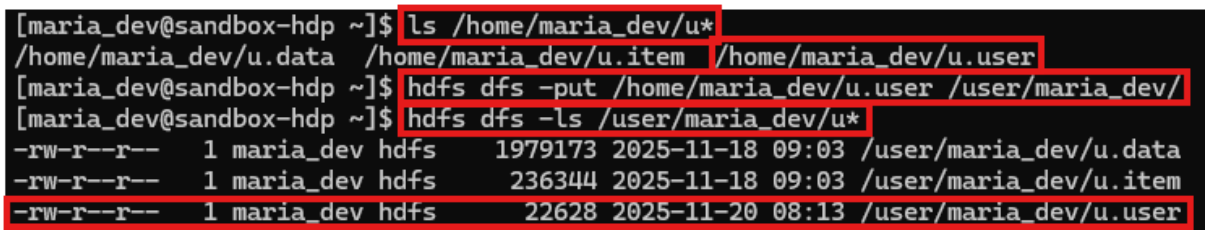
Ahora vamos a ver los resultados en los archivos de HDFS



Apartado D

1.) Implementa en PIG el script necesario para obtener la información deseada.

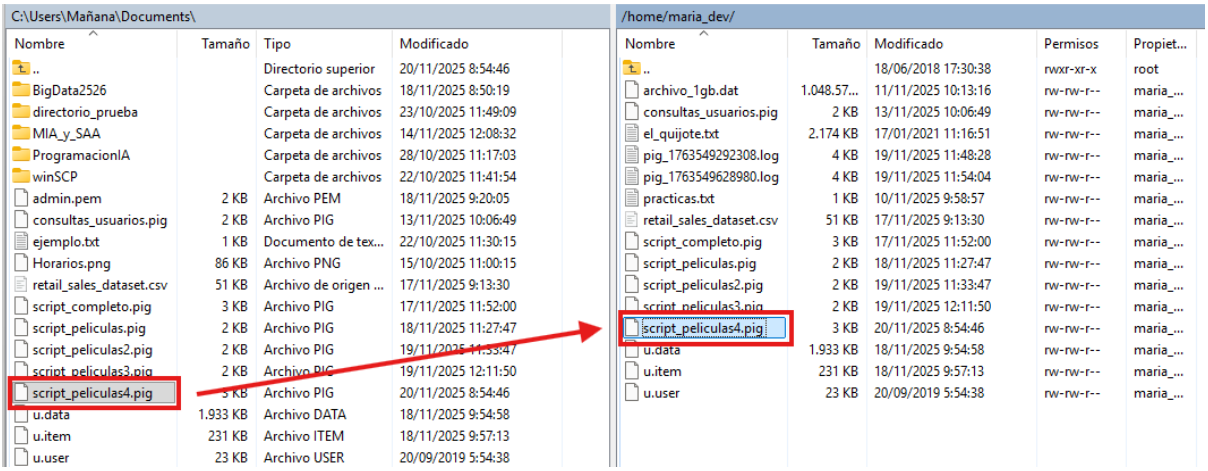
Para saber la película mejor valorada para cada ocupación, tenemos que cargar también los datos del archivo “u.user”, sabiendo esto, vamos a crear el script



Ahora creamos el script, en mi caso lo voy a hacer en un bloc de notas y sustituir el “.txt” por un “.pig”

```
-- Cargar u.data: user_id, movie_id, rating, timestamp
udata = LOAD '/user/maria_dev/u.data'
        USING PigStorage('\t')
        AS (user_id:int, movie_id:int, rating:int, timestamp:long);

-- Cargar u.user: user_id | age | gender | occupation | zip
uuser = LOAD '/user/maria_dev/u.user'
        USING PigStorage('|')
        AS (user_id:int, age:int, gender:chararray, occupation:chararray, zip:chararray);

-- Cargar u.item: movie_id | title | ...
uitem = LOAD '/user/maria_dev/u.item'
        USING PigStorage('|')
        AS (movie_id:int, title:chararray, info:chararray);

-- Unir valoraciones con usuarios para obtener ocupación
ratings_with_occ = JOIN udata BY user_id, uuser BY user_id;

-- Seleccionar movie_id, rating y ocupación
clean = FOREACH ratings_with_occ GENERATE
            udata::movie_id AS movie_id,
            udata::rating AS rating,
            uuser::occupation AS occupation;

-- Agrupar por ocupación y película
grouped = GROUP clean BY (occupation, movie_id);

-- Calcular media de rating por cada (ocupación, película)
medias = FOREACH grouped GENERATE
            group.occupation AS occupation,
            group.movie_id AS movie_id,
            AVG(clean.rating) AS media_rating;

-- Ordenar primero por ocupación y luego por rating descendente
ordenadas = ORDER medias BY occupation ASC, media_rating DESC;

-- Obtener solo la mejor película para cada ocupación
mejores = FOREACH (GROUP ordenadas BY occupation)
              GENERATE
                  group AS occupation,
                  FLATTEN(LIMIT ordenadas 1);

-- Unir con títulos de películas
joined = JOIN mejores BY movie_id, uitem BY movie_id;

-- Seleccionar campos finales
top_ocupacion = FOREACH joined GENERATE
            mejores::occupation AS occupation,
            mejores::movie_id AS movie_id,
            uitem::title AS title,
            mejores::media_rating AS media_rating;

-- Mostrar por pantalla
DUMP top_ocupacion;

-- Guardar en HDFS
STORE top_ocupacion INTO '/user/maria_dev/top_por_ocupacion'
       USING PigStorage('\t');
```

Ahora usando “WinSCP” lo vamos a copiar a la máquina virtual



```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev
archivo_1gb.dat         pig_1763549292308.log  retail_sales_dataset.csv  script_peliculas3.pig  u.data
consultas_usuarios.pig  pig_1763549628980.log  script_completo.pig       script_peliculas4.pig  u.item
el_quijote.txt          practicas.txt          script_peliculas2.pig     script_peliculas.pig   u.user
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_peliculas4.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/
Found 18 items
drwx------   - maria_dev hdfs          0 2025-11-12 10:38 /user/maria_dev/.Trash
drwx------   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/.staging
drwxr-xr-x   - maria_dev hdfs          0 2025-11-11 09:16 /user/maria_dev/datos
-rw-r--r--   1 maria_dev hdfs    2226045 2025-11-17 11:14 /user/maria_dev/el_quijote.txt
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:11 /user/maria_dev/maria_dev
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
-rw-r--r--   1 maria_dev hdfs       1401 2025-11-19 10:13 /user/maria_dev/script_peliculas.pig
-rw-r--r--   1 maria_dev hdfs       1173 2025-11-19 10:46 /user/maria_dev/script_peliculas2.pig
-rw-r--r--   1 maria_dev hdfs       1589 2025-11-19 11:16 /user/maria_dev/script_peliculas3.pig
-rw-r--r--   1 maria_dev hdfs       2122 2025-11-20 07:58 /user/maria_dev/script_peliculas4.pig
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:57 /user/maria_dev/top10_valoradas
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 11:18 /user/maria_dev/top5_antiguas_valoradas
drwxr-xr-x   - maria_dev hdfs          0 2025-11-19 10:21 /user/maria_dev/top5_peliculas
-rw-r--r--   1 maria_dev hdfs    1979173 2025-11-18 09:03 /user/maria_dev/u.data
-rw-r--r--   1 maria_dev hdfs     236344 2025-11-18 09:03 /user/maria_dev/u.item
drwxr-xr-x   - maria_dev hdfs          0 2025-11-17 10:55 /user/maria_dev/ventas_analisis
```

Ahora ya podemos ejecutar el script

```
[maria_dev@sandbox-hdp ~]$ pig script_peliculas4.pig
```

```
(student,23,Taxi Driver (1976),4.378378378378378)
(retired,54,Outbreak (1995),2.0)
(doctor,60,Three Colors: Blue (1993),4.5)
(artist,118,Twister (1996),1.5)
(salesman,231,Batman Returns (1992),4.0)
(technician,236,Citizen Ruth (1996),3.5)
(homemaker,357,One Flew Over the Cuckoo's Nest (1975),1.0)
(none,363,Sudden Death (1995),3.0)
(healthcare,376,Houseguest (1994),5.0)
(lawyer,586,Terminal Velocity (1994),2.0)
(scientist,603,Rear Window (1954),4.6)
(engineer,612,Lost Horizon (1937),4.0)
(programmer,670,Body Snatchers (1993),3.0)
(marketing,716,Home for the Holidays (1995),3.0)
(other,758,Lawnmower Man 2: Beyond Cyberspace (1996),1.5)
(administrator,765,Boomerang (1992),3.0)
(librarian,969,Winnie the Pooh and the Blustery Day (1968),4.0)
(writer,982,Maximum Risk (1996),1.0)
(entertainment,1252,Contempt (M◆pris, Le) (1963),3.0)
(educator,1275,Killer (Bulletproof Heart) (1994),1.0)
(executive,1373,Good Morning (1971),1.0)
```

2.) Muestra la salida del mismo por pantalla y almacena también su resultado en un archivo en HDFS.

Una vez ejercitado el script nos muestra toda la información y el resultado por pantalla fue este;

```
(student,23,Taxi Driver (1976),4.3783783783783783)
(retired,54,Outbreak (1995),2.0)
(doctor,60,Three Colors: Blue (1993),4.5)
(artist,118,Twister (1996),1.5)
(salesman,231,Batman Returns (1992),4.0)
(technician,236,Citizen Ruth (1996),3.5)
(homemaker,357,One Flew Over the Cuckoo's Nest (1975),1.0)
(none,363,Sudden Death (1995),3.0)
(healthcare,376,Houseguest (1994),5.0)
(lawyer,586,Terminal Velocity (1994),2.0)
(scientist,603,Rear Window (1954),4.6)
(engineer,612,Lost Horizon (1937),4.0)
(programmer,670,Body Snatchers (1993),3.0)
(marketing,716,Home for the Holidays (1995),3.0)
(other,758,Lawnmower Man 2: Beyond Cyberspace (1996),1.5)
(administrator,765,Boomerang (1992),3.0)
(librarian,969,Winnie the Pooh and the Blustery Day (1968),4.0)
(writer,982,Maximum Risk (1996),1.0)
(entertainment,1252,Contempt (M◆pris, Le) (1963),3.0)
(educator,1275,Killer (Bulletproof Heart) (1994),1.0)
(executive,1373,Good Morning (1971),1.0)
```

Ahora vamos a ver los resultados en los archivos de HDFS

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/top_por_ocupacion/
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2025-11-20 08:25 /user/maria_dev/top_por_ocupacion/_SUCCESS
-rw-r--r--   1 maria_dev hdfs        893 2025-11-20 08:25 /user/maria_dev/top_por_ocupacion/part-v009-o000-r-00000
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/top_por_ocupacion/part*
student 23      Taxi Driver (1976)      4.3783783783783783
retired 54      Outbreak (1995) 2.0
doctor  60      Three Colors: Blue (1993)       4.5
artist  118     Twister (1996)  1.5
salesman        231     Batman Returns (1992)   4.0
technician      236     Citizen Ruth (1996)     3.5
homemaker       357     One Flew Over the Cuckoo's Nest (1975)  1.0
none    363     Sudden Death (1995)     3.0
healthcare      376     Houseguest (1994)       5.0
lawyer  586     Terminal Velocity (1994)        2.0
scientist       603     Rear Window (1954)      4.6
engineer        612     Lost Horizon (1937)     4.0
programmer      670     Body Snatchers (1993)   3.0
marketing       716     Home for the Holidays (1995)    3.0
other   758     Lawnmower Man 2: Beyond Cyberspace (1996)        1.5
administrator   765     Boomerang (1992)        3.0
librarian       969     Winnie the Pooh and the Blustery Day (1968)     4.0
writer  982     Maximum Risk (1996)     1.0
entertainment   1252    Contempt (M◆pris, Le) (1963)    3.0
educator        1275    Killer (Bulletproof Heart) (1994)       1.0
executive       1373    Good Morning (1971)     1.0
```

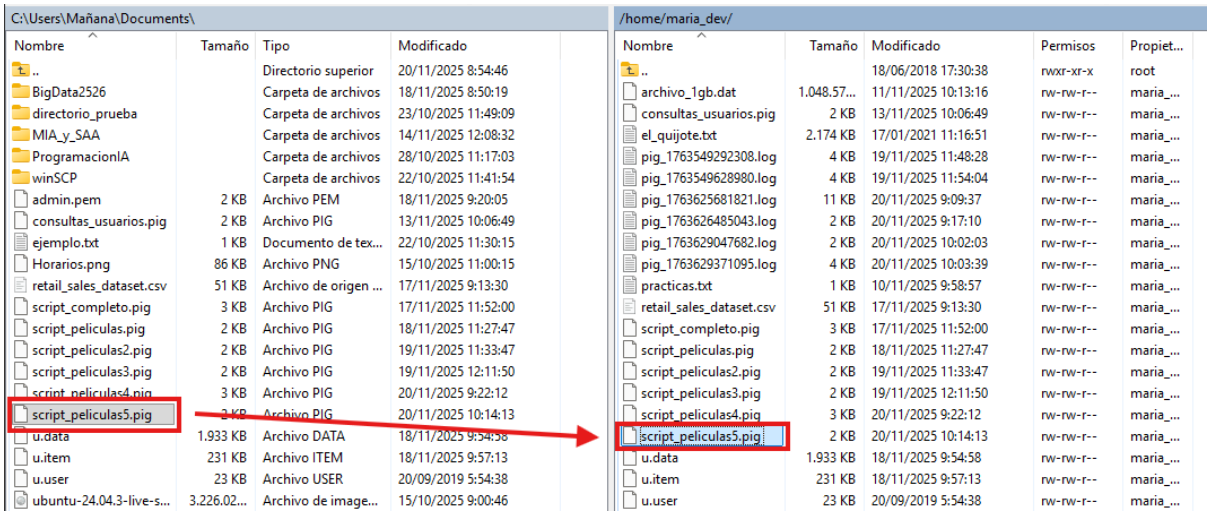
Apartado E

1.) Implementa en PIG el script necesario para obtener la información deseada.

Vamos a crear el script haciendo todo lo que nos pide este enunciado, en mi caso lo voy a hacer en un bloc de notas y sustituir el ".txt" por un ".pig"

```
-- Cargar u.data: user_id, movie_id, rating, timestamp
udata = LOAD '/user/maria_dev/u.data'
        USING PigStorage('\t')
        AS (user_id:int, movie_id:int, rating:int, timestamp:long);

-- Cargar u.item: movie_id | title | otros campos
uitem = LOAD '/user/maria_dev/u.item'
        USING PigStorage('|')
        AS (movie_id:int, title:chararray, info:chararray);

-- Extraer el año desde el título (ej: "Toy Story (1995)")
peliculas = FOREACH uitem GENERATE
        movie_id,
        title,
        (int) REGEX_EXTRACT(title, '.*\\((\\d{4})\\)', 1) AS year;

-- Eliminar registros sin año
peliculas_filtradas = FILTER peliculas BY year IS NOT NULL AND year > 0;

-- Obtener década (por ejemplo: 1995 → 1990)
peliculas_decada = FOREACH peliculas_filtradas GENERATE
        movie_id,
        title,
        (year - (year % 10)) AS decade;

-- Join entre ratings y películas con década
joined = JOIN udata BY movie_id, peliculas_decada BY movie_id;

-- Agrupar por década
grouped = GROUP joined BY decade;

-- Calcular promedio de rating por década
promedios = FOREACH grouped GENERATE
        group AS decade,
        AVG(joined.rating) AS avg_rating;

-- Ordenar por década ascendente
valoraciones_decadas = ORDER promedios BY decade ASC;

-- Mostrar resultado por pantalla
DUMP valoraciones_decadas;

-- Guardar en HDFS como CSV
STORE valoraciones_decadas INTO '/user/maria_dev/valoraciones_por_decada'
      USING PigStorage(',');
```

Ahora usando "WinSCP" lo vamos a copiar a la máquina virtual




```
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev/script*
/home/maria_dev/script_completo.pig    /home/maria_dev/script_peliculas3.pig   /home/maria_dev/script_peliculas5.pig
/home/maria_dev/script_peliculas2.pig  /home/maria_dev/script_peliculas4.pig  /home/maria_dev/script_peliculas.pig
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put /home/maria_dev/script_peliculas5.pig /user/maria_dev/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/script*
-rw-r--r--   1 maria_dev hdfs       1401 2025-11-19 10:13 /user/maria_dev/script_peliculas.pig
-rw-r--r--   1 maria_dev hdfs       1173 2025-11-19 10:46 /user/maria_dev/script_peliculas2.pig
-rw-r--r--   1 maria_dev hdfs       1589 2025-11-19 11:16 /user/maria_dev/script_peliculas3.pig
-rw-r--r--   1 maria_dev hdfs       2489 2025-11-20 08:24 /user/maria_dev/script_peliculas4.pig
-rw-r--r--   1 maria_dev hdfs       1438 2025-11-20 09:16 /user/maria_dev/script_peliculas5.pig
```

Ahora ya podemos ejecutar el script

```
[maria_dev@sandbox-hdp ~]$ pig script_peliculas5.pig
```

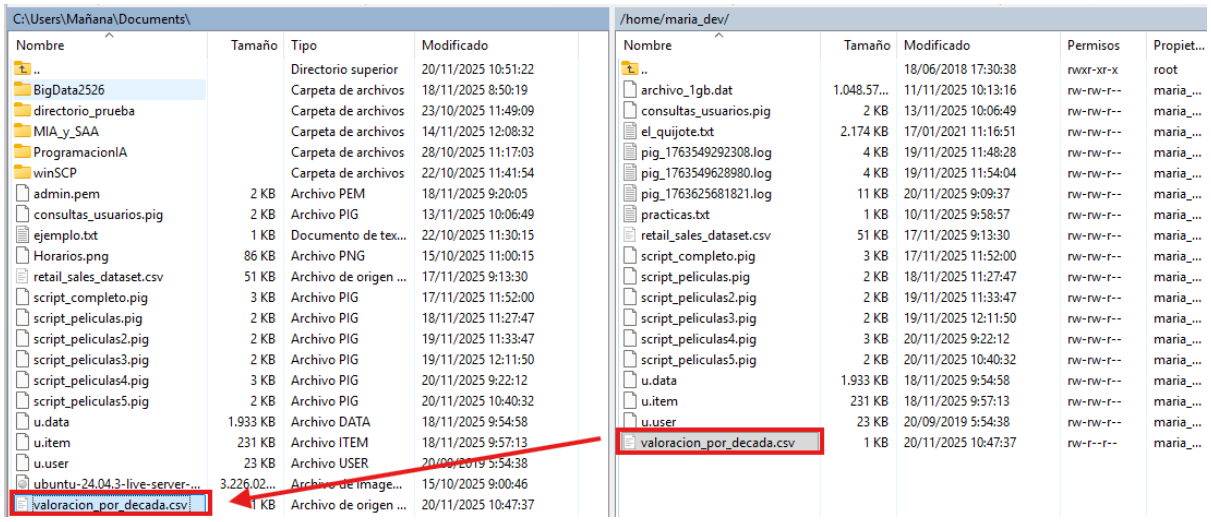
```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/valoraciones_por_decada/
Found 2 items
-rw-r--r--   1 maria_dev hdfs          0 2025-11-20 09:18 /user/maria_dev/valoraciones_por_decada/_SUCCESS
-rw-r--r--   1 maria_dev hdfs        188 2025-11-20 09:18 /user/maria_dev/valoraciones_por_decada/part-v006-o000-r-00000
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/valoraciones_por_decada/part*
1920,3.5357142857142856
1930,3.9251336898395723
1940,4.01067140951534
1950,3.9381910972497876
1960,3.8769506267587617
1970,3.8734491315136474
1980,3.766791335515038
1990,3.389869299861815
```

2.) Abrir el archive en Excel y generar un gráfico de barras con los datos.

Ahora para abrir el archivo en Excel tengo que pasarlo de la máquina virtual al Windows, para ello hago lo siguiente

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -get /user/maria_dev/valoraciones_por_decada/part-v006-o000-r-00000 ~/valoracion_por_deca
da.csv
[maria_dev@sandbox-hdp ~]$ ls /home/maria_dev/valo*
/home/maria_dev/valoracion_por_decada.csv
```

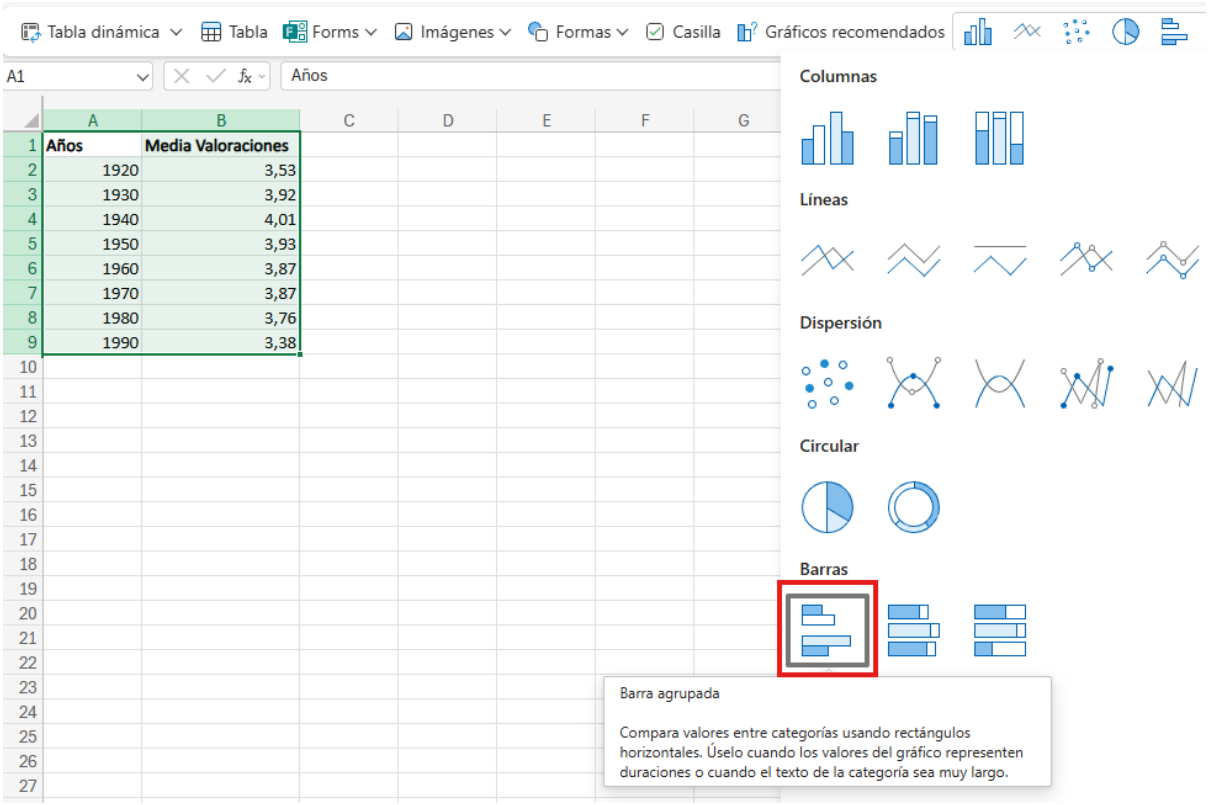
Ahora utilizo el “WinSCP” para pasarlo a mi Windows



Ahora vamos a abrir excel y cargar el documento “.csv”

	A	B
1	Años	Media Valoraciones
2	1920	3,53
3	1930	3,92
4	1940	4,01
5	1950	3,93
6	1960	3,87
7	1970	3,87
8	1980	3,76
9	1990	3,38

Para crear la gráfica tenemos que seleccionar los datos e ir a la sección “Insertar” y después darle a la opción de “graficas”



Años	Media Valoraciones
1920	3,53
1930	3,92
1940	4,01
1950	3,93
1960	3,87
1970	3,87
1980	3,76
1990	3,38

También hay una opción en la cual te muestran graficas recomendadas la cual te puede ser más util si no estás acostumbrado a trabajar con Excel como es mi caso

	A	B	C	D	E	F	G	H	I
1	**Años**	**Media Valoraciones**							
2	1920	3,53							
3	1930	3,92							
4	1940	4,01							
5	1950	3,93							
6	1960	3,87							
7	1970	3,87							
8	1980	3,76							
9	1990	3,38							
10									
11									
12									
13									
14									
15									
16									
17									
18									
19