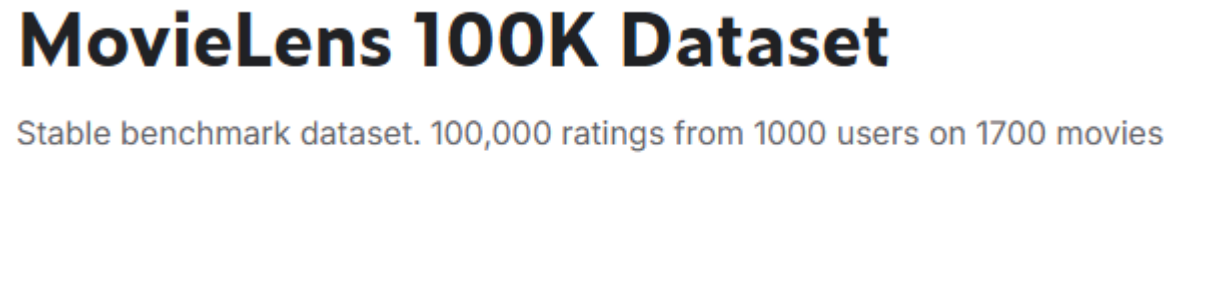
# PR_06.1 Dani Gayol Rodríguez

# Apartado A

**Antes de nada, hay que descargar el conjunto de datos “Movielens” que hay en Kaggle**




**Nos conectamos a la máquina virtual y usamos el comando “hive”**

```
C:\Users\Mañana>ssh maria_dev@localhost -p 2222
maria_dev@localhost's password:
Last login: Tue Dec  2 08:15:57 2025 from 172.18.0.3
[maria_dev@sandbox-hdp ~]$ hive
log4j:WARN No such property [maxFileSize] in org.apache.log4j.DailyRollingFileAppender.

Logging initialized using configuration in file:/etc/hive/2.6.5.0-292/0/hive-log4j.properties
```

**El “hive” no se inicia desde la consola de comando asique lo voy a hacer desde “Ambari”**

## 1.) Crea una base de datos que llamaremos “movielens” para almacenar las tablas necesarias. Para cada una de las consultas deberás crear previamente las tablas y cargar los datos necesarios para poder realizarlas.

**Creamos la base de datos desde “Ambari”**

```
1 CREATE DATABASE IF NOT EXISTS movielens;
2 USE movielens;
```

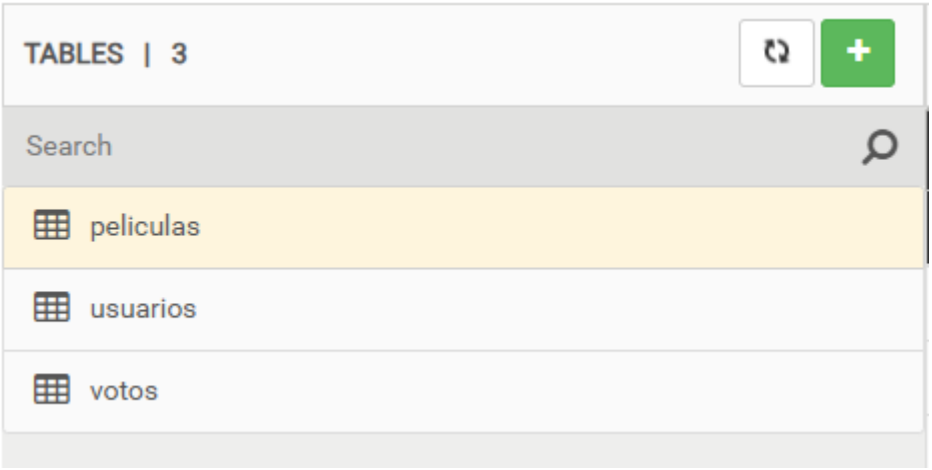
**Una vez creada, la seleccionamos**


**Ahora voy a crear las tablas y cargarle los datos. Para los archivos de “u.user”, “u.data” y “u.item” al ya tenerlos almacenados en una carpeta con la ruta “/user/maria_dev/movielens” de la práctica anterior no hace falta que lo copiemos de nuevo, los voy a pillar de esa ruta**

```
CREATE TABLE usuarios (
    user_id INT,
    age INT,
    gender STRING,
    occupation STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|';
```

```
CREATE TABLE peliculas (
    movie_id INT,
    title STRING,
    anio STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|';
```

```
CREATE TABLE votos (
    user_id INT,
    movie_id INT,
    rating INT,
    fecha BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';
```




**Ahora voy a cargarle los datos a cada tabla**

```
LOAD DATA INPATH '/user/maria_dev/movielens/u.item' INTO TABLE peliculas;
LOAD DATA INPATH '/user/maria_dev/movielens/u.user' INTO TABLE usuarios;
LOAD DATA INPATH '/user/maria_dev/movielens/u.data' INTO TABLE votos;
```

## 2.) Encontrar las 10 ocupaciones más frecuentes entre los votantes

```
SELECT occupation, COUNT(*) AS total
FROM usuarios
GROUP BY occupation
ORDER BY total DESC
LIMIT 10;
```

| occupation | total |
|---|---|
| student | 196 |
| other | 105 |

## 3.) Y luego el número de hombres y mujeres

```
SELECT gender, COUNT(*) AS total
FROM usuarios
GROUP BY gender;
```

| gender | total |
|---|---|
| F | 273 |
| M | 670 |

## 4.) Muestra la edad media por géneros.

Si quieres redondear el número, puedes usar “ROUND” y el número de decimales al que quieres redondear, por ejemplo; ROUND(AVG(age), 2) y te lo redondea a 33.81

```sql
SELECT gender, AVG(age) AS edad_media
FROM usuarios
GROUP BY gender;
```

| gender | edad_media |
|---|---|
| F | 33.81318681318681 |
| M | 34.149253731343286 |

## 5.) Muestra la edad media por ocupaciones.

```sql
SELECT occupation, AVG(age) AS edad_media
FROM usuarios
GROUP BY occupation
ORDER BY edad_media DESC;
```

| occupation | edad_media |
|---|---|
| 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 |

## 6.) Encontrar las cinco películas (código, título y número de votos) más votadas (recuento de votos, no media).

```sql
SELECT p.movie_id, p.title, COUNT(v.movie_id) AS num_votos
FROM peliculas p
JOIN votos v ON p.movie_id = v.movie_id
GROUP BY p.movie_id, p.title
ORDER BY num_votos DESC
LIMIT 5;
```

| p.movie_id | p.title | num_votos |
|---|---|---|
| 50 | Star Wars (1977) | 583 |
| 258 | Contact (1997) | 509 |