

**Diseño y explotación de la base de datos MoviesDB**

Luis Aguilar, laaguilar21@utpl.edu.ec **(estudiante líder)**

Steven Vega, srvega4@utpl.edu.ec

Marco Herrera, mshm1296@hotmail.com

15 de enero de 2025

**Índice de contenidos**

[1. Diseño y explotación de la base de datos MoviesDB 3](#_Toc187847308)

[1.1. Obtención de datos 3](#_Toc187847309)

[1.1.1. Análisis descriptivo del dataset 3](#_Toc187847310)

[1.1.2. Descripción de datos del dataset 3](#_Toc187847311)

[1.1.3. Descripción de los requisitos de datos para MoviesDB 3](#_Toc187847312)

[1.2. Modelado Conceptual 4](#_Toc187847313)

[1.2.1. Proceso de modelado conceptual 4](#_Toc187847314)

[1.2.2. Resumen del modelo conceptual 4](#_Toc187847315)

[1.3. Modelado Lógico 4](#_Toc187847316)

[1.3.1. Proceso de modelado lógico 4](#_Toc187847317)

[1.3.2. Diccionario de Datos 4](#_Toc187847318)

[1.4. Modelado Físico 5](#_Toc187847319)

[1.4.1. Creación de la base de datos 5](#_Toc187847320)

[1.4.2. Inserción de datos 5](#_Toc187847321)

[1.5. Validación y Explotación de la base de datos 5](#_Toc187847322)

[1.5.1. Consultas 5](#_Toc187847323)

[1.5.2. Inserción de datos 5](#_Toc187847324)

[2. Referencias Bibliográficas 5](#_Toc187847325)

# Diseño y explotación de la base de datos MoviesDB

El proyecto consiste en diseñar y desarrollar una base de datos llamada \*\*MoviesDB\*\* a partir de un dataset en formato CSV con información sobre películas. Las etapas incluyen la extracción y análisis de datos, el diseño del modelo conceptual para identificar entidades y relaciones, la creación del modelo lógico con tablas y relaciones normalizadas, el desarrollo del modelo físico mediante la implementación de la base de datos en SQL, y finalmente la validación y explotación de la base con consultas e inserciones para garantizar su funcionalidad y utilidad.

## Obtención de datos

### Análisis descriptivo del dataset

|  |  |
| --- | --- |
| **Nombre del Dataset** | Movies pi\_movies\_complete |
| **Objetivo del Dataset** | Gestionar y analizar información sobre películas, actores, equipos técnicos, calificaciones y palabras clave para estudios de popularidad, rentabilidad y otros análisis relacionados con la industria cinematográfica. |
| **Cuántas películas están incluidas** | Por determinar (se debe contar el total de registros únicos en la tabla Pelicula). |
| **Cuáles son los géneros más comunes y menos presentados** | Los géneros más comunes y menos presentados pueden analizarse explorando el campo genres en la tabla Pelicula y calculando la frecuencia de cada género. |
| **Cuántas películas se producen por año** | Se puede obtener al agrupar los datos por año usando el campo release\_date. Esto permite identificar tendencias anuales en la producción cinematográfica. |
| **Se define si hay más de un director de películas incluido** | Sí, se puede verificar revisando los datos en la tabla Crew donde el campo Job es "Director". |
| **Cuáles otros datos o metadatos se incluyen en el Dataset** | Información sobre presupuesto, recaudación, duración, idiomas, países de producción, palabras clave, calificaciones de usuarios, popularidad y colecciones asociadas. |
| **Qué información podría analizar con este Dataset (al menos 3 ideas)** | 1. Relación entre presupuesto y recaudación.  2. Comparación de géneros más populares por región o año.  3. Análisis de tendencias en duración de películas a lo largo de los años. |
| **Con qué datos de películas que encuentran en sitios como - Netflix, IMBD, TMBD- se relaciona el dataset** | Se relaciona con métricas como popularity, vote\_average, genres, cast, crew, y otros metadatos comunes en plataformas como IMDb y TMDB. |

### Descripción de datos del dataset

| **Columna** | **Tipo de Dato** | **Descripción** |
| --- | --- | --- |
| Adult | boolean | Identifica si la pelicula es para adultos o no. |
| belong\_to\_collection | JSON | Identifica a que grupo de peliculas pertenece una pelicula |
| budget | Double | Presupuesto de la pelicula |
| genres | JSON | Género o categorías asociadas. |
| homepage | String | Ruta de la pelicula a una pagina web con informacion sobre ella |
| id | String | Identificador de la pelicula |
| imdb\_id | String | Identificador de la base de datos de la pelicula |
| original\_languague | String | Idioma original de la pelicula |
| original\_title | String | Titulo original de la pelicula |
| overview | String | Descripcion de la pelicula |
| popularity | String | Popularidad de la pelicula |
| poster\_Path | String | Ruta del poster de la coleccion |
| production\_companies | JSON | Compañias productoras de la pelicula |
| production\_countries | JSON | Pais o paises en el que se produce la pelicula |
| release\_date | String | Fecha de lanzamiento de la pelicula |
| revenue | String | Ganancias de la pelicula |
| runtime | String | Duracion de la pelicula |
| spoken\_languagues | JSON | Idiomas al que esta doblada la pelicula |
| status | String | Estado de la pelicula |
| tagline | String | Representacion leve de la pelicula |
| title | String | Ruta del póster de la colección. |
| video | boolean | Si contiene o no un video |
| vote\_average | Double | Promedio de votos |
| vote\_count | Int | Conteo de votos |
| keywords | JSON | Las palabras clave de la pelicula |
| cast | JSON | Los actores que intervienen dentro de la pelicula con su respectivo orden, identificadores y personaje al que interpretan |
| crew | JSON | El equipo tecnico detras de la pelicula que ayuda en su produccion y desarrollo |
| ratings | JSON | La validacion y aceptacion por parte de los usuarios de la pelicula |

### Descripción de los requisitos de datos para MoviesDB

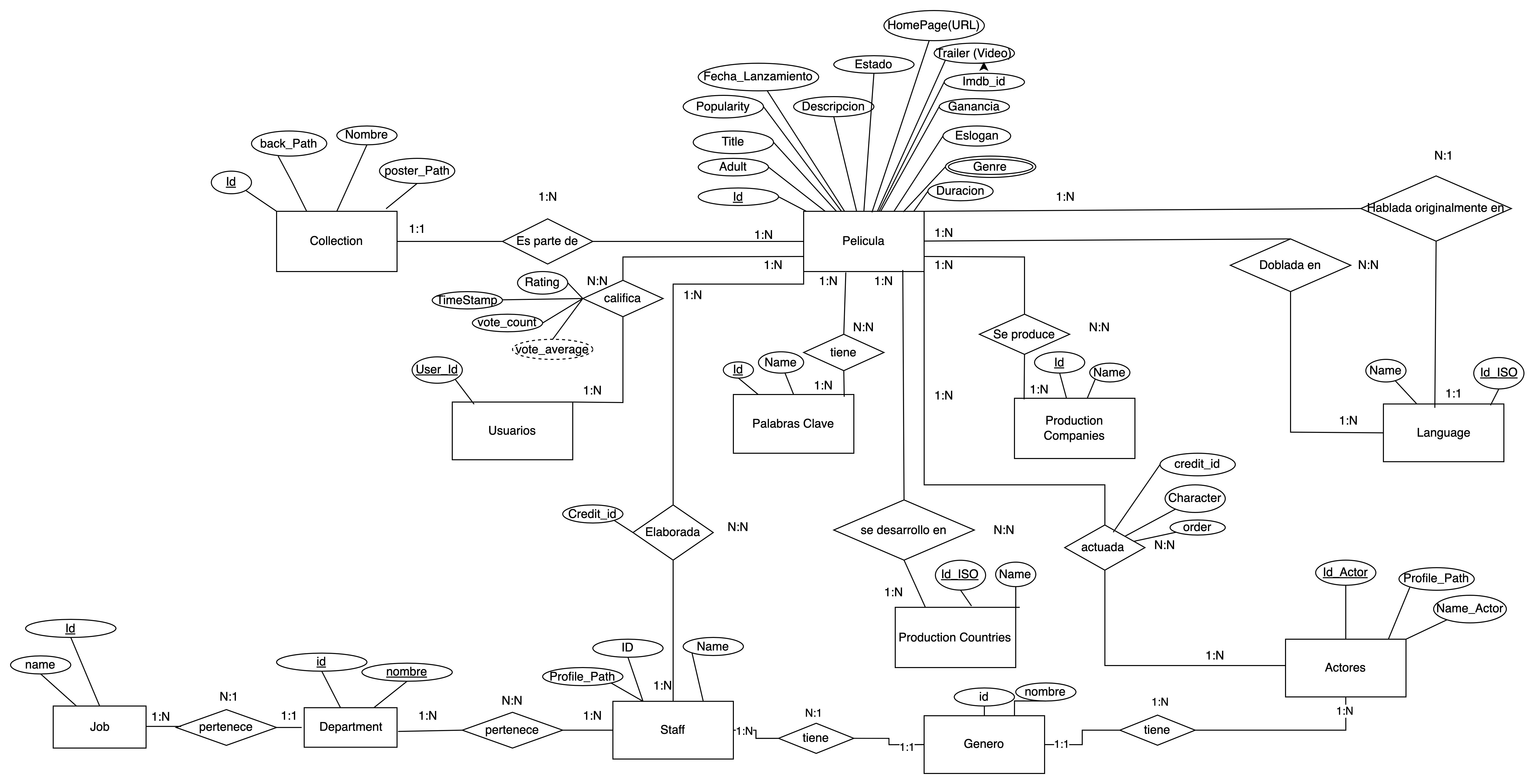
Para crear una base de datos a partir del dataset, es esencial identificar tablas principales: películas, géneros, elenco, equipo técnico, producción, palabras clave y calificaciones. Las relaciones entre estas tablas deben ser normalizadas, destacando relaciones muchos a muchos, como entre películas y géneros, o películas y elenco. Es necesario extraer datos anidados en formato JSON para tablas intermedias, como movie\_genres o movie\_cast, asegurando integridad con claves primarias y foráneas en todas las tablas relacionadas.

Además, la base debe garantizar consistencia mediante validación de formatos (como fechas o identificadores únicos) y la creación de índices en campos clave para optimizar consultas. Tablas como ratings requieren relaciones uno a muchos con películas, mientras que producción y países involucrarán múltiples relaciones. Este diseño permitirá manejar datos complejos como palabras clave y miembros del equipo técnico de manera estructurada, facilitando consultas robustas y escalabilidad en la gestión de información cinematográfica.

## Modelado Conceptual

### Proceso de modelado conceptual

Este modelo conceptual se creó analizando cuidadosamente los datos disponibles, identificando las entidades principales como "Películas", "Géneros", "Palabras Clave", "Actores" y "Staff", que representan los conceptos clave del sistema. Cada una de estas entidades contiene información específica, como el título de la película, su popularidad, o los nombres de actores y miembros del equipo técnico. Para representar las conexiones entre estas entidades, se identificaron relaciones importantes, como "una película tiene varios géneros" o "una película puede ser actuada por muchos actores". Además, se procesaron datos más complejos, como las listas de géneros o palabras clave, separándolos en tablas propias para organizar mejor la información. Esto asegura que los datos estén bien estructurados, sin duplicaciones y listos para ser utilizados.



### Resumen del modelo conceptual

|  |  |
| --- | --- |
| **Cantidad** | |
| Entidades | 11 |
| Relaciones | 10 |

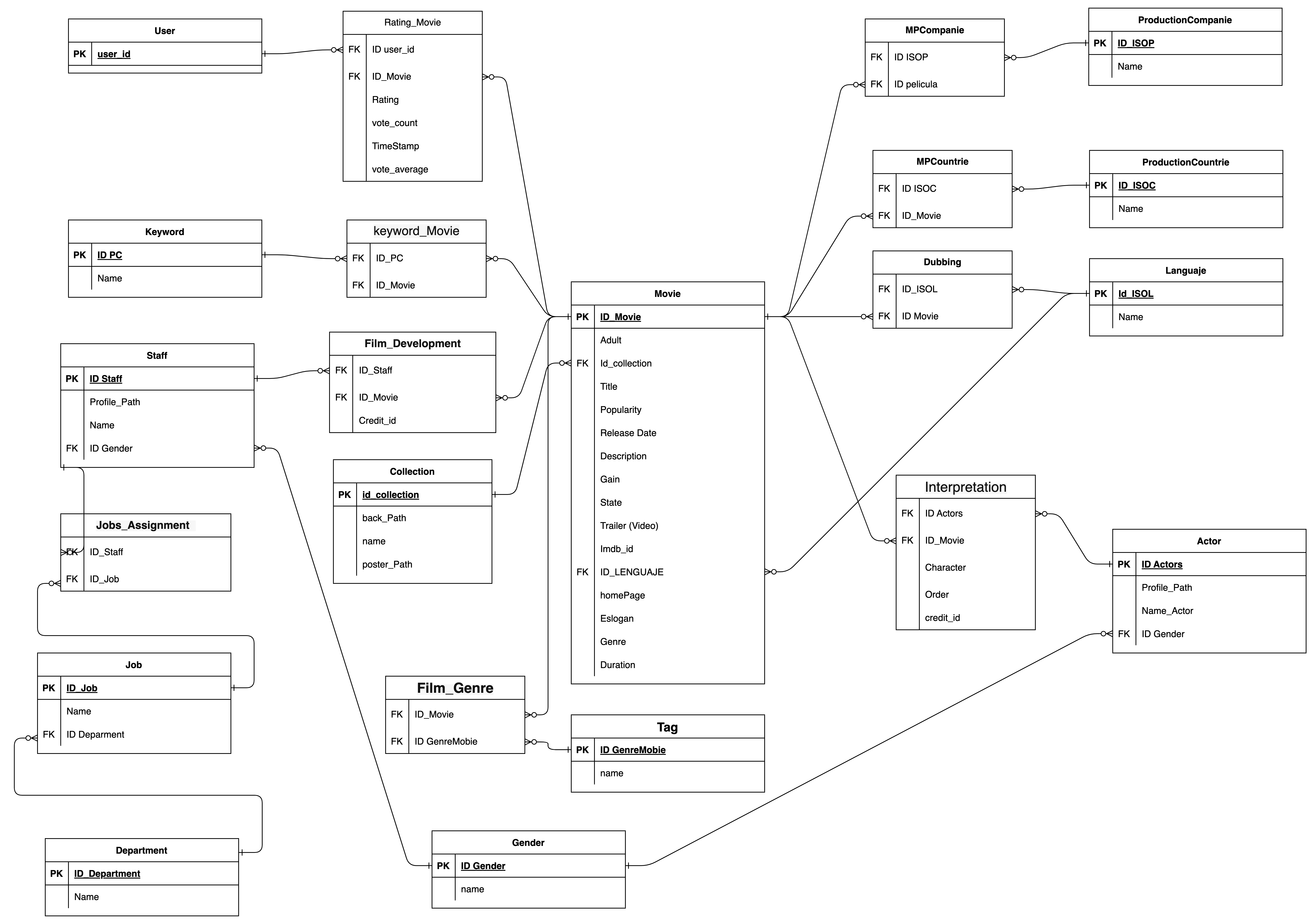
|  |  |
| --- | --- |
| **Atributos de entidades** | |
| **Entidad** | **Atributos** |
| Película | Id, Title, Descripción, Fecha\_Lanzamiento, Popularity, Eslógan, Duración, Estado, Adult, Imdb\_id, HomePage, Ganancia, trailer, Genre |
| Collection | Id, Nombre, back\_Path, poster\_Path |
| Palabras Clave | Id, Name |
| Production Companies | Id, Name |
| Production Countries | Id\_ISO, Name |
| Language | Id\_ISO, Name |
| Actores | Id\_Actor, Name\_Actor, Profile\_Path |
| Staff | ID, Name, Profile\_Path |
| Usuarios | User\_Id |
| Department | id, nombre |
| Job | Id, name |

|  |  |
| --- | --- |
| **Atributos de relación** | |
| Relación | Atributos |
| Es parte de | Id, Nombre |
| Tiene | Id, Name |
| Se produce | Id, Name |
| Se desarrolló en | Id\_ISO, Name |
| Doblada en | Id\_ISO, Name |
| Hablada originalmente en | Id\_ISO, Name |
| Actuada | Credit\_id, Character, order, Id\_Actor |
| Elaborada | Credit\_id |
| Califica | TimeStamp, vote\_count, vote\_average, Rating |
| Pertenece | Id, name |

## Modelado Lógico

### Proceso de modelado lógico

Cada entidad del modelo conceptual (como "Película", "Género", "Actor") se convirtió en una tabla en el modelo lógico. Los atributos identificados en el modelo conceptual (como Title, Popularity o Name) ahora son columnas en las tablas correspondientes. Las relaciones identificadas en el modelo conceptual, como "una película tiene muchos géneros" o "una película es actuada por varios actores", fueron representadas mediante tablas intermedias o "join tables". Por ejemplo, las relaciones de muchos a muchos entre películas y géneros resultaron en la tabla Film\_Genre, mientras que la relación entre películas y palabras clave generó la tabla keyword\_Movie. Las entidades y tablas intermedias se enriquecieron con llaves primarias (por ejemplo, ID en las tablas principales) y llaves foráneas (como movie\_id en Film\_Genre) para garantizar la integridad referencial entre las tablas.



Se aplicó un proceso de normalización para descomponer datos complejos y eliminar redundancias. Por ejemplo, en lugar de almacenar directamente los países o compañías de producción en la tabla de películas, se crearon tablas separadas como ProductionCountrie y ProductionCompanie vinculadas a la tabla principal de películas mediante relaciones.Algunas relaciones más específicas, como el personal (Staff) y sus asignaciones de trabajo (Jobs\_Assignment), se manejaron mediante la creación de tablas adicionales que modelan detalles específicos, como departamentos y roles en las películas.

### Diccionario de Datos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[Nombre de Tabla]** | | | | |
| **Atributo** | **Tipo de Dato** | **Dominio** | **Descripción** | **Restricciones** |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **[Nombre de Tabla]** | | | | |
| **Atributo** | **Tipo de Dato** | **Dominio** | **Descripción** | **Restricciones** |
|  |  |  |  |  |
|  |  |  |  |  |

## Modelado Físico

### Creación de la base de datos

Explicación breve del proceso

Sentencias DDL de la creación de la base de datos

### Inserción de datos

Explicación breve del proceso

Adjuntar (en Canvas) un sólo script que contenga tanto la creación y la población de la base de datos.

## Validación y Explotación de la base de datos

Validar con 10 operaciones: 9 consultas y sus resultados y 1 inserción completa de datos de películas.

### Consultas

Deben ser:

* multitabla mínimo combinar tres tablas
* incluir subconsultas
* consultas con grupos de datos (GROUP BY)

Por cada consulta coloque descripción, SQL y resultado.

### Inserción de datos

* Realizar la inserción de datos de una película completa

Finalmente, genere un script de las consultas e inserción y anéxelo a la entrega en canvas.

# Referencias Bibliográficas

Coloque aquí las referencias bibliográficas utilizadas en formato [APA 7ma edición](https://drive.google.com/file/d/1DKSsMRg4eD4-9CctrqEgnK_YrebwYZAA/view?usp=drive_link)