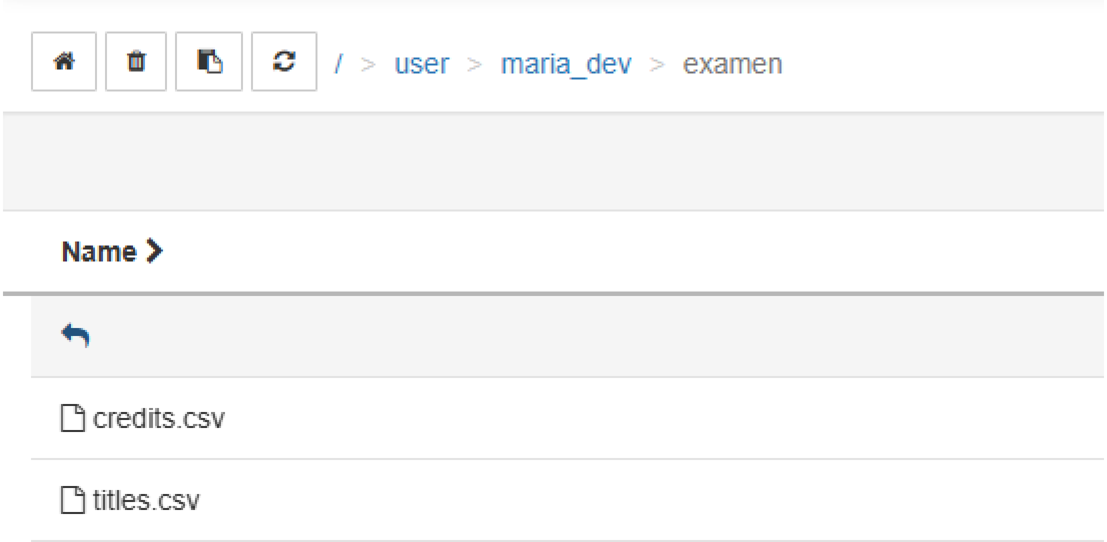
Examen Hive y Pig

Antes de nada subí los dos archivos dados desde entorno grafico (ya que es más sencillo) a hadoop:



Los archivos están en “/user/maria_dev/examen”

Ejercicio 1:

Código:

Crear database:
CREATE DATABASE IF NOT EXISTS netflix; USE netflix;

Creación de tablas:

1. Table Titulos:

```
CREATE TABLE titulos (id STRING,
                                 title STRING,
                                 type STRING,
                                 release_year INT,
                                 age_certification STRING,
                                 runtime INT,
                                 genres STRING,
                                 porduction_countries ARRAY<STRING>,
                                 seasons INT,
                                 imdb_id INT,
                                 imbd_score FLOAT,
                                 imdb_votes FLOAT,
                                 tmdb_popularity FLOAT,
                                 tmdb_score FLOAT) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' LOCATION '/users/maría_dev/examen/titles.csv';
```

2. Tabla creditos

```
CREATE TABLE creditos (person_id INT,
                                 id STRING,
                                 name STRING,
                                 character STRING,
                                 role STRING) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ',' LOCATION '/users/maría_dev/examen/credits.csv';
```

Ahora voy a darles datos:

1. Cargar datos a la tabla títulos:

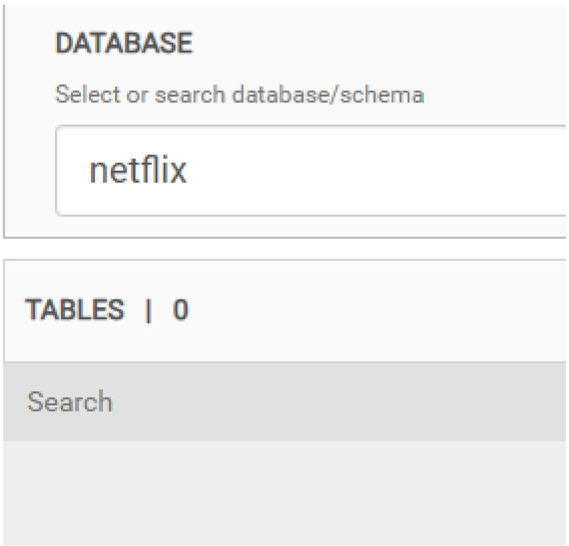
```
LOAD DATA INPATH '/user/maria_dev/examen/titles.csv' INTO TABLE titulos;
```

2. Cargar datos a la tabla créditos:

LOAD DATA INPATH '/user/maria_dev/examen/credits.csv' INTO TABLE creditos;

Salida:

Crear database:



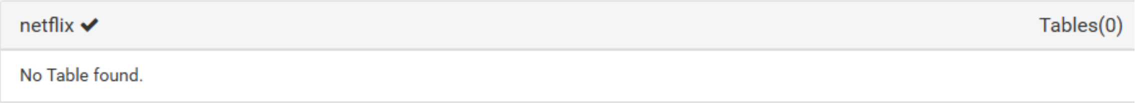
DATABASE

Select or search database/schema

netflix

TABLES | 0

Search

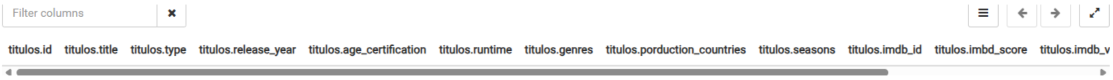


netflix ✓ Tables(0)

No Table found.

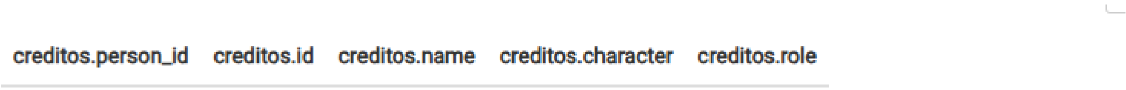
Creacion de tablas:

1. Tabla títulos:



titulos.id	titulos.title	titulos.type	titulos.release_year	titulos.age_certification	titulos.runtime	titulos.genres	titulos.porduction_countries	titulos.seasons	titulos.imdb_id	titulos.imbd_score	titulos.imdb_v

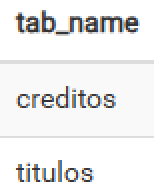
2. Tabla créditos:



creditos.person_id	creditos.id	creditos.name	creditos.character	creditos.role

Las dos tablas:

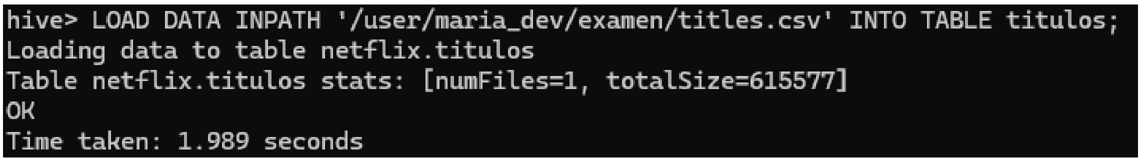
Hice un show tables; para mostrar las tablas



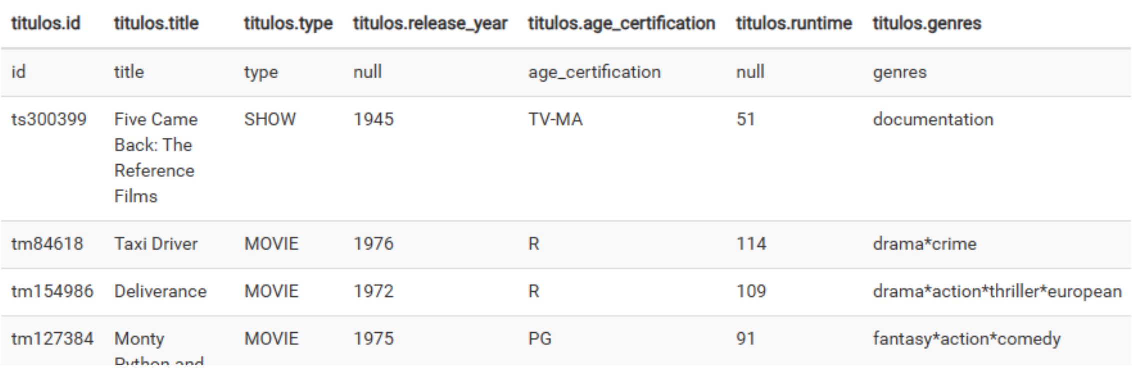
tab_name
creditos
titulos

Dar datos a las tablas:

1. Tabla títulos:

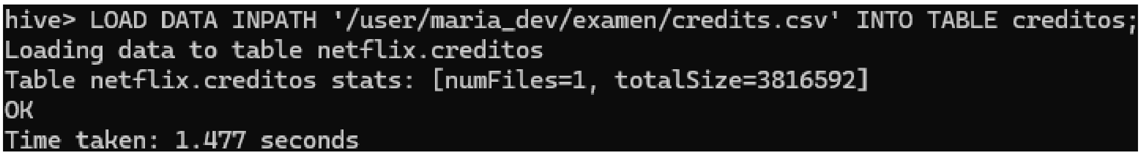


```
hive> LOAD DATA INPATH '/user/maria_dev/examen/titles.csv' INTO TABLE titulos;
Loading data to table netflix.titulos
Table netflix.titulos stats: [numFiles=1, totalSize=615577]
OK
Time taken: 1.989 seconds
```

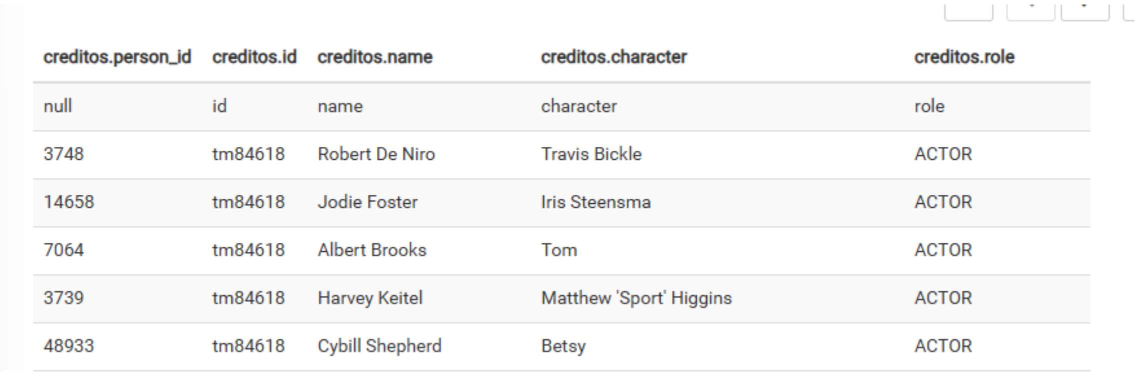


titulos.id	titulos.title	titulos.type	titulos.release_year	titulos.age_certification	titulos.runtime	titulos.genres
id	title	type	null	age_certification	null	genres
ts300399	Five Came Back: The Reference Films	SHOW	1945	TV-MA	51	documentation
tm84618	Taxi Driver	MOVIE	1976	R	114	drama*crime
tm154986	Deliverance	MOVIE	1972	R	109	drama*action*thriller*european
tm127384	Monty Python and	MOVIE	1975	PG	91	fantasy*action*comedy

2. Tabla créditos:



```
hive> LOAD DATA INPATH '/user/maria_dev/examen/credits.csv' INTO TABLE creditos;
Loading data to table netflix.creditos
Table netflix.creditos stats: [numFiles=1, totalSize=3816592]
OK
Time taken: 1.477 seconds
```



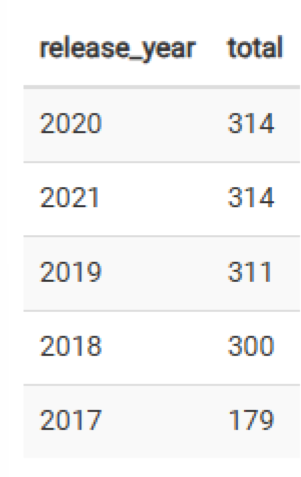
creditos.person_id	creditos.id	creditos.name	creditos.character	creditos.role
null	id	name	character	role
3748	tm84618	Robert De Niro	Travis Bickle	ACTOR
14658	tm84618	Jodie Foster	Iris Steensma	ACTOR
7064	tm84618	Albert Brooks	Tom	ACTOR
3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
48933	tm84618	Cybill Shepherd	Betsy	ACTOR

Ejercicio 2:

Codigo:

```
SELECT release_year, COUNT(*) as total  from titulos WHERE type == "SHOW"
GROUP BY release_year ORDER BY total DESC LIMIT 5;
```

Salida:



release_year	total
2020	314
2021	314
2019	311
2018	300
2017	179

Ejercicio 3:

Código:

```
SELECT genre, COUNT(*) FROM titulos LATERAL

VIEW EXPLODE(SPLIT(genres,'*')) g AS genre WHERE type == 'MOVIE' GROUP BY
genre LIMIT 10;
```

Salida:

Me salta un error en la línea de la peli “Taxi Driver”

Ejercicio 4:

FALTA FILTAR PARA QUE APAREZCAN LOS QUE TIENEN 2 O MAS

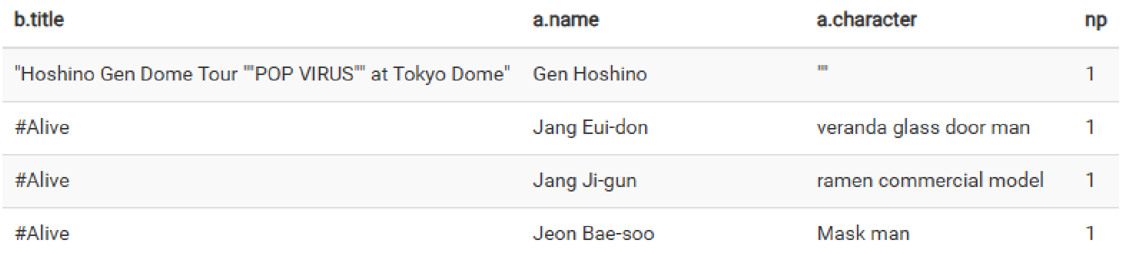
Código:

```
select b.title,a.name,a.character,COUNT(*) as np from creditos as a JOIN titulos as
b on a.id = b.id

WHERE a.role == 'ACTOR'

GROUP BY b.title,a.name,a.character;
```

Salida:



b.title	a.name	a.character	np
"Hoshino Gen Dome Tour ""POP VIRUS"" at Tokyo Dome"	Gen Hoshino	""	1
#Alive	Jang Eui-don	veranda glass door man	1
#Alive	Jang Ji-gun	ramen commercial model	1
#Alive	Jeon Bae-soo	Mask man	1

Ejercicio 5:

NO ME DIO TIEMPO A EJECUTAR

Código:

```
select b.title,COUNT(*) as na from creditos as a JOIN titulos as b on a.id = b.id

WHERE a.role == 'ACTOR'

GROUP BY b.title LIMIT 5;
```

Salida: