

**Análisis Movie Data**

**Integrantes:**

Carlos Contreras Rojas.

Fernando Gutiérrez Toro.

Mónica Palma Segovia.

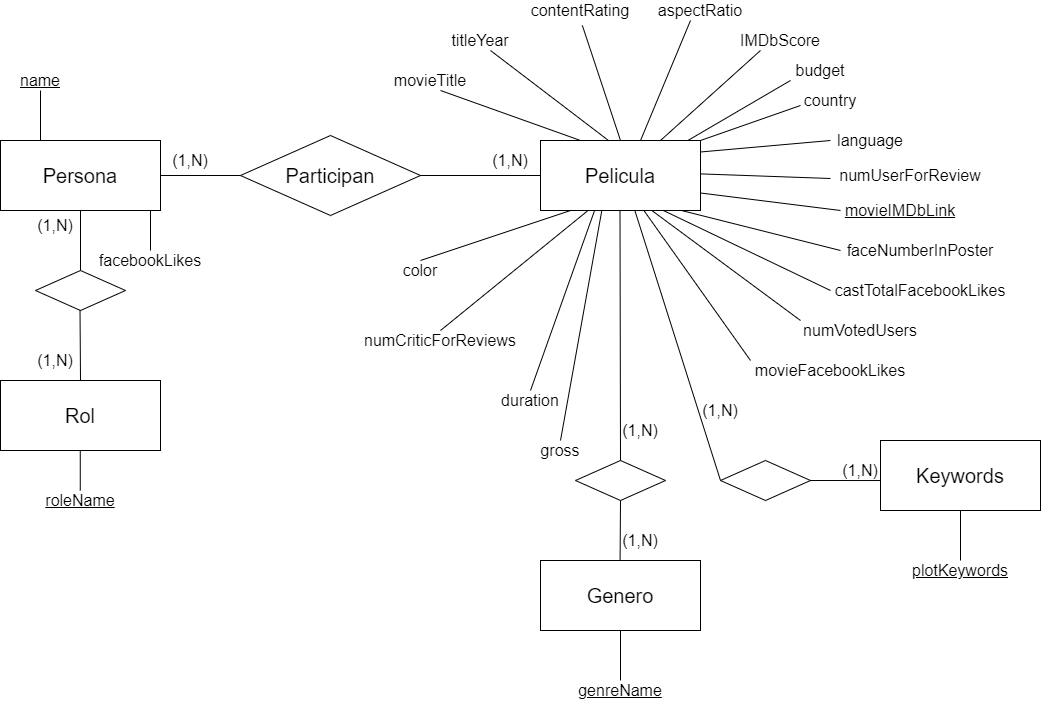
**Profesor:**

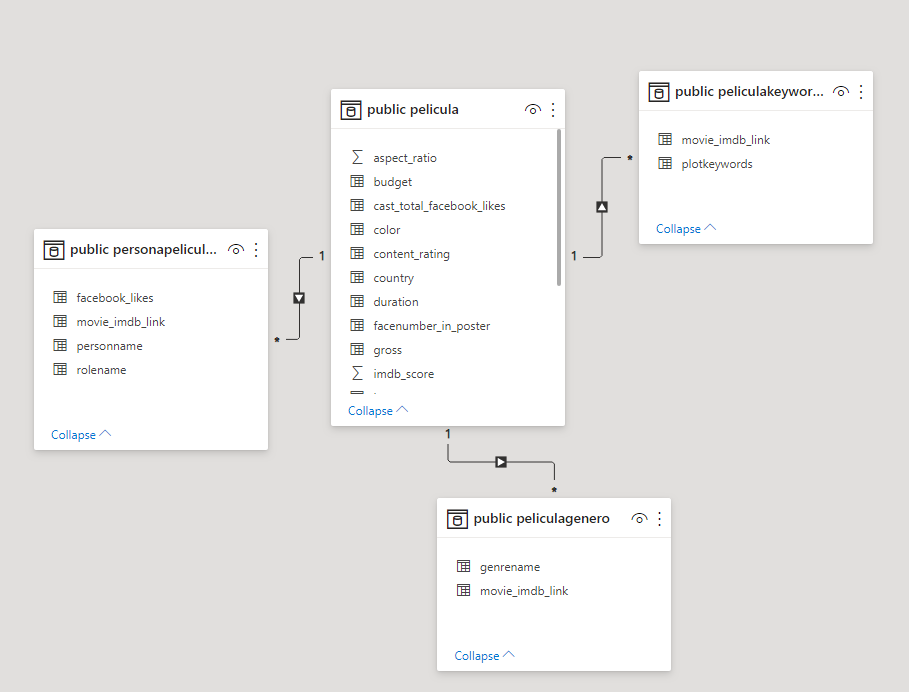
Mario Ortiz Bonilla.

**Asignatura:**

Pr. Procesamiento Analítico de Datos.

1. **Modelo**

****

Decidimos juntar las tablas “personas” y “rol” en una sola tabla.

1. **PROCESO ETL**

**Análisis por columna**

Por cada columna del excel “moviedata.xlsx”, podemos encontrar los siguientes datos, por ejemplo, en la columna “color”, existen las palabras: Black and White – #ERROR – Color – N/A

A continuación, se muestran los datos que podemos encontrar por cada columna y los errores que puedan tener:

director\_name: #NODATA – palabras con caracteres especiales (Ã³) – palabras con caracteres normales (letras, puntos)

num\_critic\_for\_reviews: N/A – números enteros

duration: #NODATA – números

director\_facebook\_likes: N/I – numeros enteros

actor\_3\_facebook\_likes: N/I – numeros enteros

actor\_2\_name: N/A – palabras con caracteres especiales (Ã³) - palabras con caracteres normales (letras, puntos, numeros)

actor\_1\_facebook\_likes: números enteros – vacíos

gross: números enteros – vacíos

genres: palabras normales separadas por (|) – palabras normales

actor\_1\_name: N/A – palabras con caracteres especiales (Ã³) - palabras con caracteres normales (letras, puntos, numeros) – vacíos

movie\_title: palabras con caracteres especiales (Ã³) - palabras con caracteres normales (letras, puntos, numeros) – palabras que terminan con Â

num\_voted\_users: numeros enteros

cast\_total\_facebook\_likes: numeros enteros

actor\_3\_name: palabras con caracteres especiales (Ã³) - palabras con caracteres normales (letras, puntos, numeros) – vacíos

facenumber\_in\_poster: numeros enteros – vacios

plot\_keywords: palabras normales separadas por (|) – palabras normales – vacíos

movie\_imdb\_link: todos son un link url

num\_user\_for\_reviews: numeros enteros – vacios

language: N/A – None – palabras normales

country: el signo “?” – palabras normales

content\_rating: palabras con caracteres normales - el signo “?” no es valido

budget: numeros enteros – el numero “-1” no es valido

title\_year: numeros enteros – el numero “0” no es valido

actor\_2\_facebook\_likes: numeros enteros (el 0 es válido) – vacíos

imdb\_score: numeros enteros y decimales, pueden ser entre el 10 y el 1

aspect\_ratio: numeros enteros y decimales – vacíos

movie\_facebook\_likes: numeros enteros (el 0 es valido)

También encontramos las siguientes inconsistencias:

- Algunas filas se repiten.

- Distintos datos que representan lo mismo en la columna content\_rating (ver tabla de conversion).

**Limpiar los datos**

Se les realizaran los siguientes cambios a los datos:

- Borrar las filas sin director (director = #NODATA).

- Quitarle el último carácter al título de las películas (Â).

- Columna Budget: reemplazar "-1" con NULL.

- Columna content\_rating: usar la taba para correcciones.

- Columna aspect\_ratio: corregir los valores inválidos.

- Columnas con likes en vacío, se cambian a NULL

- hay años 0, se cambian a NULL