ANÁLISIS DE TENDENCIAS Y ÉXITO EN PELÍCULAS DE TERROR: UN ESTUDIO BASADO EN DATOS DE TMDB

JUAN ESTEBAN MONSALVE CASTRILLÓN

TECNOLÓGICO DE ANTIOQUIA INSTITUCIÓN UNIVERSITARIA

BASES DE DATOS AVANZADAS

MEDELLÍN

2024

TABLA DE CONTENIDO

TABLA DE CONTENIDO 2
ÍNDICE DE FIGURAS 3
ÍNDICE DE TABLAS 5
INTRODUCCIÓN 6
JUSTIFICACIÓN 7
OBJETIVOS 8
CAPÍTULO 1: COMPRENSIÓN DEL NEGOCIO. 9
CAPÍTULO 2: ENTENDIMIENTO DE LOS DATOS. 11
CAPÍTULO 3: PREPARACIÓN DE LOS DATOS. 14
CAPÍTULO 4: MODELADO. 22
CAPÍTULO 5: EVALUACIÓN. 23
CAPÍTULO 6: EVALUACIÓN Y DESPLIEGUE 42
CONCLUSIONES 50
REFERENCIAS. 52

ÍNDICE DE FIGURAS

FIGURA 1. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 1 14
FIGURA 2. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 2 14
FIGURA 3. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 3 15
FIGURA 4. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 4 16
FIGURA 5. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 5 16
FIGURA 6. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 17
FIGURA 7. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 17
FIGURA 8. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 17
FIGURA 9. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 18
FIGURA 10. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 18
FIGURA 11. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 6 18
FIGURA 12. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 7 19
FIGURA 13. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 8 19
FIGURA 14. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 8 20
FIGURA 15. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 8 20
FIGURA 16. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 8 21
FIGURA 17. CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA. PASO 9 21
FIGURA 18. MODELADO ANALÍTICO DE DATOS 22
FIGURA 19. DIAGRAMA DE LA BASE DE DATOS 23
FIGURA 20. FLUJO DE CONTROL 24
FIGURA 21. TAREA A EJECUTAR SQL. QUERY 25
FIGURA 22. ETL_DIM_LANGUAGE 26
FIGURA 23. ADMINISTRADOR DE ARCHIVOS PLANOS DIM_LANGUAGE. GENERAL 27
FIGURA 24. ADMINISTRADOR DE ARCHIVOS PLANOS DIM_LANGUAGE. COLUMNAS 28
FIGURA 25. ADMINISTRADOR DE ARCHIVOS PLANOS DIM_LANGUAGE. OPCIONES AVANZADAS 29
FIGURA 26. ORIGEN DE ARCHIVO PLANO 30
FIGURA 27. LOOKUP TRASNFORMATION. ETL_DIM_LANGUAGE 31
FIGURA 28. LOOKUP TRASNFORMATION. ETL_DIM_LANGUAGE 32
FIGURA 29. DESTINO OLE DB. ASIGNACIONES 33
FIGURA 30. ELT_DIM_GENRE 34
FIGURA 31. ETL_DIM_DATE 34
FIGURA 32. ETL_FACTS_MOVIES 35
FIGURA 33. ORIGEN DEL ARCHIVO PLANO FACTS_MOVIES. COLUMNAS 36
FIGURA 34. LOOKUP TRANSFORMATION DIM_DATE 37
FIGURA 35. LOOKUP TRANSFORMATION DIM_GENRE 38
FIGURA 36. LOOKUP TRANSFORMATION DIM_LANGUAGE 39
FIGURA 37. DESTINO OLE DB PARA FACTS_MOVIES 40
FIGURA 38. DESTINO OLE DB PARA FACTS_MOVIES. ASIGNACIONES 41
FIGURA 39. PROYECTO MULTIDIMENSIONAL 42
FIGURA 40. DIAGRAMA DEL CUBO 43

FIGURA 41. DWHORRORMOVIES1 Y PROYECTOMULTIDIMENSIONAL_HORROR_MOVIES EN SQL 44
FIGURA 42. FILTROS PARA LOS DATOS EN POWER BI 45
FIGURA 43. DIMENSIONES Y TOTALES FINALES 46
FIGURA 44. DASHBOARD. PÁGINA 1 47
FIGURA 45. DASHBOARD. PÁGINA 2 47
FIGURA 46. DASHBOARD. PÁGINA 3 48
FIGURA 47. DASHBOARD. PÁGINA 4 48
FIGURA 48. DASHBOARD. PÁGINA 5 49
FIGURA 49. DASHBOARD. PÁGINA 6 49

ÍNDICE DE TABLAS

DICCIONARIO DE LA BASE DE DATOS……..…….12

INTRODUCCIÓN

El presente proyecto se centra en la limpieza y preparación de un conjunto de datos relacionado con películas de terror, extraído mediante la API de The Movie Database (TMDB). El propósito de este proyecto es organizar, filtrar y depurar los datos obtenidos para garantizar su utilidad en análisis futuros.

En este proyecto se abordarán varios aspectos clave, como la identificación y eliminación de valores nulos, vacíos o inconsistentes, la corrección de datos erróneos mediante el promedio, y la normalización de información crítica como id, títulos, fechas de lanzamiento, y géneros. Además, se hará especial énfasis en la preparación de los datos para su eventual utilización en tareas como análisis de tendencias, recomendaciones de películas, y estudios de popularidad dentro del género de terror.

A lo largo del documento, se presentará la justificación, el objetivo general y específicos de la limpieza de datos, la comprensión de negocio, el entendimiento de los datos, las técnicas empleadas para llevar a cabo la limpieza, el procedimiento hecho mediante código y los resultados obtenidos, con el fin de proporcionar un conjunto de datos limpio, consistente y adecuado para cualquier tipo de análisis posterior.

JUSTIFICACIÓN

El análisis y limpieza de datos de películas de terror es un paso fundamental en el proceso de Inteligencia de Negocios (BI) dentro de la industria del entretenimiento. En un mercado donde las preferencias de los espectadores están en constante evolución, disponer de datos precisos y consistentes es clave para tomar decisiones informadas que puedan impactar positivamente en la rentabilidad y competitividad de las empresas.

El género de terror, en particular, posee una audiencia dedicada y específica que presenta patrones de comportamiento únicos. Entender estos patrones a través de datos limpios y bien estructurados permite a las empresas de entretenimiento anticipar tendencias, personalizar recomendaciones, y optimizar estrategias de marketing. Además, el análisis de estos datos puede revelar insights valiosos sobre las preferencias de los espectadores, facilitando la creación de contenido que resuene mejor con su audiencia objetivo.

La limpieza de datos asegura que las decisiones basadas en estos análisis sean confiables, evitando que errores o inconsistencias en los datos lleven a conclusiones incorrectas. Esto es especialmente relevante en la implementación de modelos predictivos y algoritmos de recomendación, donde la calidad de los datos es directamente proporcional a la precisión y efectividad de los resultados obtenidos.

Por tanto, este proyecto de limpieza de datos no solo es un requisito técnico, sino una necesidad estratégica para maximizar el valor que la Inteligencia de Negocios puede aportar a la industria del entretenimiento, permitiendo a las empresas capitalizar de manera más efectiva el creciente interés en el género de terror.

OBJETIVOS

OBJETIVO GENERAL

Realizar una limpieza exhaustiva y estructuración de la base de datos de películas de terror extraída de la API de TMDB, con el fin de optimizar la calidad de los datos para su uso en análisis de Inteligencia de Negocios, permitiendo así la extracción de insights precisos y la toma de decisiones estratégicas informadas en la industria del entretenimiento.

OBJETIVOS ESPECÍFICOS

1. **Identificar y eliminar valores nulos, duplicados o inconsistentes** en el conjunto de datos para asegurar la integridad y fiabilidad de la información.
2. **Corregir y completar datos erróneos o incompletos**, utilizando fuentes externas como Google Colab, para mejorar la precisión del conjunto de datos.
3. **Preparar y estructurar los datos** para facilitar su integración en sistemas de análisis de BI (Inteligencia de negocios), asegurando que los datos estén listos para su uso en modelos predictivos y mecanismos de recomendación mediante dashboards, diagramas u otros elementos gráficos.
4. **Normalizar los campos relevantes** como títulos de películas, fechas de lanzamiento y géneros, garantizando uniformidad y estandarización en el conjunto de datos.
5. **Documentar el proceso de limpieza de datos** de manera detallada, proporcionando un registro claro y replicable de los pasos realizados, las decisiones tomadas y las herramientas utilizadas.

CAPÍTULO 1: COMPRENSIÓN DEL NEGOCIO.

COMPRENSIÓN DEL NEGOCIO

Este proyecto se enmarca en una empresa del sector del entretenimiento, específicamente una plataforma de streaming que se especializa en ofrecer una amplia gama de contenido audiovisual, con un enfoque destacado en el género de terror. La empresa busca posicionarse como el principal destino para los aficionados al cine de terror, proporcionando una experiencia de usuario personalizada y altamente satisfactoria.

La plataforma utiliza datos recolectados de diversas fuentes, incluida la API de TMDB (The Movie Database), una base de datos cinematográfica abierta y colaborativa que proporciona información detallada sobre películas, series y actores. TMDB es conocida por su extensa y actualizada colección de datos, que incluye sinopsis, imágenes, calificaciones y detalles de producción. Al integrar estos datos en su catálogo, la plataforma puede comprender mejor las preferencias de su audiencia y ofrecer recomendaciones más precisas, lo que contribuye a mejorar la retención y satisfacción del cliente.

OBJETIVOS INSTITUCIONALES

1. Ampliar el catálogo de películas de terror con contenido exclusivo y de alta calidad, atrayendo a un público cada vez más amplio y diverso.
2. Incrementar la fidelización del usuario mediante recomendaciones personalizadas basadas en análisis de datos avanzados.
3. Mejorar continuamente la experiencia del usuario en la plataforma, garantizando un acceso fácil y rápido al contenido más relevante.
4. Utilizar la Inteligencia de Negocios para identificar nuevas oportunidades de crecimiento y expansión en mercados emergentes.
5. Fomentar una comunidad activa y comprometida de entusiastas del género de terror, facilitando la interacción y el intercambio de opiniones a través de la plataforma.

INTERESES CON LOS DATOS RECOLECTADOS

Los datos recolectados tienen un valor estratégico significativo para la empresa, ya que permiten:

> **Optimizar la recomendación de películas:** Los datos limpios y estructurados permiten desarrollar algoritmos de recomendación más precisos, que mejoran la satisfacción del usuario y reducen la tasa de abandono.

Identificar tendencias emergentes: Analizar los datos para detectar patrones de consumo y nuevas tendencias en el género de terror, lo cual es crucial para mantenerse al día con la industria.

Personalizar campañas de marketing: Utilizar insights obtenidos de los datos para crear campañas de marketing dirigidas y efectivas, que atraigan a nuevos suscriptores y retengan a los existentes.

Mejorar la toma de decisiones: Apoyar la toma de decisiones informadas en cuanto a adquisiciones de contenido, desarrollo de nuevos productos y expansión a nuevos mercados.

CAPÍTULO 2: ENTENDIMIENTO DE LOS DATOS.

¿DE DONDE PROVIENEN LOS DATOS?

La base de datos utilizada en este proyecto fue encontrada en **Kaggle** y contiene información sobre películas de terror que datan de la década de 1950. Este conjunto de datos fue extraído de **The Movie Database (TMDB)** a través de su API, utilizando la herramienta *httr* de R para realizar las consultas. TMDB es una base de datos abierta y colaborativa que proporciona detalles extensos y actualizados sobre películas, series y actores. El conjunto incluye alrededor de 35,000 registros de películas, lo que permite una amplia exploración y análisis del género de terror a lo largo del tiempo.

DICCIONARIO DE LA BASE DE DATOS

Variable	Definición	Tipo	Importancia
id	ID único de la película	int	Identifica de forma única cada película en la base de datos.
original_title	Título original de la película	char	Permite mantener el nombre original, relevante para la autenticidad y búsqueda.
title	Título de la película	char	Es crucial para la identificación y búsqueda de la película.
original_language	Idioma original de la película	char	Importante para el análisis de la audiencia y la localización del contenido.
overview	Sinopsis/Descripción de la película	char	Ayuda a describir la trama, clave para recomendaciones y análisis de contenido.
tagline	Lema	char	Útil para marketing y captación del interés del público.

release_date	Fecha de lanzamiento	date	Permite análisis de tendencias y patrones de popularidad a lo largo del tiempo.
poster_path	URL de la imagen	char	Importante para la visualización y promoción del contenido.
popularity	Popularidad	num	Indica la aceptación y el interés del público, esencial para recomendaciones.
vote_count	Total de votos	int	Refleja el nivel de interacción del público con la película.
vote_average	Calificación promedio	num	Es clave para determinar la calidad percibida por los usuarios.
budget	Presupuesto de la película	int	Permite analizar la relación entre inversión y éxito comercial.
revenue	Ingresos de la película	int	Indica el éxito financiero de la película, útil para estudios de rentabilidad.
runtime	Duración de la película (minutos)	int	Importante para analizar la estructura narrativa y preferencias del público.
status	Estado de la película	char	Informa si la película está disponible o en desarrollo.
adult	La película está clasificada como contenido para adultos o no	char	Permite filtrar o restringir películas según las restricciones de edad del usuario
genre_names	Lista de géneros	char	Clave para la segmentación y análisis de contenido específico del género de terror.

collection	ID de la colección	num	Permite agrupar películas dentro de una franquicia o serie.
collection_name	Título de la colección	char	Facilita la identificación de sagas o franquicias, relevante para análisis de contenido agrupado.
backdrop_path	URL de la imagen de fondo	char	proporciona la ubicación de una imagen que representa visualmente la película en un formato de fondo

CAPÍTULO 3: PREPARACIÓN DE LOS DATOS.

PROBLEMAS POR RESOLVER

1. ¿Qué subgéneros de terror han sido más populares a lo largo de las décadas?
2. ¿Cuándo es el mejor momento para lanzar una película de terror?
3. ¿Qué duración de una película es la que más popularidad tiene?
4. ¿En qué idiomas producen las películas de terror con mejor recibimiento por los espectadores?
5. ¿Cuál ha sido el año que más películas fueron sacadas?

CAMBIOS PARA REALIZAR SEGÚN EL PROBLEMA

A continuación, se presentará de manera detallada el proceso llevado a cabo para la limpieza de los datos.

1. Primero se pasó la base de datos del formato CSV a un formato XLSX para subirla de forma sencilla a la plataforma de GoogleColab. Allí se importó el Pandas, una biblioteca de Python utilizada para la manipulación y limpieza de datos. Y se procedió a subir el archivo con la base de datos con el siguiente código:

```
df = pd.read_excel('/content/drive/MyDrive/Horror_Movies.xlsx')
print(df.shape)
df.head()
```

Figura 1. Cambios para realizar según el problema. Paso 1

2. Se utiliza un df.info() para mirar las columnas y luego un se extraen los subniveles de cada columna con el siguiente código:

```
cols_cat = [ 'id', 'original_title', 'title', 'original_language', 'overview', 'tagline', 'release_date',
            'poster_path', 'popularity', 'vote_count', 'vote_average', 'budget', 'revenue', 'runtime',
             'status', 'adult', 'backdrop_path', 'genre_names', 'collection', 'collection_name']

for col in cols_cat:
  print(f'Columna {col}:{df[col].nunique()} subniveles')
```

Figura 2. Cambios para realizar según el problema. paso 2

En este se observa que los subniveles en la columna ‘adult’ es 1, por lo dato, al ser un único valor, se tiene en cuenta para borrar la columna al momento de actualizar el DataFrame.

3. Se hace un análisis que dé hay muchos valores en la columna ‘status’, lo cual me indica que no todas las películas que aparecen han sido lanzadas, por lo tanto, se procede a eliminar los datos de todas las películas que no han sido lanzadas, y esta columna también se tiene en cuenta para ser eliminada, ya que quedará con un solo registro. Para eliminar los datos que no tengan el estado de ‘Released’, se aplica el siguiente código:

```
df = df[df['status'] == 'Released']
print(df.shape)
```

Figura 3. Cambios para realizar según el problema. paso 3

4. Hacemos un análisis de las columnas que voy a eliminar con el conocimiento obtenido en clase y la información que quiero obtener para resolver los problemas, aquí el análisis de las columnas eliminadas:
    - Id: Se eliminará ya que, al borrar algunas filas en el futuro, los datos perderán su secuencia, por lo que se piensa agregar una columna de id al finalizar la limpieza.
    - Original_title: En esta columna se encuentran algunos nombres de películas en lenguajes no legibles, teniendo en cuenta que ya existe una columna con el nombre traducido al inglés, se vuelve innecesaria.
    - Overview: Llamado sinopsis en español, ayudar a describir la trama de las películas, y es necesaria para el espectador, pero no para la inteligencia negocios.
    - Tagline: Llamado lema en español ayuda a llamar la atención del público, pero de nuevo, no es necesaria para la inteligencia de negocios.
    - Poster_path: URL de la imagen del poster, no es ningún dato relevante para el BI.
    - Status: Como se había comentado, se aplicó una limpieza y solo quedó un dato, por lo que se vuelve innecesaria.
    - Adult: Al analizar los subniveles se vio que esta tenía un solo dato, por lo que se vuelve inútil para la inteligencia de negocios.
    - Backdrop_path: Llamado URL de la imagen de fondo, el cual no es útil ya que es una dirección URL.

- Collection: Código único de las colecciones, no serán utilizadas por lo que se borra este código. Además, que muchas películas no pertenecen a una colección en específico, por lo que no se utilizará.
- Collection_name: Contiene el nombre de la colección y no se utilizará tampoco por la misma razón que la columna 'collection'

Excluyendo estas 10 columnas, nos quedan otras 10 para armar el DataFrame, de la siguiente forma:

```
df = df[[ 'title', 'original_language', 'release_date', 'popularity',
          'vote_count', 'vote_average','budget', 'revenue','runtime',
           'genre_names']]
df.head()
```

Figura 4. Cambios para realizar según el problema. paso 4

5. Luego se proceden a borrar las filas con datos faltantes y borrar los registros duplicados:

```
#Eliminar filas con datos faltantes
df.dropna(inplace=True)
print(df.shape)
```

```
(32322, 10)
```

```
#Borrar los registros duplicados
df.drop_duplicates(inplace=True)
print(df.shape)
```

```
(32322, 10)
```

Figura 5. Cambios para realizar según el problema. paso 5

6. Para el siguiente paso de la limpieza, se empezarán a modificar o eliminar los datos que tienen 0 en alguna posición de la fila. Ya sea promediando los datos en base a la columna o eliminando la fila directamente. Para ver los datos en 0 de cada columna, se utilizarán las siguientes líneas, las cuales mostrarán los nombres de cada columna

```
zero_counts_columns = (df == 0).sum()
print(zero_counts_columns)
```

Figura 6. Cambios para realizar según el problema. paso 6

Se eliminarán las filas donde ‘vote_count’ es 0, ya que no nos interesan las películas con ningún voto.

```
# Eliminar las filas donde la columna 'vote_count' sea 0
df = df[df['vote_count'] != 0]
```

Figura 7. Cambios para realizar según el problema. paso 6

Se observa que la cantidad de 0 en la columna ‘vote_average’ ha disminuido ya que gran parte de estas eran las mismas que no tenía ningún voto. No solo estas columnas disminuyeron de cantidad de 0, sino algunas otras también lo hicieron.

Se observa que quedaron solo 39 datos en la columna ‘vote_average’, por lo que teniendo en cuenta que aún se tienen 20917 registros, se ha decidió promediarlo, ya que son datos insignificantes para la dimensión de datos con los que se trabaja.
Para promediarlos, se excluyen los números que son diferentes a 0, se promedian, y se guardan en una variable, luego se reemplaza cada 0 con el valor del promedio, de la siguiente forma:

```
# Calcular el promedio de los valores de 'vote_average' que no son 0
mean_vote_average = df[df['vote_average'] != 0]['vote_average'].mean()
# Reemplazar los valores 0 en la columna 'vote_average' con el promedio calculado
df['vote_average'] = df['vote_average'].replace(0, mean_vote_average)
```

Figura 8. Cambios para realizar según el problema. paso 6

Para la columna de ‘runtime’, la cantidad de 0 que quedaron fue de 884, por lo que al solo representar el 4.23% de los datos totales, se decidió promediar los otros datos diferentes de 0 y reemplazar cada dato en 0, al igual que en el paso anterior. La única diferencia en este caso es que los valores fueron pasados a enteros para así tener minutos completos y no a mitades. A continuación, el código usado:

```
# Calcular el promedio de los valores de 'runtime' que no son 0
mean_runtime = df[df['runtime'] != 0]['runtime'].mean()
# Convertir a entero
mean_runtime = int(mean_runtime)
# Reemplazar los valores 0 en la columna 'runtime' con el promedio calculado
df['runtime'] = df['runtime'].replace(0, mean_runtime)
```

Figura 9. Cambios para realizar según el problema. paso 6

A diferencia del caso anterior, en la columna ‘budget’ que es el presupuesto, la cantidad de filas con datos en 0 representa el 82.02%, mientras que en la columna ‘revenue’ que son las ganancias de la película, la cantidad de filas con datos en 0 representa el 92.82%. teniendo en cuenta que es la gran cantidad de datos, promediarlas solo me daría información poco confiable, por lo que se ha decidido eliminarlas del análisis.

Para eliminarlas, se hace los mismo que en el caso anterior, asignando a mi DataFrame nuevo, mi DataFrame original, pero omitiendo las columnas no deseadas, de la siguiente forma:

```
df = df[[ 'title', 'original_language', 'release_date', 'popularity',
          'vote_count', 'vote_average', 'runtime', 'genre_names']]
```

Figura 10. Cambios para realizar según el problema. paso 6

Para finalizar, vamos a verificar que todos los datos sean los mismos en todas las filas, se contarán los 0 de cada columna, se va a mirarán cuantos datos quedaron en total y se mirará el DataFrame como queda al finalizar, con el siguiente código:

```
#Confirmar que todos los datos poseen la misma cantidad
print(df.shape)
print('-------------------------------------')
# Contar cuántos ceros hay en cada columna
zero_counts_columns = (df == 0).sum()
print(zero_counts_columns)
print('-------------------------------------')
#Cantidad de datos y columnas
df.info()
df.tail()
```

Figura 11. Cambios para realizar según el problema. paso 6

7. A este DataFrame se le va a agregar un índice que se llamará id y se posicionará como primera columna. Esté funcionará como el dato único que diferencia a cada una de las películas en la base de datos ya que no se repite. Se agrega de la siguiente forma:

```
# Reiniciar el índice y eliminar el anterior
df.reset_index(drop=True, inplace=True)
# Crear la nueva columna 'id' que comienza desde 1
df['id'] = range(1, len(df) + 1)
# Reordenar las columnas para que 'id' sea la primera
cols = ['id'] + [col for col in df.columns if col != 'id']
df = df[cols]
# Verificar la tabla con el nuevo id como primera columna
df.tail()
```

Figura 12. Cambios para realizar según el problema. paso 7

8. En este paso se empieza a formar las dimensiones y la tabla de los hechos para realizar el modelo analítico de datos.
   Se empieza con la dimensión para los lenguajes, donde se extraen los lenguajes únicos y se crea un DataFrame con estos. Luego se hace un mapeo de cada uno de los datos para conectarlos con una nueva columna que se utilizará en la tabla de los hechos para ser conectada con en la nueva dimensión que hemos creado.

```
# Extraer los lenguajes únicos
unique_languages = df['original_language'].unique()
# Crear un DataFrame con los lenguajes únicos
dim_language = pd.DataFrame({'language_id': range(1, len(unique_languages) + 1),
                             'original_language': unique_languages})
# Mostrar la tabla de lenguajes únicos
print(dim_language)
print('-----------------------------------')
# Crear un diccionario para mapear cada lenguaje con su ID
language_map = dict(zip(dim_language['original_language'], dim_language['language_id']))
# Crear una nueva columna en tu DataFrame original con el ID del lenguaje
df['language_id'] = df['original_language'].map(language_map)
# Mostrar el DataFrame original con la nueva columna de lenguaje_id
print(df[['id', 'original_language', 'language_id']])
```

Figura 13. Cambios para realizar según el problema. paso 8

Se hace el mismo proceso con la Dim_genre. Se extraen los géneros únicos y se crea un Dataframe con estos. Luego se hace un mapeo de cada uno de los datos para conectarlos con una nueva columna.

```
# Obtener los géneros únicos
unique_genres = df['genre_names'].unique()
# Crear la tabla de dimensión de géneros
dim_genre = pd.DataFrame({'genre_id': range(1, len(unique_genres) + 1),
                          'genre_names': unique_genres})
# Mostrar la tabla de géneros únicos
print(dim_genre)
print('----------------------------')
# Crear un diccionario para mapear cada género con su genre_id
genre_map = dict(zip(dim_genre['genre_names'], dim_genre['genre_id']))

# Crear una nueva columna en tu DataFrame original con el ID del lenguaje
df['genre_id'] = df['genre_names'].map(genre_map)
# Mostrar el DataFrame original con la nueva columna de lenguaje_id
print(df[['id', 'genre_names', 'genre_id']])
```

Figura 14. Cambios para realizar según el problema. paso 8

Luego se agregó un nuevo dataframe en el que se utilizó la fecha como clave primaria, y luego se procedió a dividir la fecha en varias partes para hacer la tabla calendario. Luego esta se exportó de igual forma como la vez anterior en formato csv

```
df['release_date'] = pd.to_datetime(df['release_date'])  # Convertir la columna 'release_date' a formato datetime
# Extraer las fechas únicas del DataFrame original
unique_dates = df['release_date'].unique()
# Crear un DataFrame con las fechas únicas y nombrarlo dim_date
dim_date = pd.DataFrame({'date': unique_dates})
# Agregar columnas con detalles adicionales sobre las fechas
dim_date['date'] = pd.to_datetime(dim_date['date'])
dim_date['year'] = dim_date['date'].dt.year
dim_date['month'] = dim_date['date'].dt.month
dim_date['month_name'] = dim_date['date'].dt.strftime('%B')
dim_date['day'] = dim_date['date'].dt.day
dim_date['day_name'] = dim_date['date'].dt.strftime('%A')
dim_date['week_of_year'] = dim_date['date'].dt.isocalendar().week
dim_date['quarter'] = dim_date['date'].dt.quarter

# Mostrar la tabla dim_date
print(dim_date)
print('-----------------------------------')
# Crear un diccionario para mapear cada fecha con su propia fecha (clave primaria)
date_map = dict(zip(dim_date['date'], dim_date['date']))
# Crear una nueva columna en tu DataFrame original con la clave de la fecha
df['date_key'] = df['release_date'].map(date_map)
# Mostrar el DataFrame original con la nueva columna de date_key
print(df[['id', 'release_date', 'date_key']])
dim_date
```

Figura 15. Cambios para realizar según el problema. paso 8

Por último, se creará la tabla para los hechos que será llamada ‘facts_movies’, en la cual se ingresaran las id para conectar a las otras tablas, así como los valores totales o numéricos para realizar el análisis en futuros modelos de análisis:

```
# Crear la tabla de hechos
fact_movies = df[['id', 'title', 'popularity', 'vote_count',
                  'vote_average', 'runtime', 'language_id',
                  'release_date']]
fact_movies
```

Figura 16. Cambios para realizar según el problema. paso 8

9. Se exportan cada una de las dimensiones y la tabla de hechos para un Excel donde será cargadas para posteriormente ser unidas.

```
dim_language.to_csv('dim_language.csv', index=False)
dim_date.to_csv('dim_date.csv', index=False)
dim_genre.to_csv('dim_genre.csv', index=False)
fact_movies.to_csv('facts_movies.csv', index=False)
```

Figura 17. Cambios para realizar según el problema. paso 9

CAPÍTULO 4: MODELADO.

MODELADO ANALÍTICO DE DATOS

En este modelado analítico, se hizo en forma de estrella donde la tabla de hechos conecta con todas las dimensiones: con la dimensión lenguaje mediante el 'lenguaje_id', con la dimensión del date mediante 'Date', y con la dimensión género que se conecta con 'genre_id'.

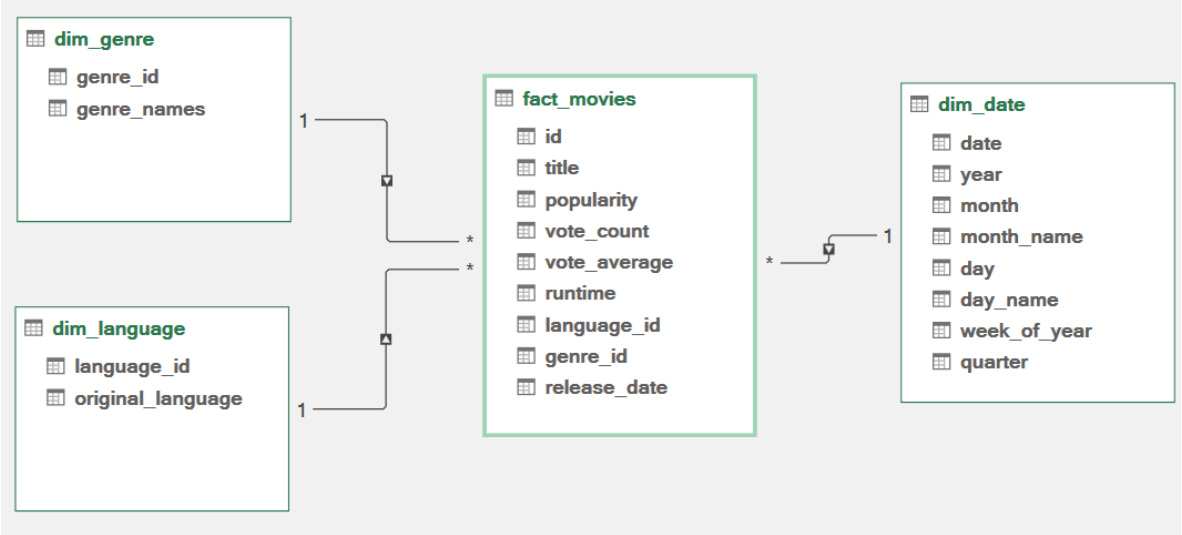

Figura 18. Modelado analítico de datos

CAPÍTULO 5: EVALUACIÓN.

Con el diagrama anterior en Power Pivot se hace una base de datos adaptándola las columnas a SQL. Se usará la fecha como String para la clave primaria, así evitar errores de conexión

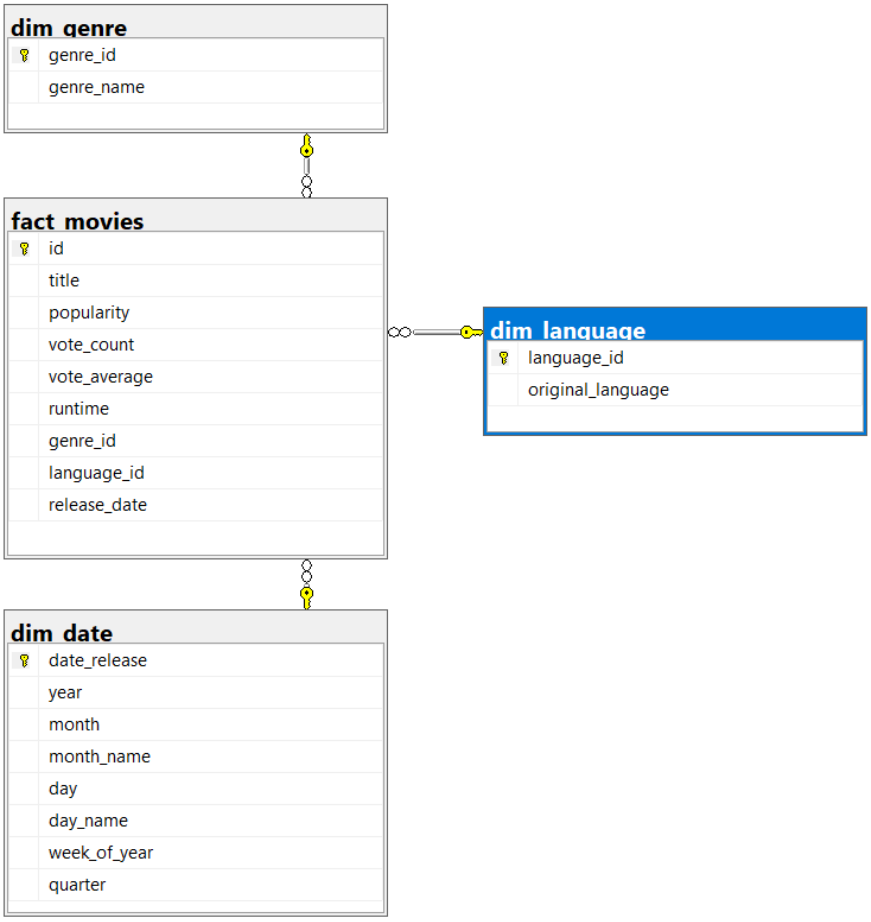

Figura 19. Diagrama de la base de datos

Aquí se observa el flujo de control del ETL funcionando correctamente, y a continuación se verá el proceso de cada elemento.

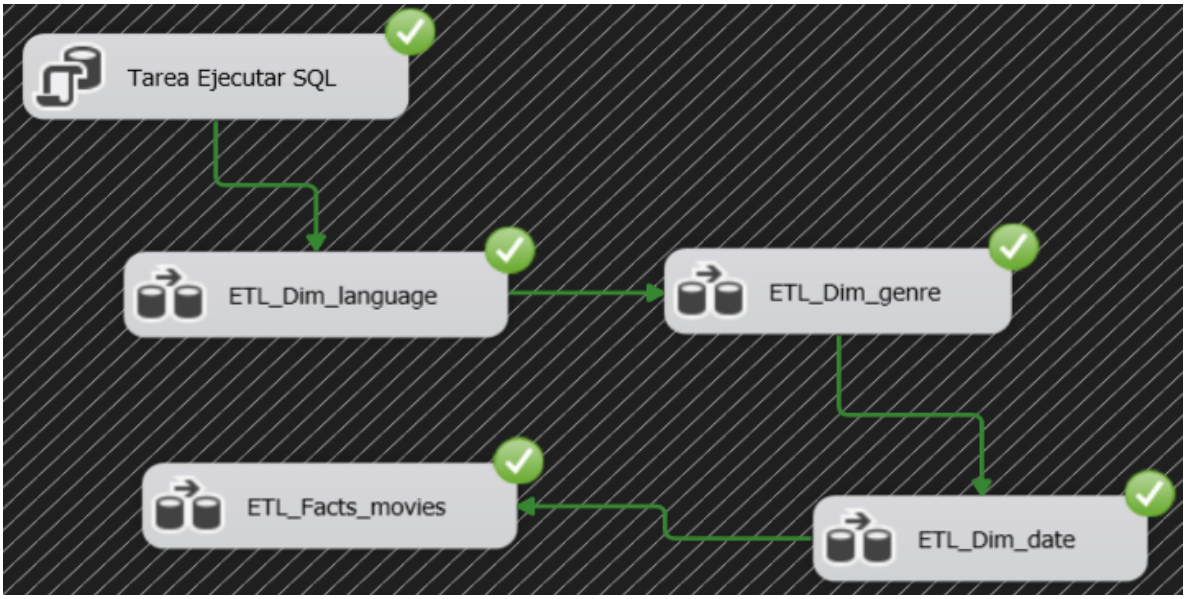

Figura 20. Flujo de control

Para empezar, se borran los datos existentes en la base de datos con un Query dentro de las configuraciones

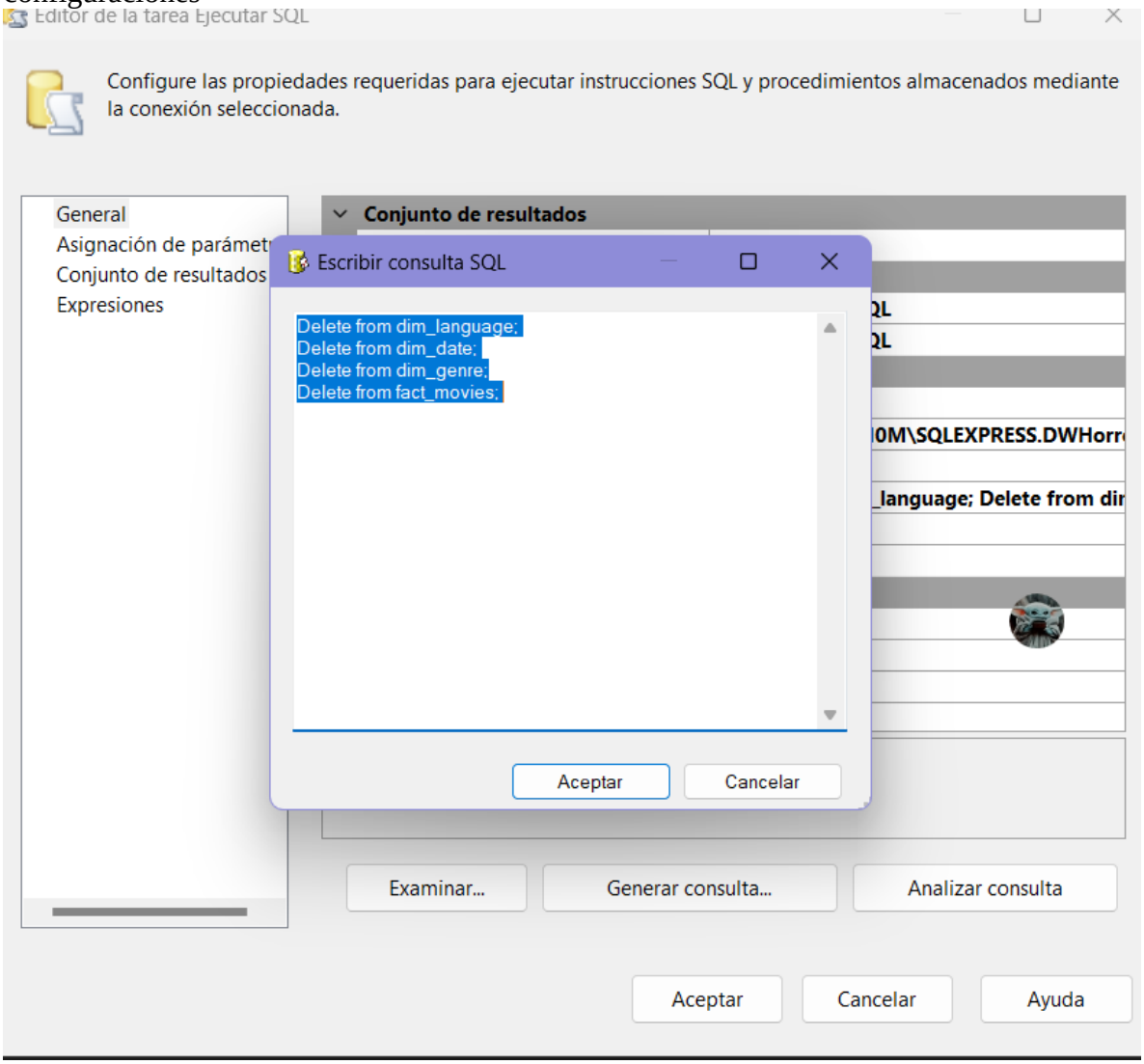

Figura 21. Tarea a ejecutar SQL. Query

Luego, pasamos a ETL_Dim_language y allí veremos el flujo de datos y cada una de las configuraciones de los elementos de este

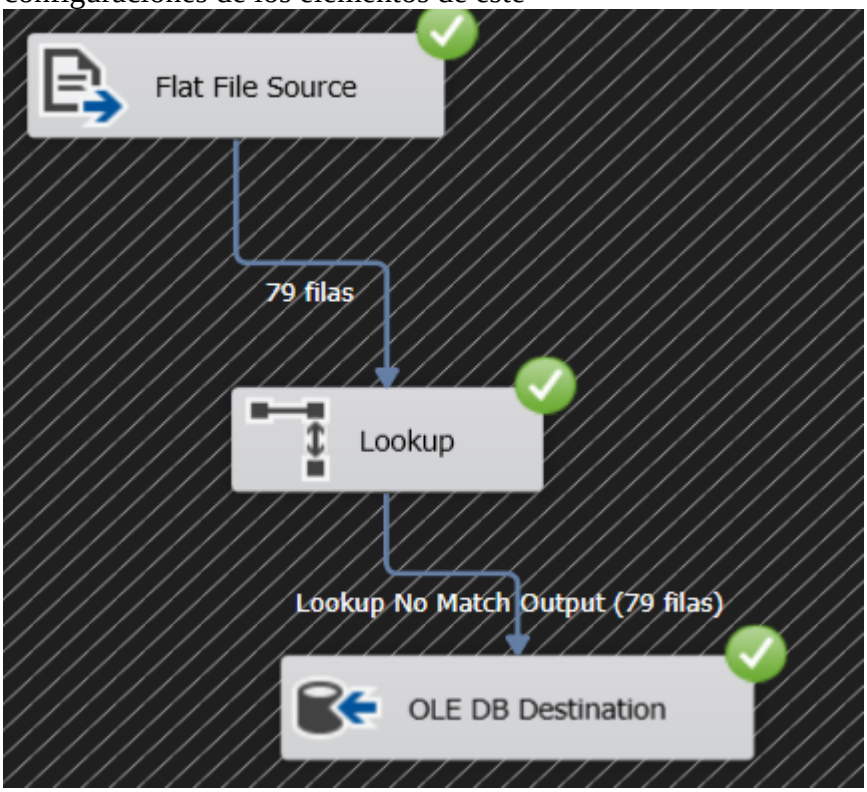

Figura 22. ETL_Dim_language

En la siguiente imagen, se crea una conexión con un archivo plano para ser asignada al Origen del archivo plano. En este se elige la codificación, se mira que las columnas estén como se debe, en caso de que no, se modifica el determinador de filas. Aquí también se modifica el tipo de las variables como aparecen en la base de datos en la pestaña de Opciones avanzadas.

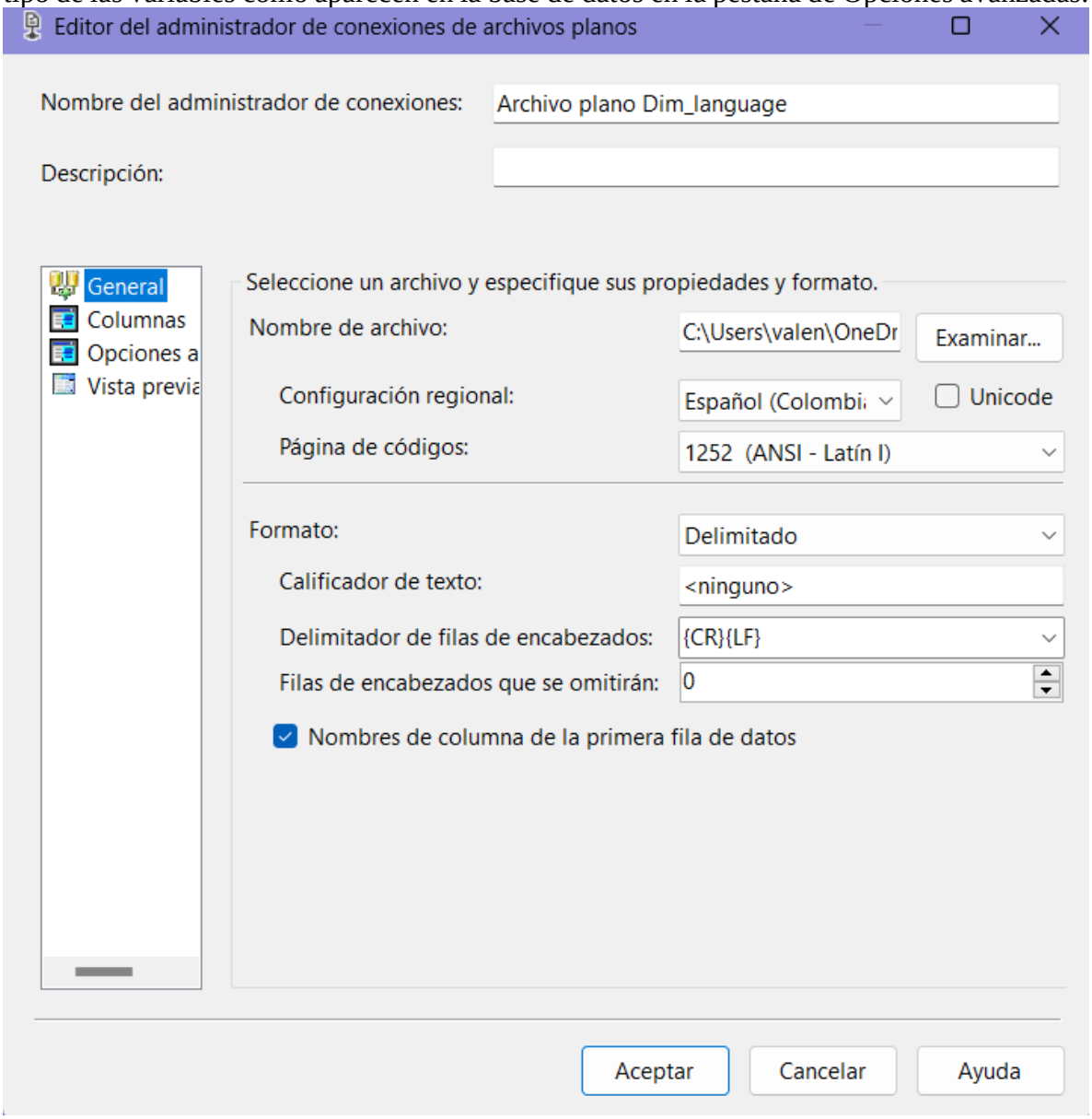

Figura 23. Administrador de archivos planos Dim_language. General

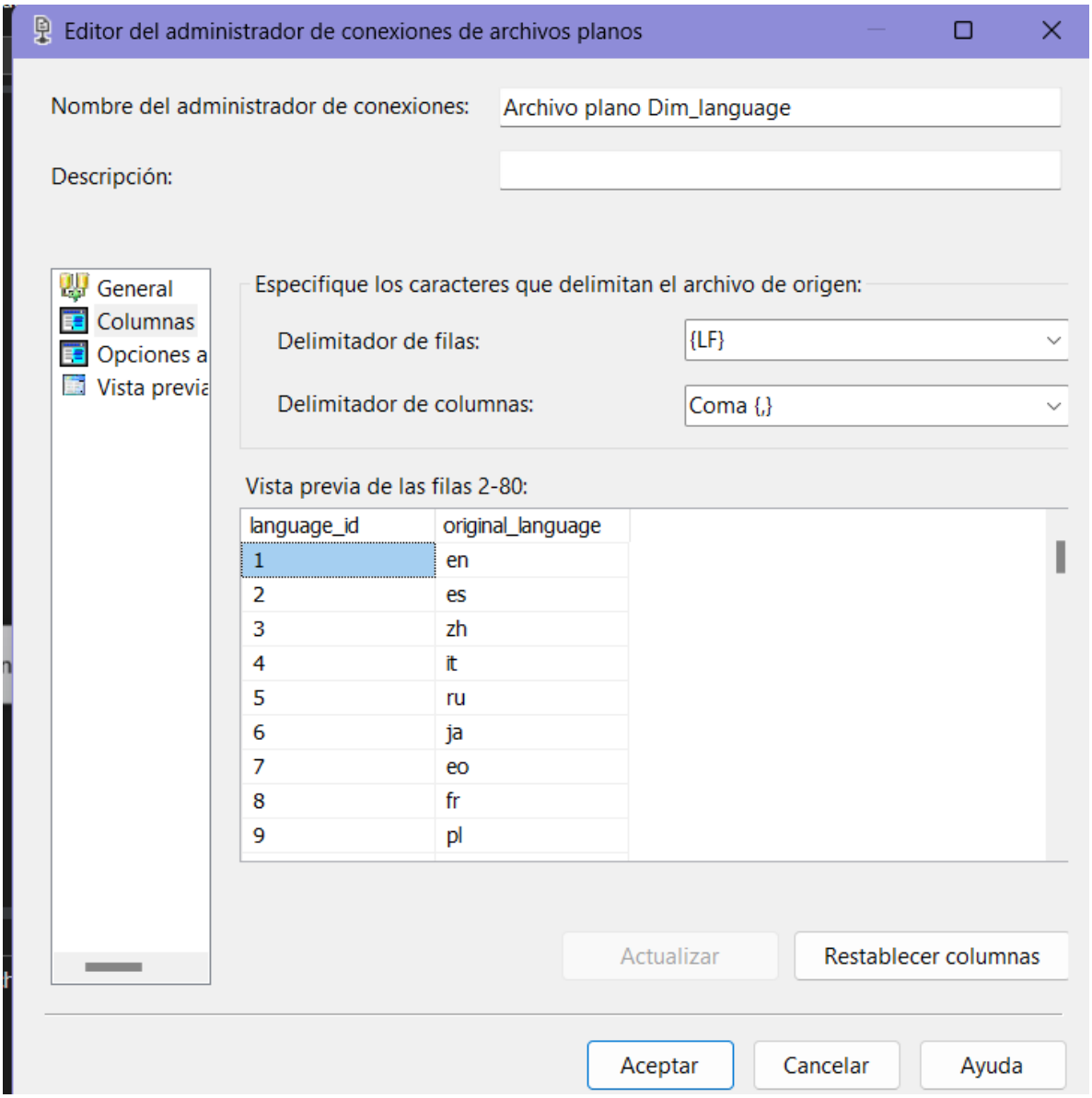

Figura 24. Administrador de archivos planos Dim_language. Columnas

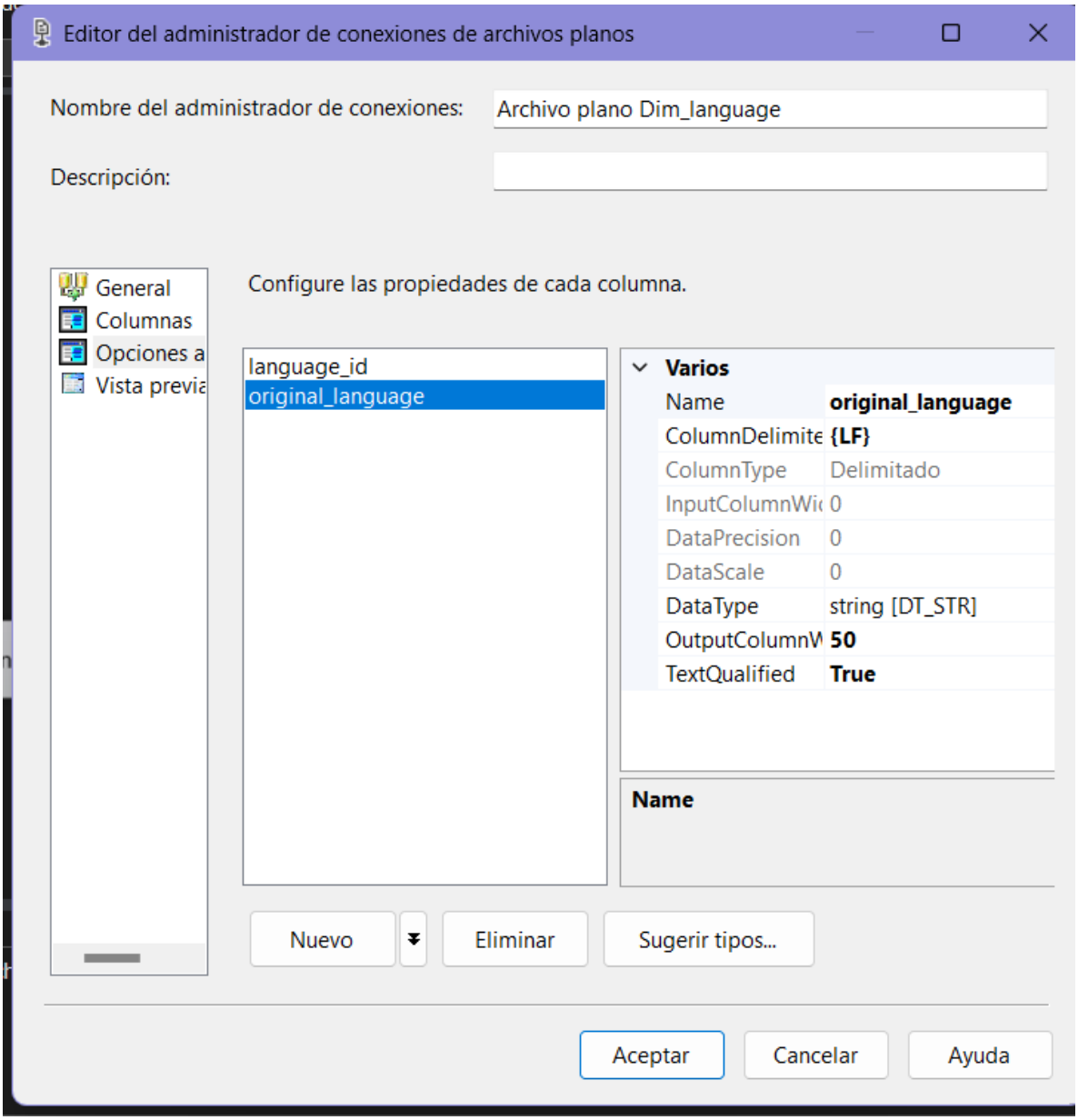

Figura 25. Administrador de archivos planos Dim_language. Opciones avanzadas

Luego se le asigna la conexión al archivo plano, luego en el Lookup ponemos una asignación para redirigir las filas sin coincidencias, conectándolo con la tabla que le corresponde. Por último, se conecta el destino a mi base de datos.

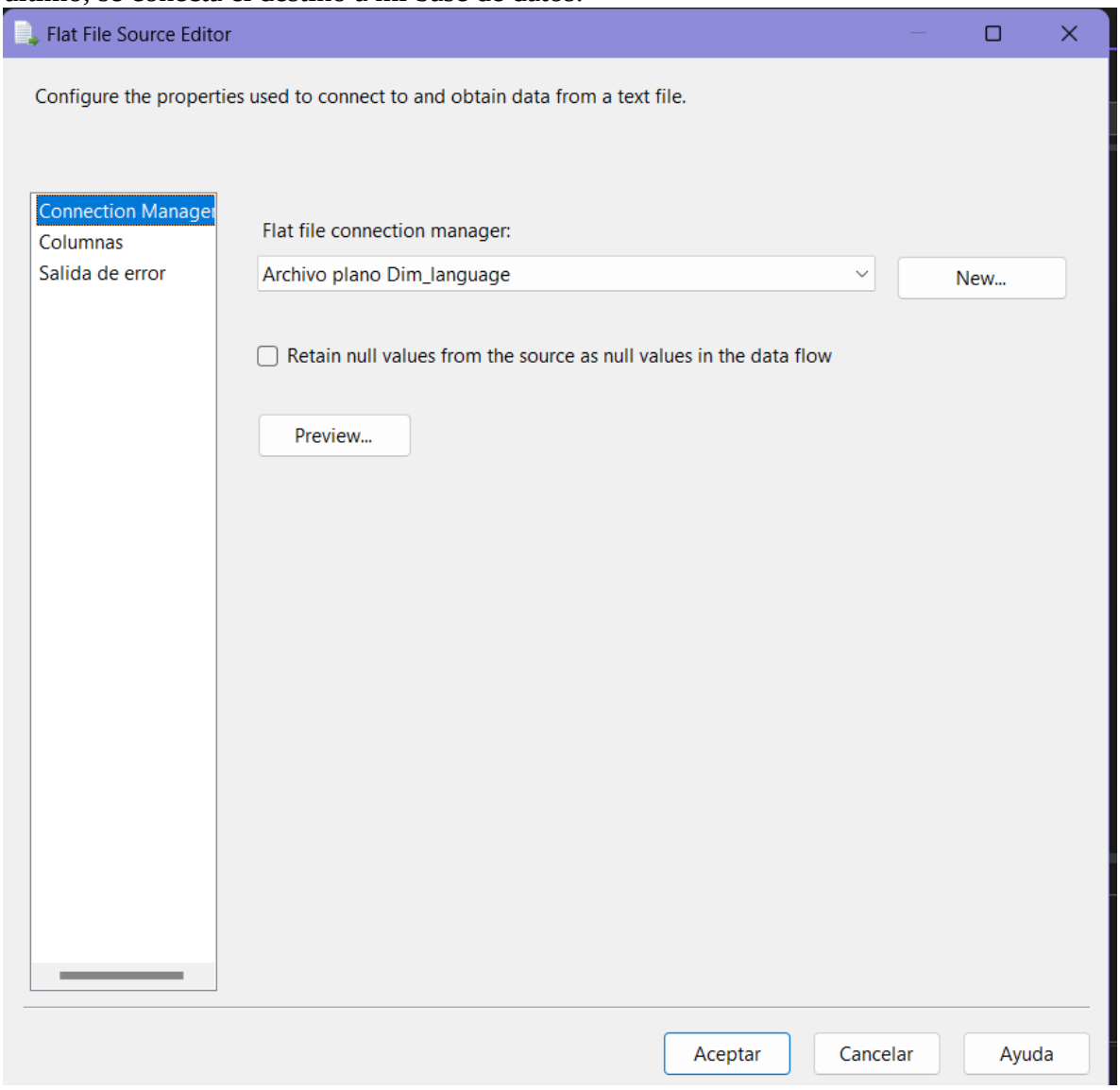

Figura 26. Origen de archivo plano

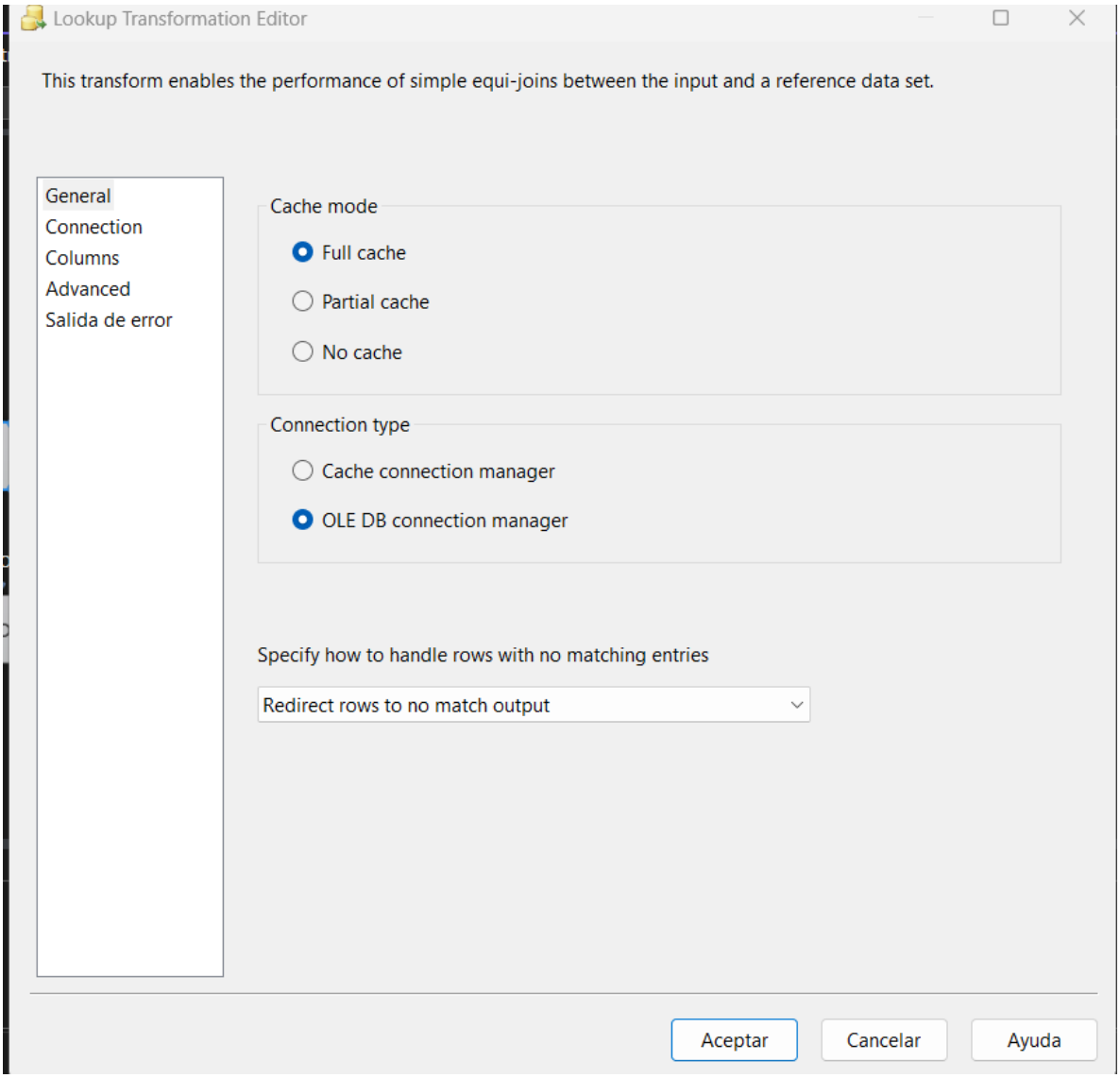

Figura 27. Lookup trasnformation. ETL_Dim_language

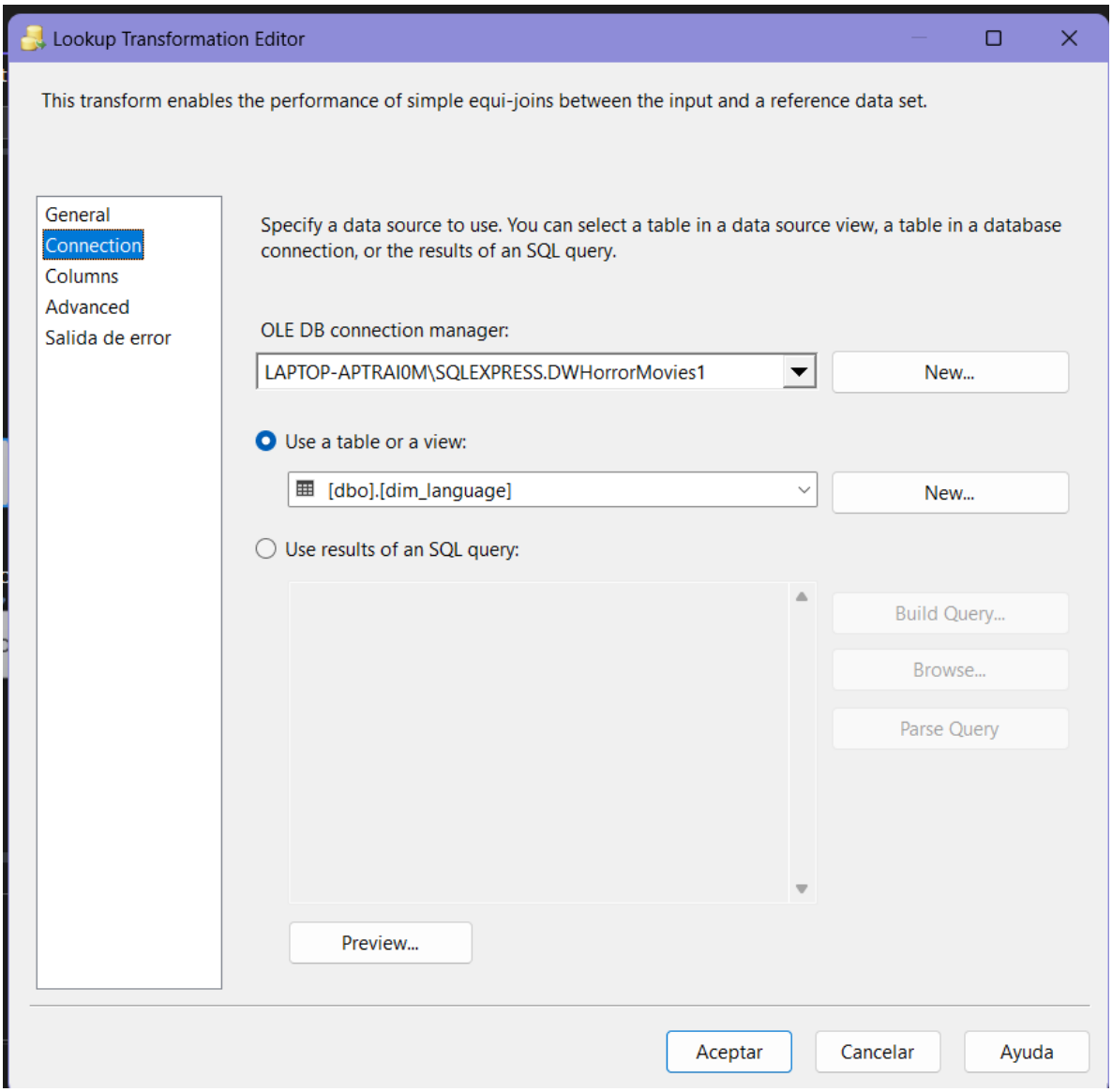

Figura 28. Lookup trasnformation. ETL_Dim_language

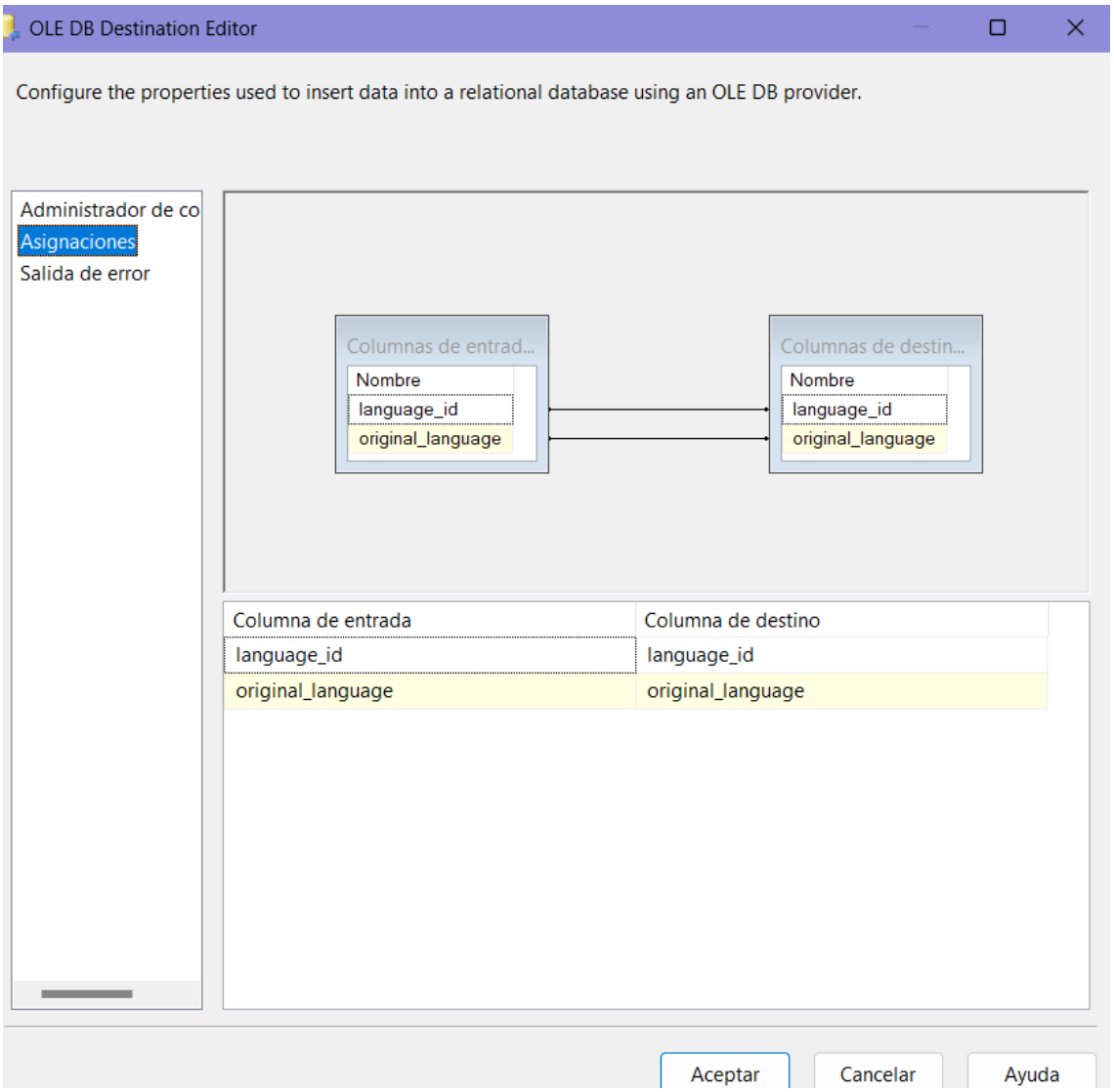

Figura 29. Destino OLE DB. Asignaciones

Se hace el mismo proceso anterior para Dim_genre y Dim_date.

Dim_genre:

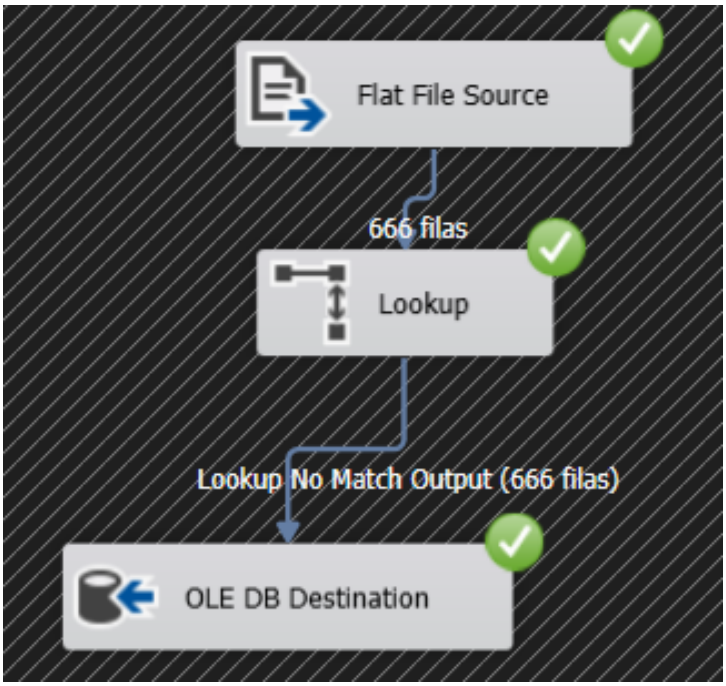

Figura 30. ELT_Dim_genre

Dim_date:

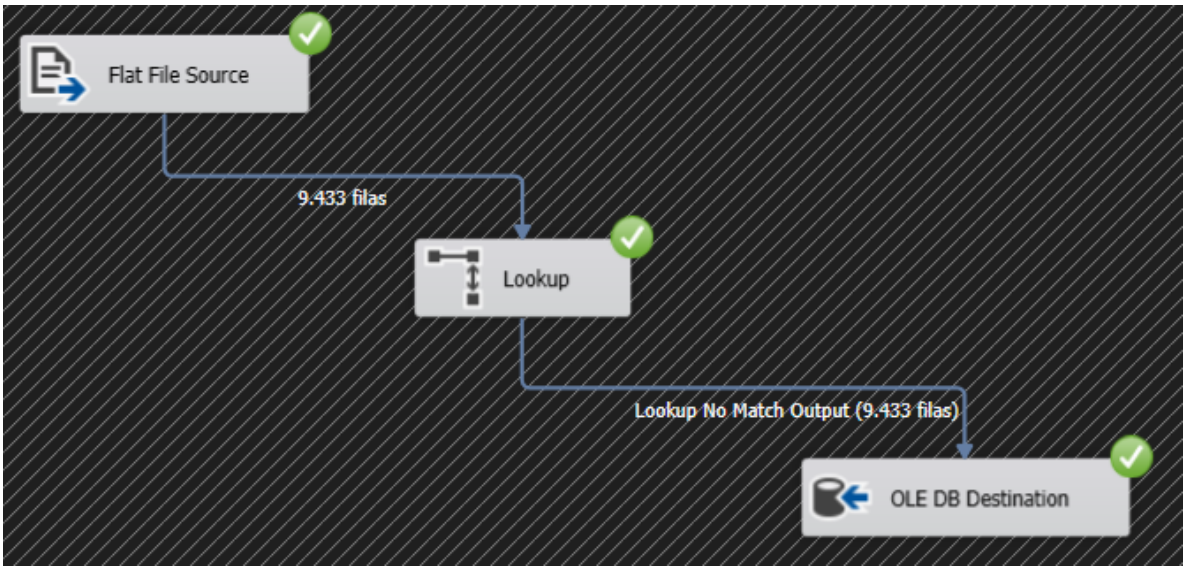

Figura 31. ETL_Dim_date

Para finalizar, se hará el ETL para la tabla facts_movies en el que se agarrará los datos de cada una de las otras tablas y se hará una conexión entre ellos, por último, se llevará a su respectiva tabla de en la base de datos. Hay algunas filas que no son reconocidas por la codificación 1252, por lo tanto, se optó por omitirlos ya que son la gran minoría.

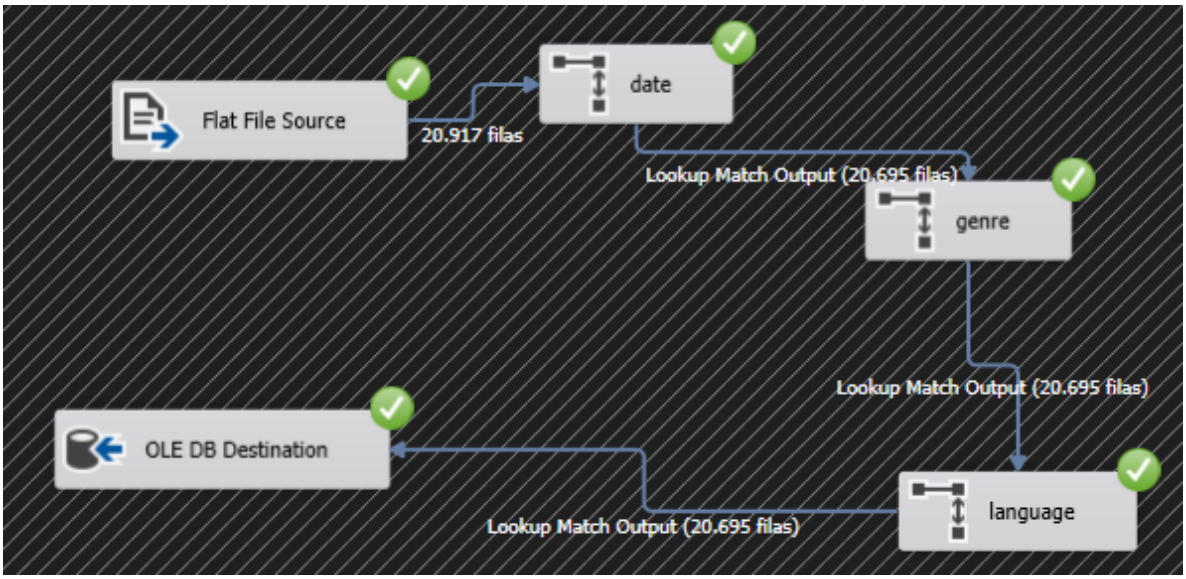

Figura 32. ETL_facts_movies

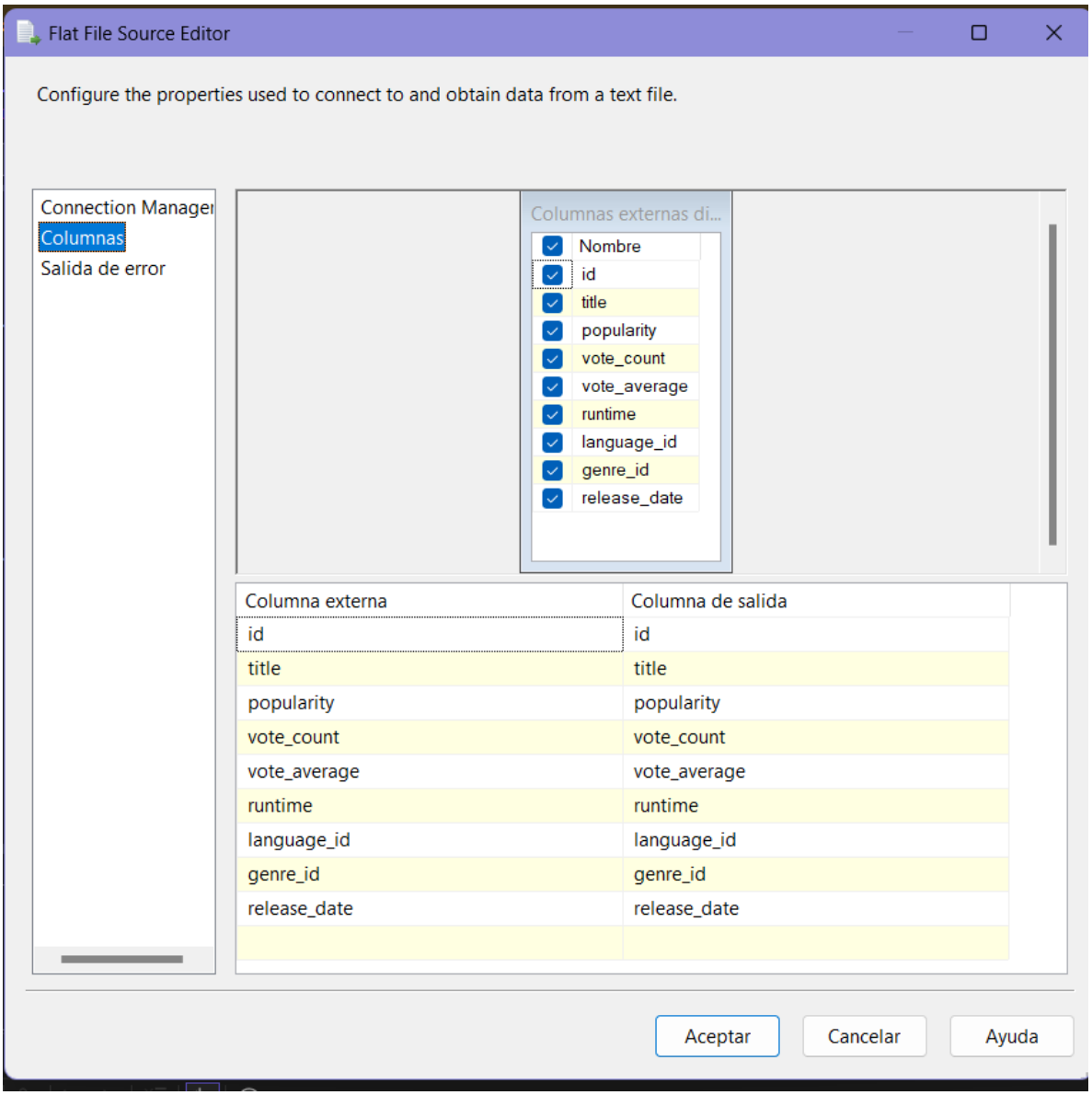

Figura 33. Origen del archivo plano facts_movies. Columnas

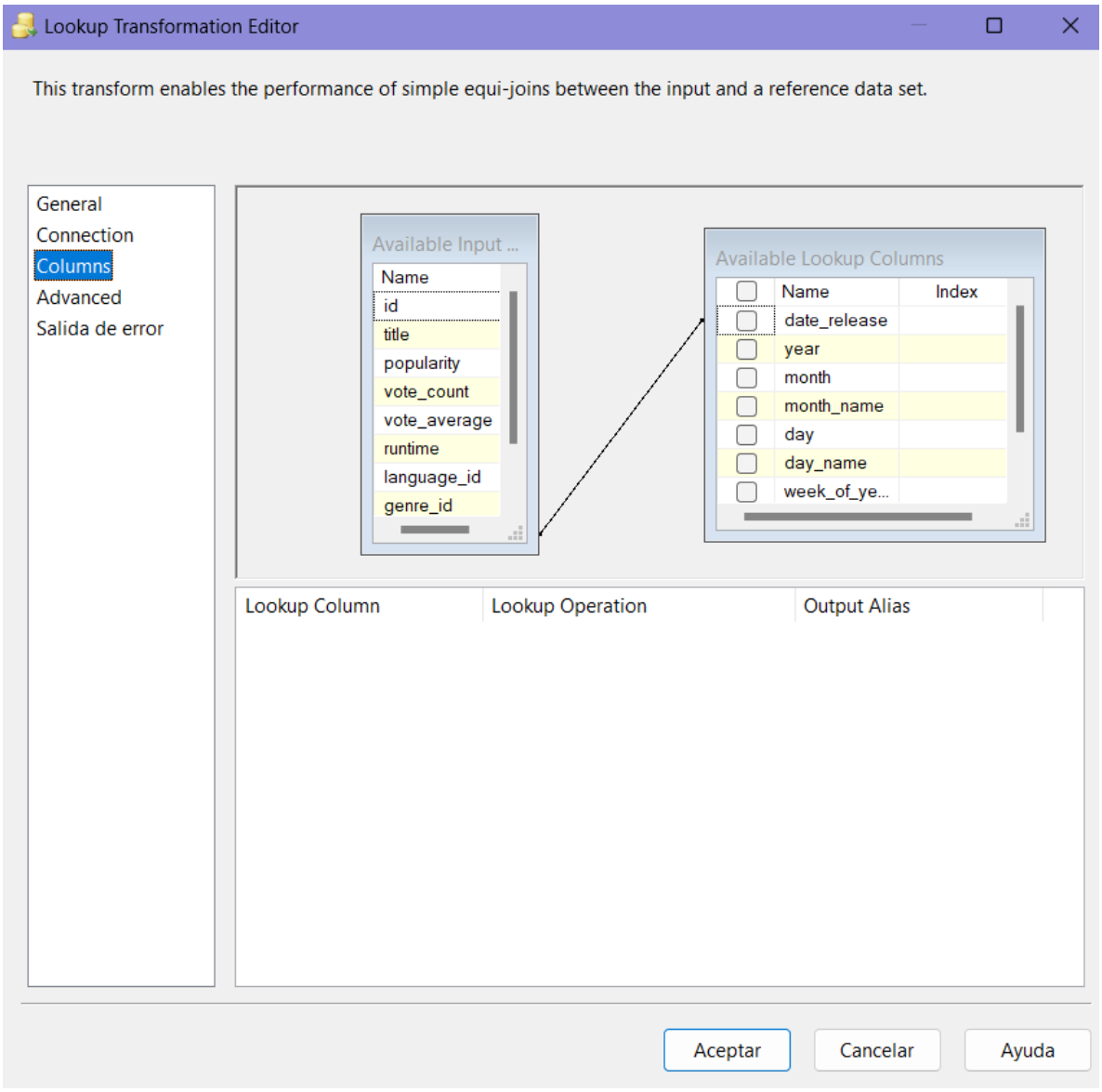

Figura 34. Lookup transformation Dim_date

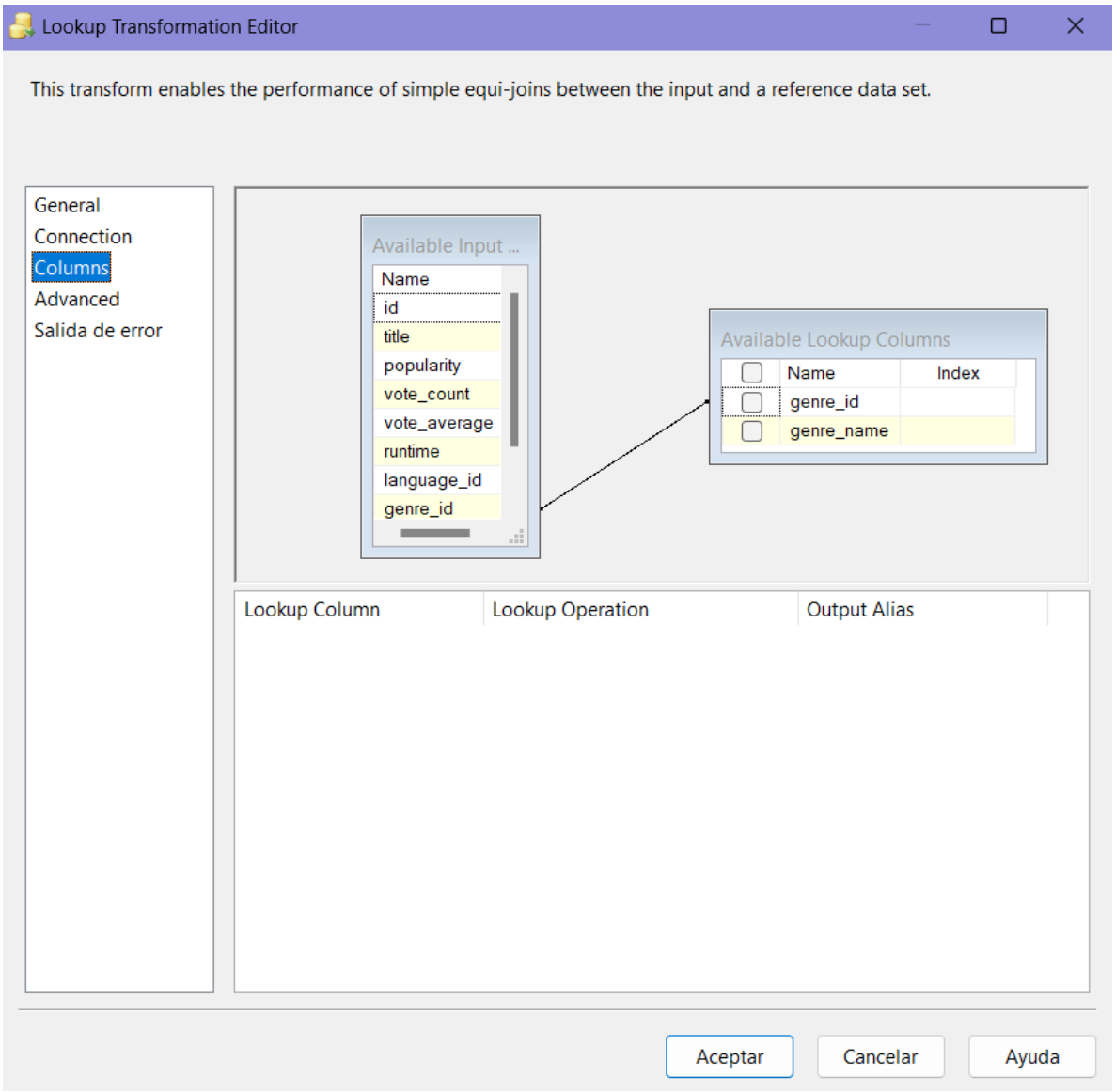

Figura 35. Lookup transformation Dim_genre

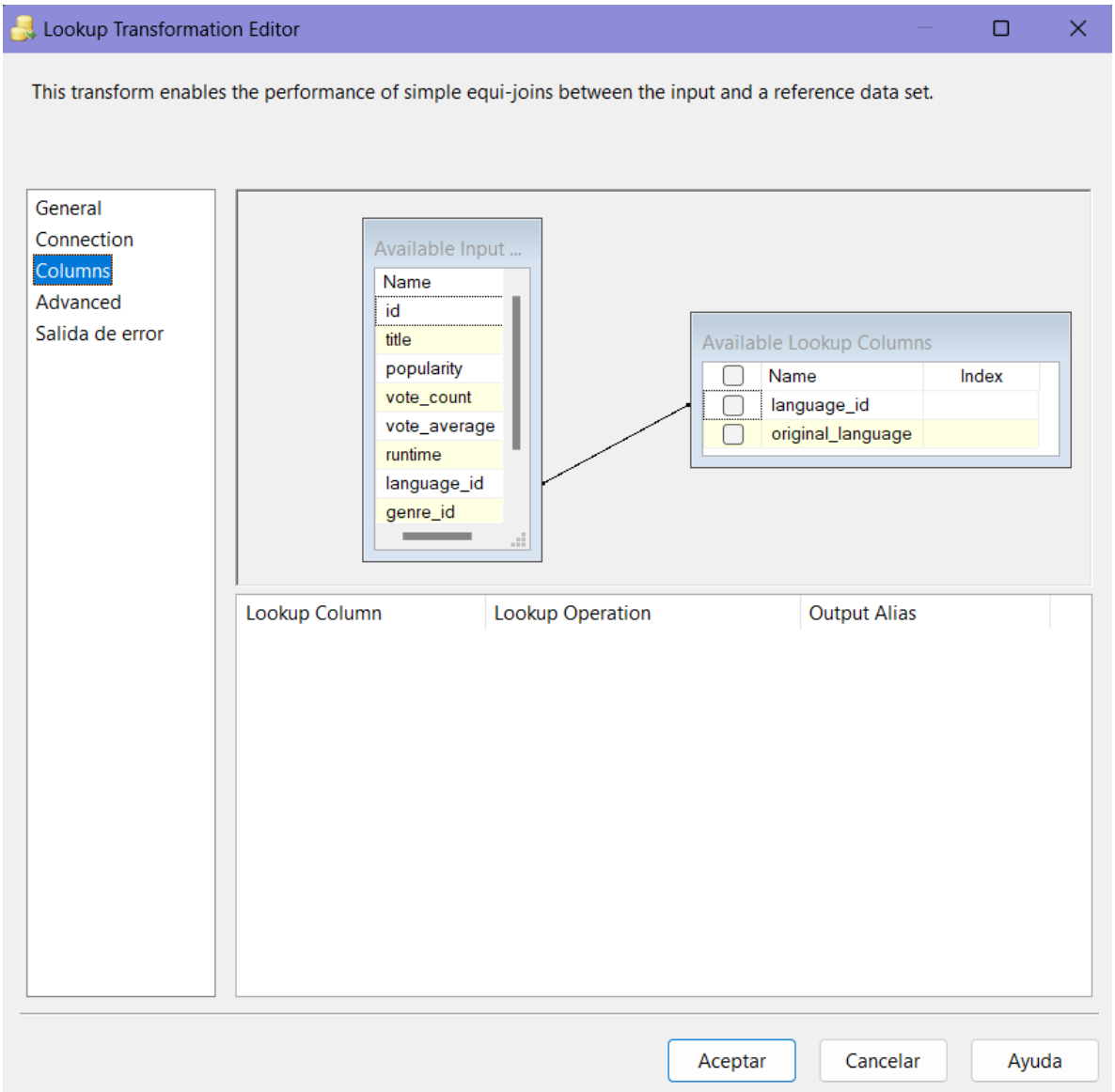

Figura 36. Lookup transformation Dim_language

Por último, se agrega un destino a base de datos en el cuál se hará la conexión de todas las columnas del archivo plano con las columnas de la tabla en la base de datos.

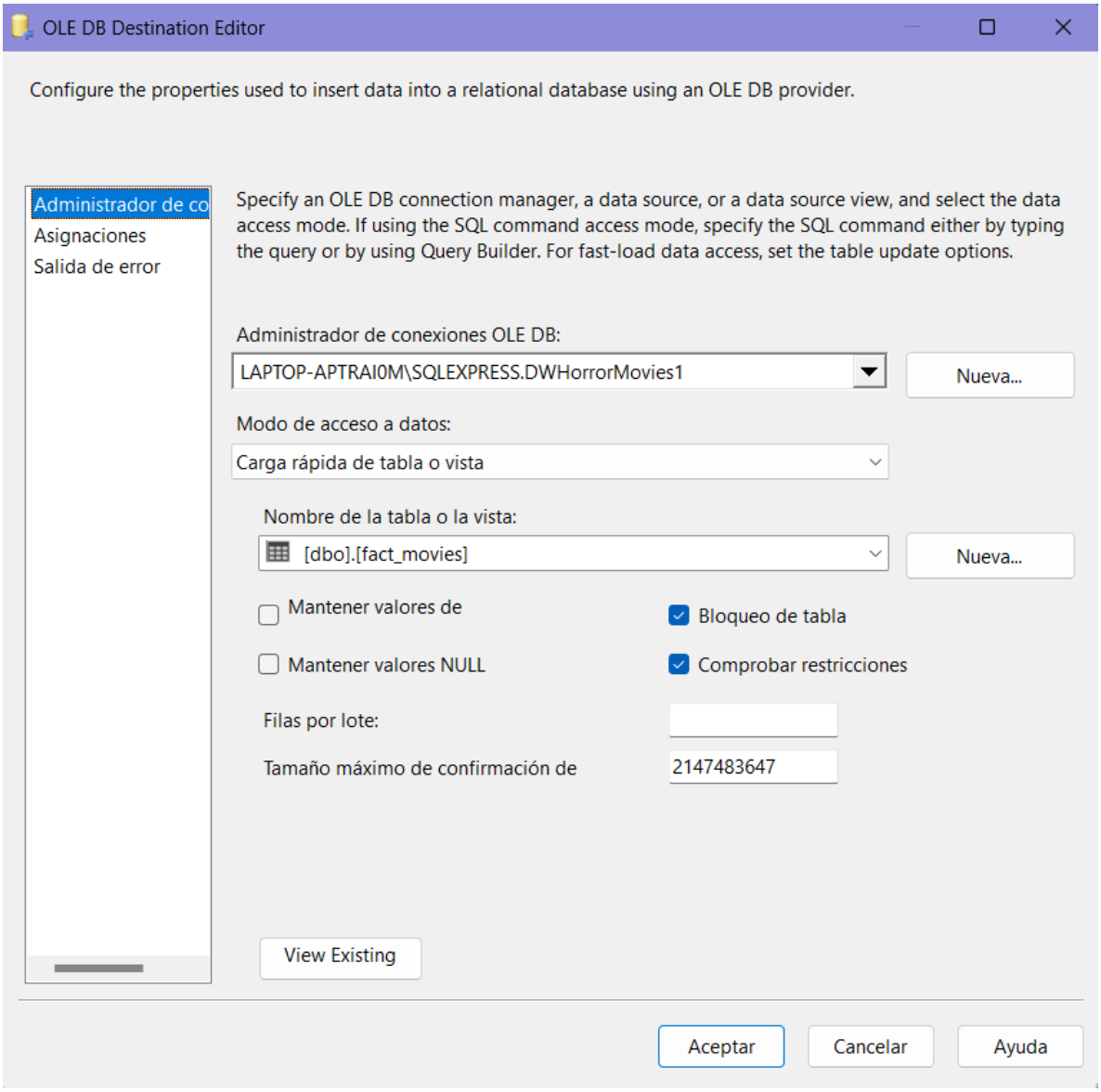

Figura 37. Destino OLE DB para facts_movies

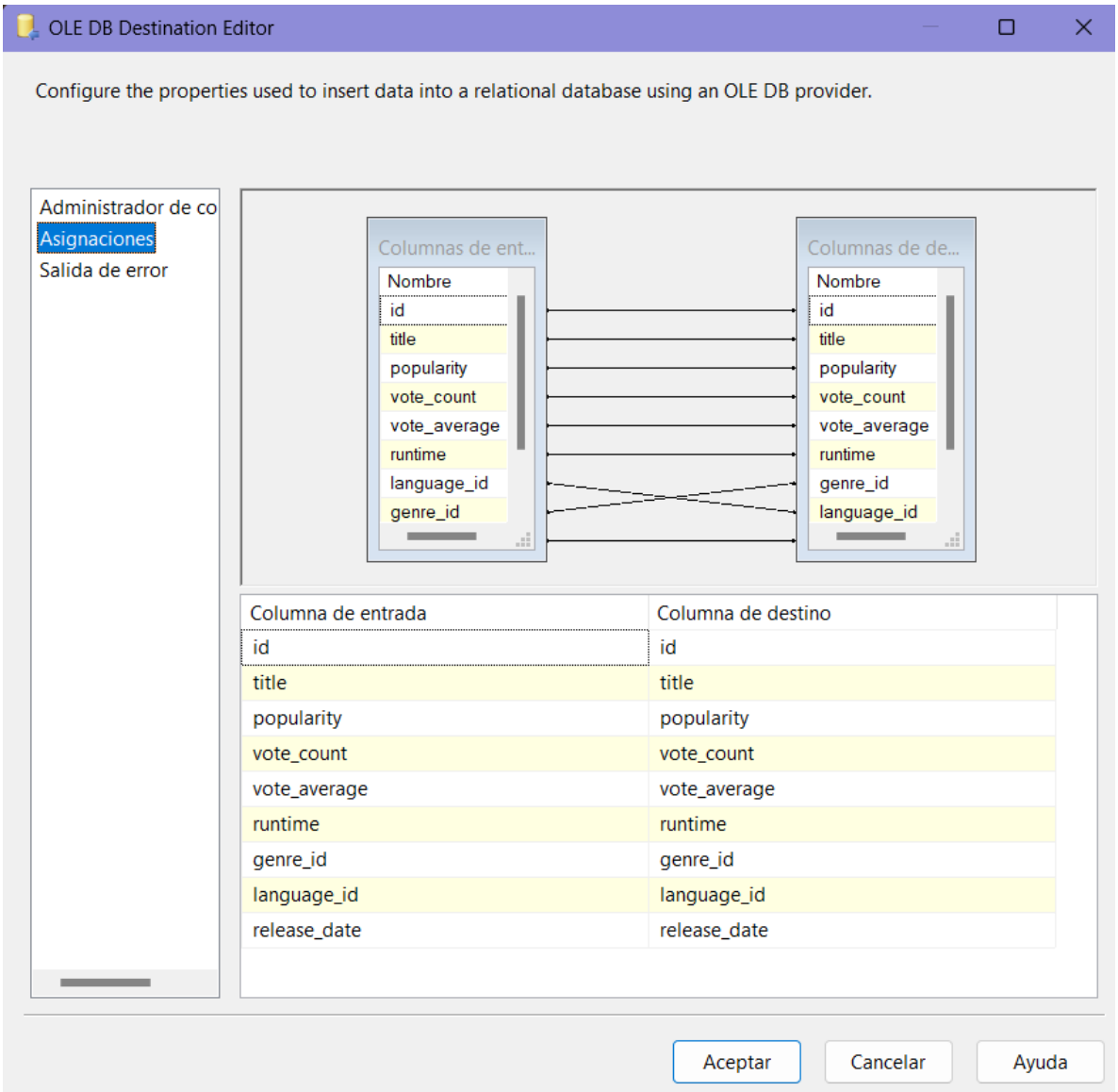

Figura 38. Destino OLE DB para facts_movies. Asignaciones

Se empieza llenando los datos en la columna izquierda con la base de datos y se asigna la tabla de hechos, así como la tabla de dimensiones.

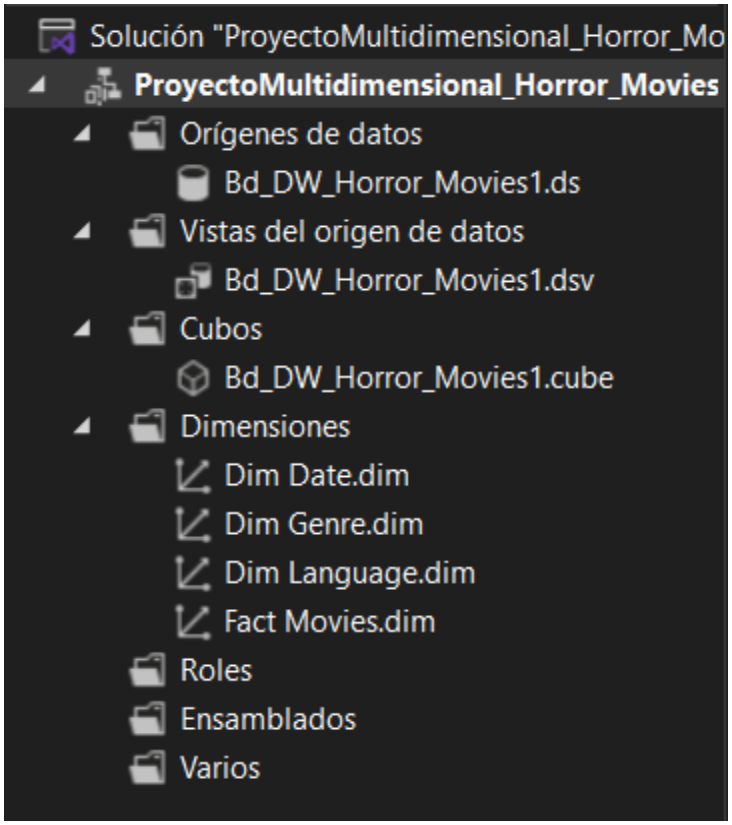

Figura 39. Proyecto Multidimensional

Lo anterior me ayuda a generar el siguiente diagrama representativo del cubo. Se procede a elegir las dimensiones que se utilizarán para filtrar los datos, pero como fue un proceso hecho en el GoogleColab, se omitirá ya que no se hicieron modificaciones.

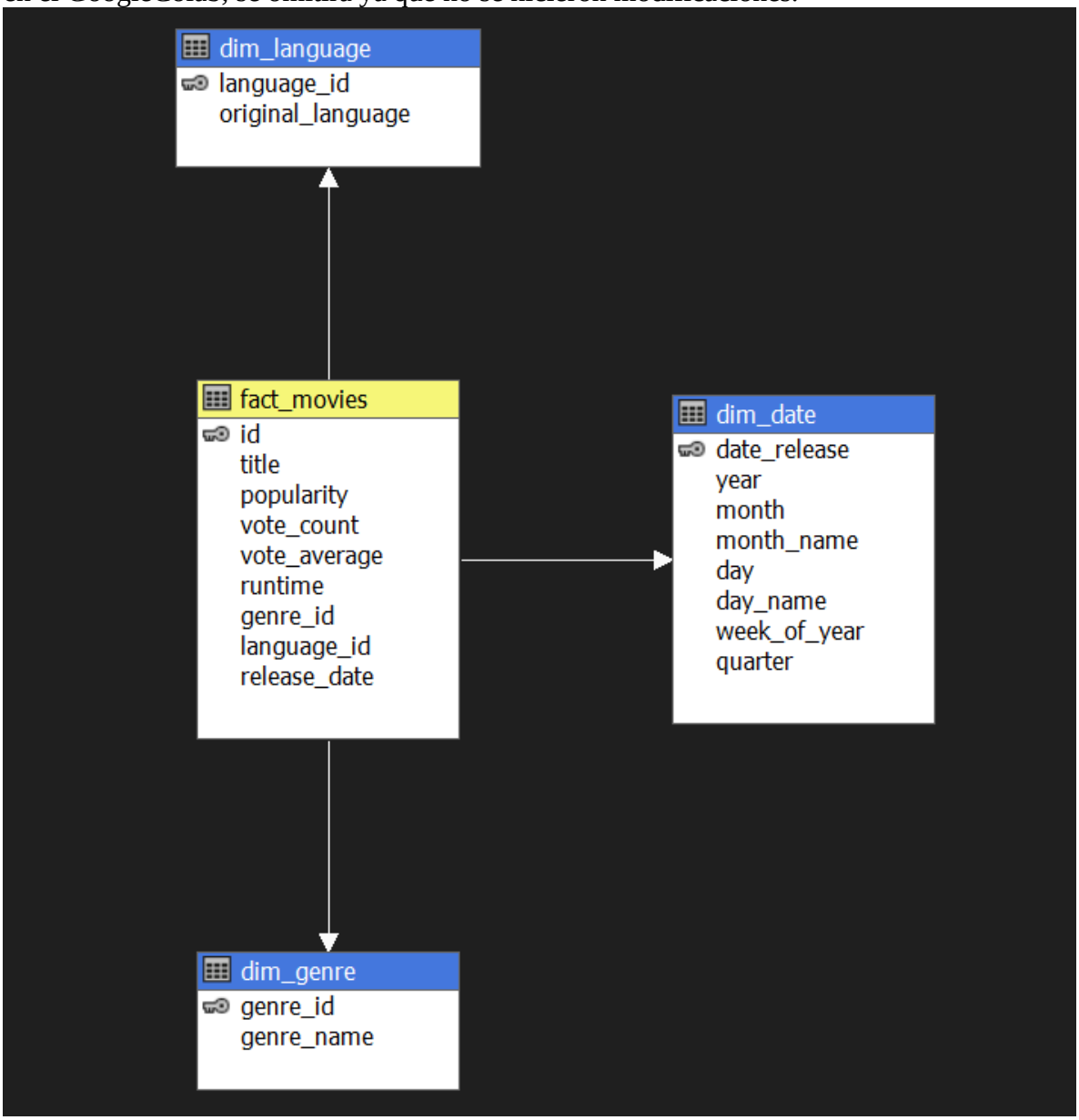

Figura 40. Diagrama del cubo

Se hace el proceso para exportar el cubo al análisis server, y si todo está correcto, se instala de forma exitosa apareciendo de esta forma:

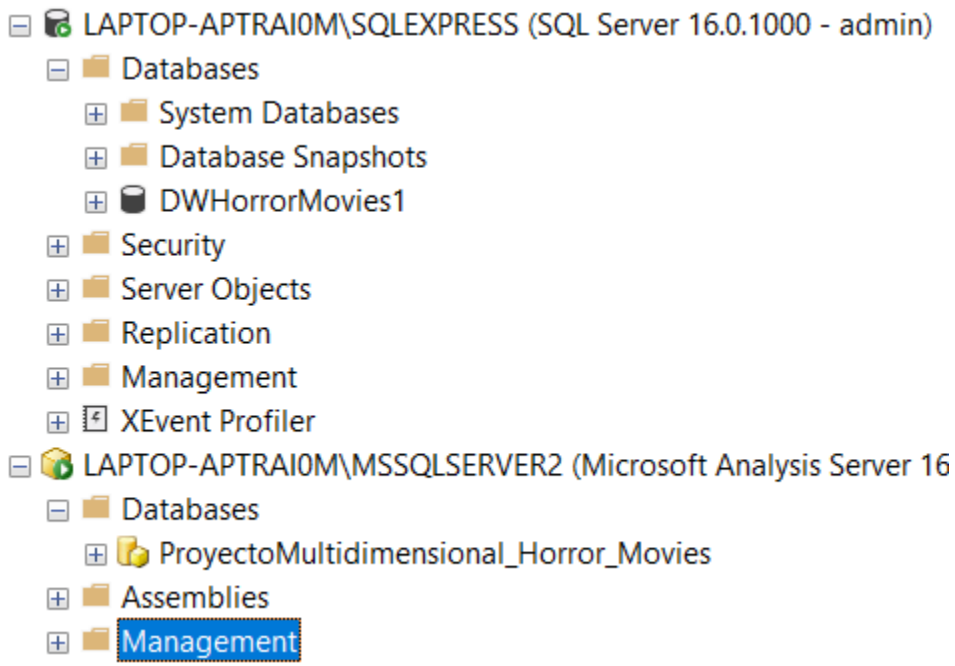

Figura 41. DWHorrorMovies1 y ProyectoMultidimensional_Horror_Movies en SQL

Luego se carga en Power BI y se seleccionan los filtros que se utilizarán para el Dashboard

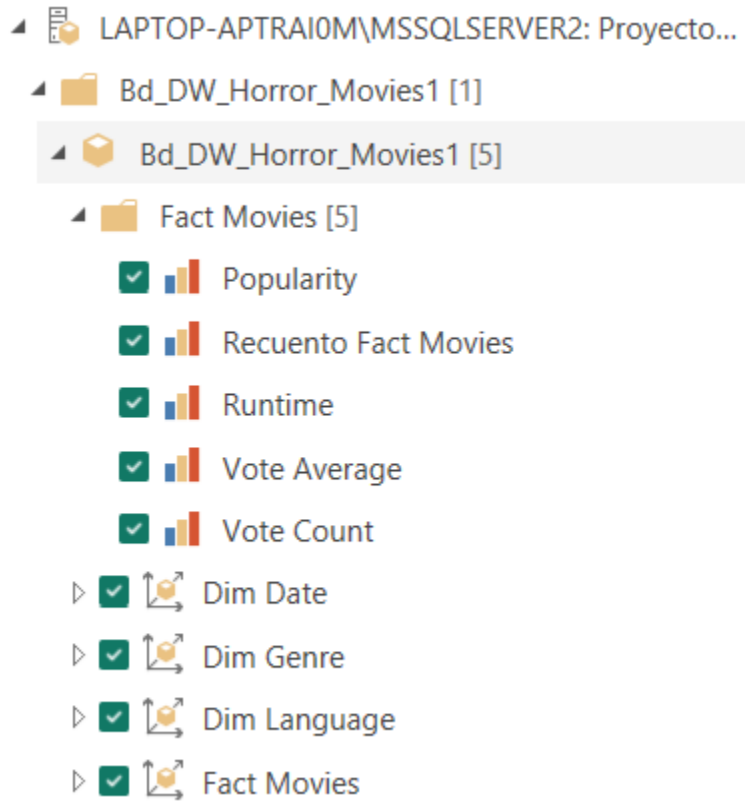

Figura 42. Filtros para los datos en Power BI

Así queda la lista a la derecha de las dimensiones y los totales por los cuales se pueden filtrar los diferentes diagramas

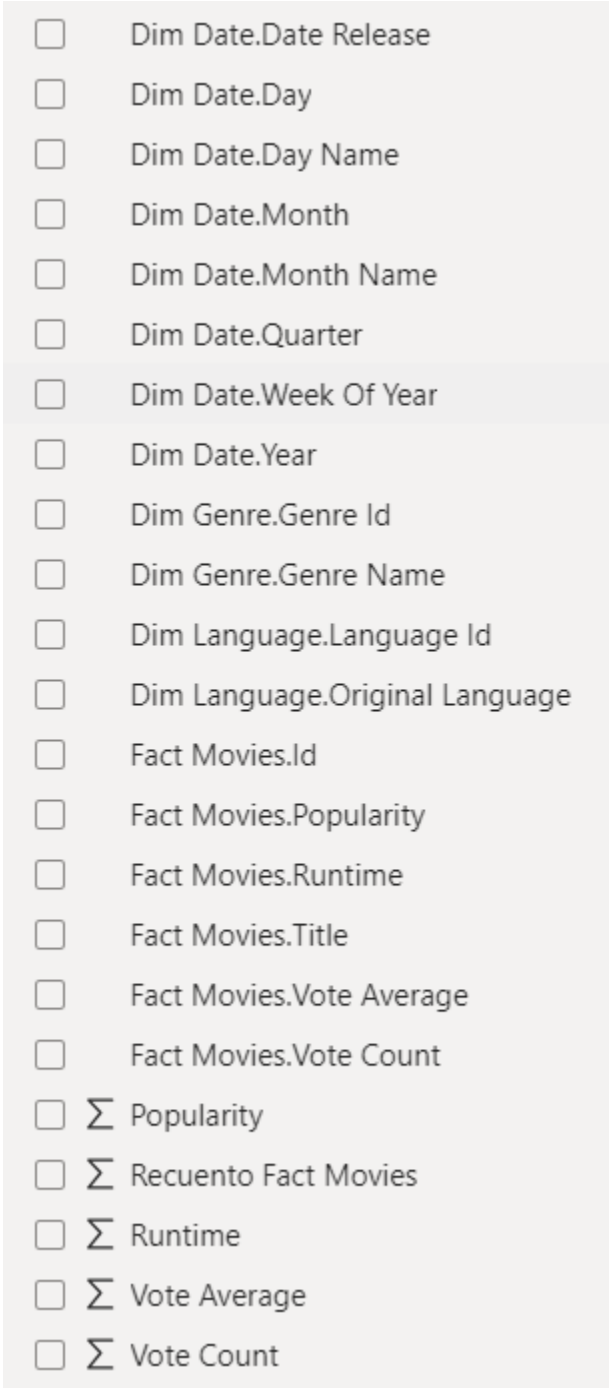

Figura 43. Dimensiones y totales finales

Para finalizar, este es el Dashboard final con las respectivas preguntas a resolver desde un inicio:

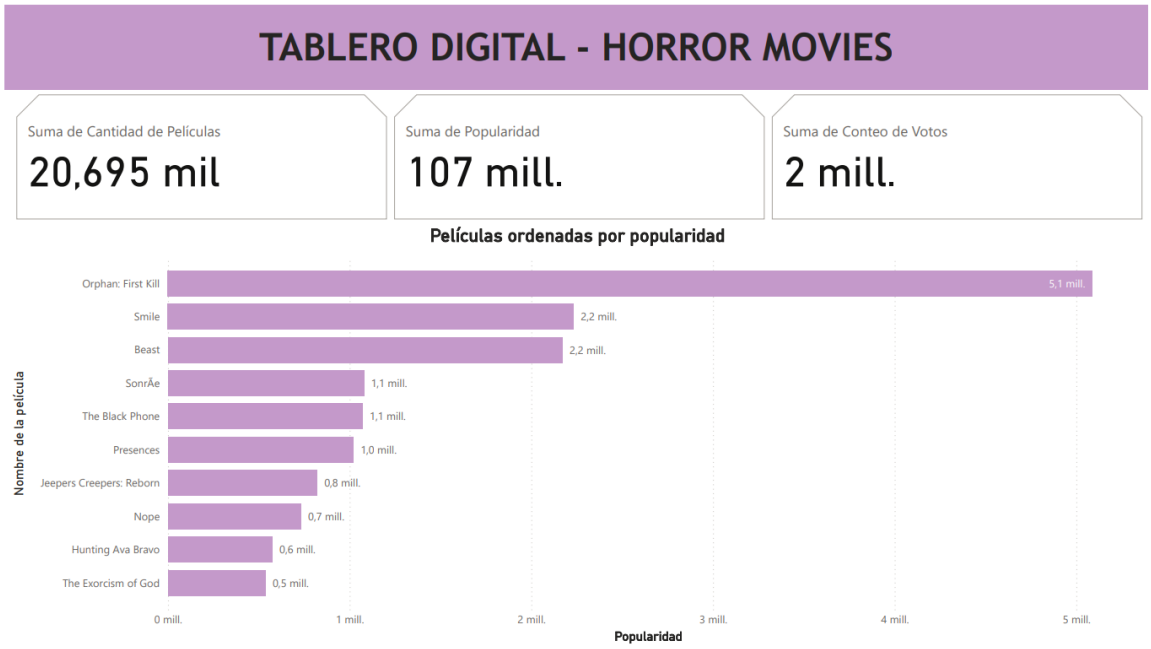

Figura 44. Dashboard. Página 1

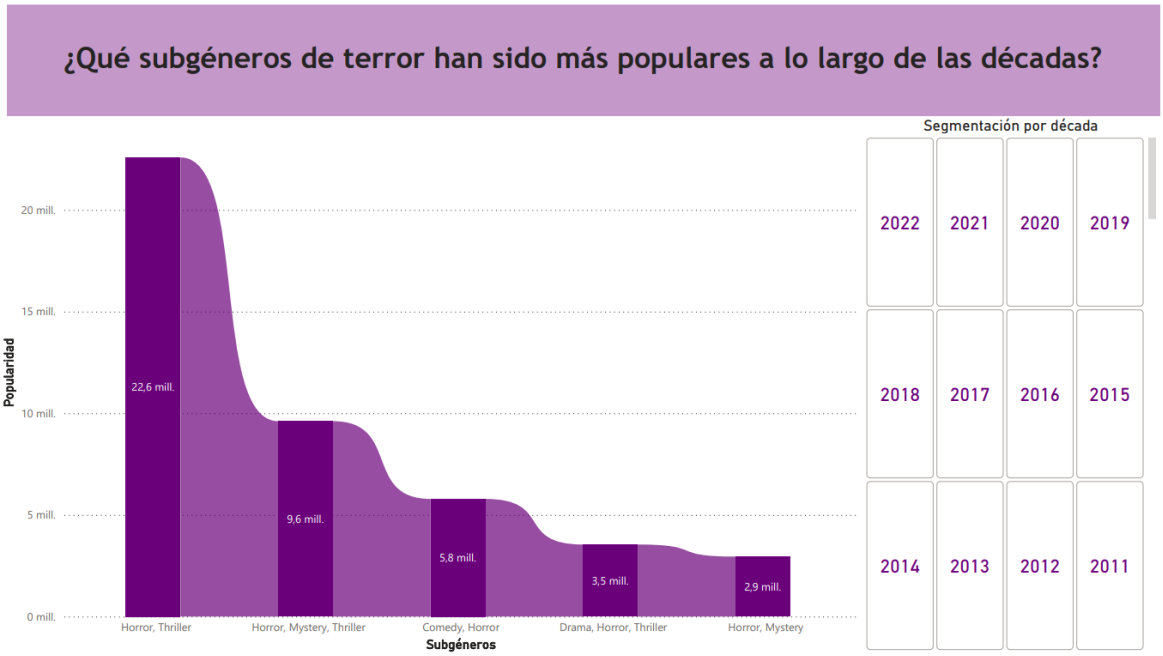

Figura 45. Dashboard. Página 2

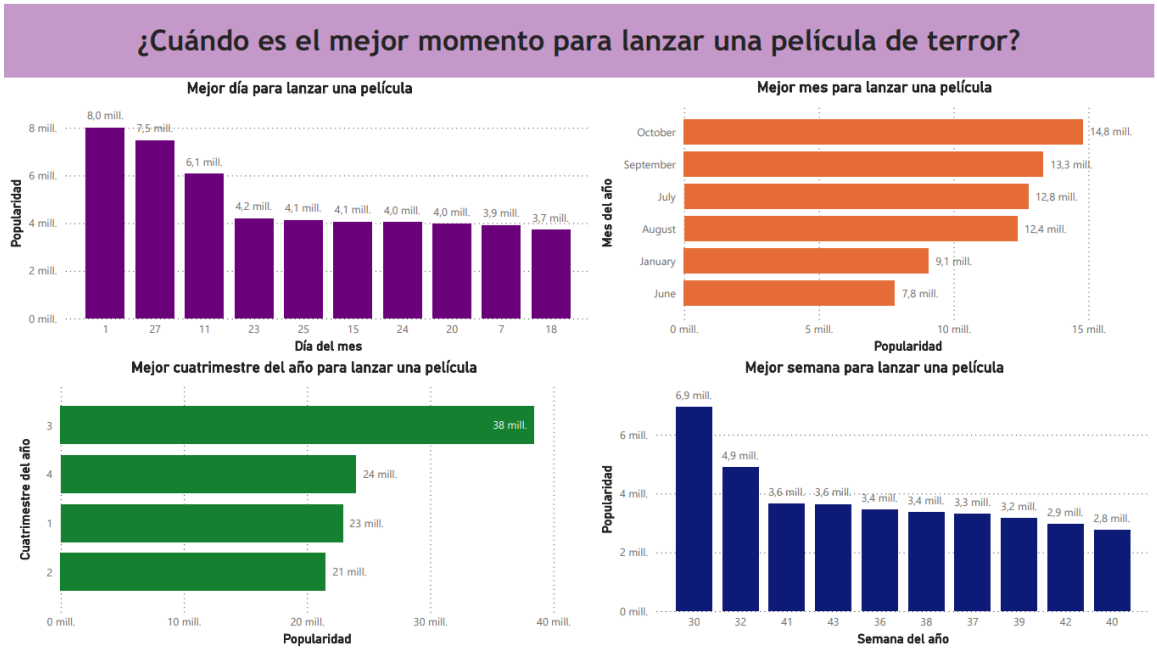

Figura 46. Dashboard. Página 3

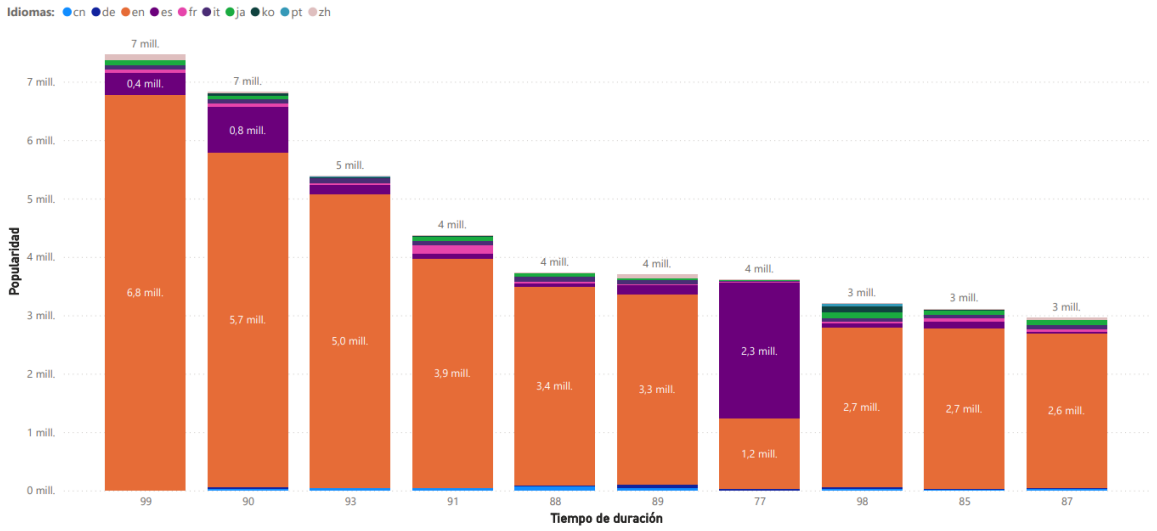

Figura 47. Dashboard. Página 4

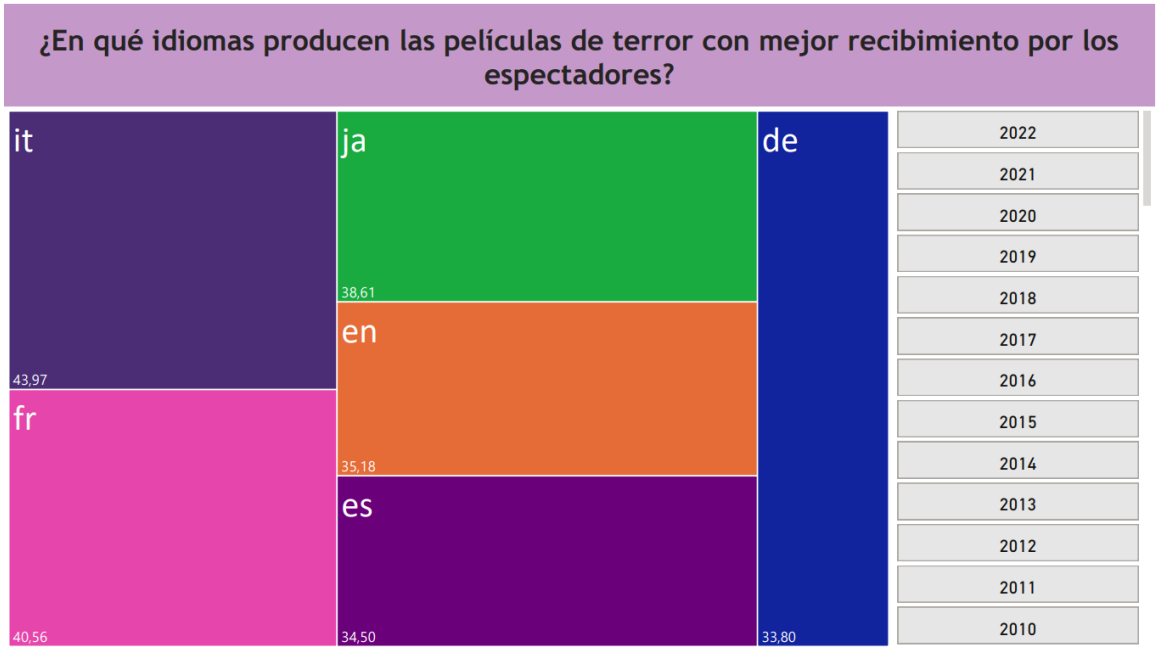

Figura 48. Dashboard. Página 5

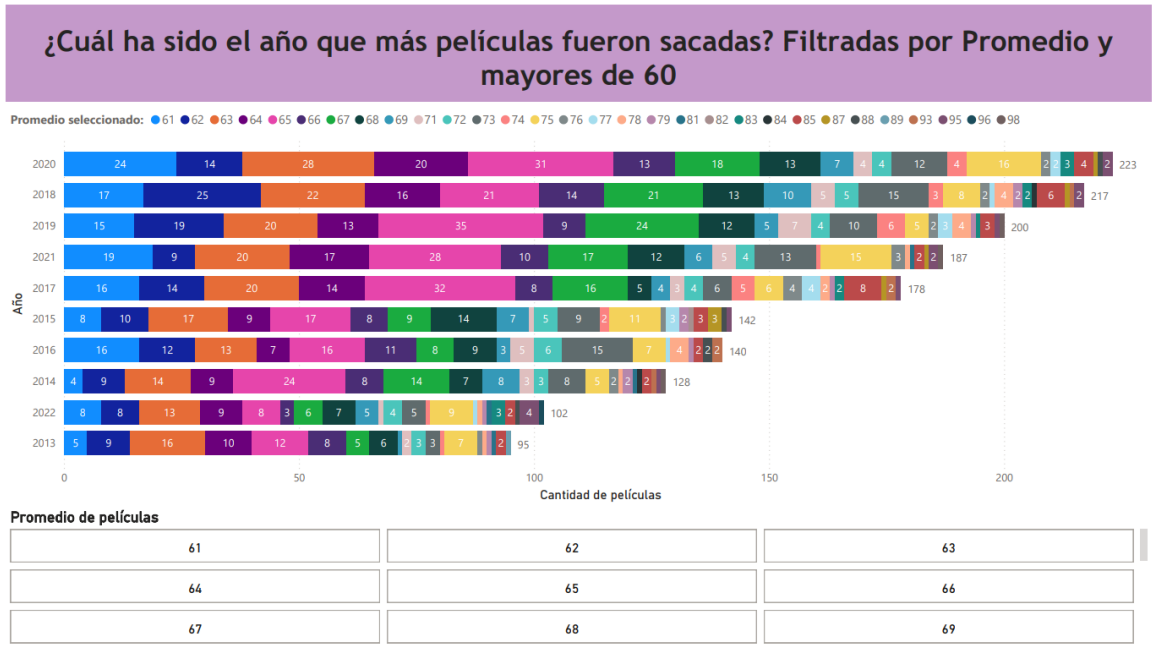

Figura 49. Dashboard. Página 6

CONCLUSIONES

- La película de terror con más popularidad en la plataforma es “Orphan: First Kill” en primer lugar, seguido por Smile y Beast.
- Los subgéneros más populares del terror son: “Horror, Thriller” en primer lugar, “Horror, Misterio y Thriller” en segundo lugar, “Comedia, Horror” en tercer lugar
- Lo mejores momentos para sacar una película de terror son:
    - Por día del mes: el 1, 27 y 11 son los 3 días que más popularidad han tenido entre la comunidad.
    - Por mes del año: Octubre, septiembre y julio son los meses que más popularidad han tenido entre los fans de las películas de terror.
    - Por cuatrimestre: el tercer (3) cuatrimestre del año ha sido el que ha tenido más popularidad. Esto significa que sacar películas de terror entre los meses agosto, septiembre y octubre implica mejores números en cuanto a la popularidad.
    - Por semana del año: La semana 30 y la 32 destacan sobre las otras semanas como las semanas con más popularidad entre las películas de terror.
- Las películas de terror que duren 99 minutos (1 hora y 39 minutos) y fueron producidas en el idioma inglés, tienden a ser más populares. Aunque es curioso que las películas en español que duran 77 minutos (1 hora y 17 minutos) tienden a ser más populares, incluso por encima que las películas en inglés de esta misma duración.
- Las películas de terror en italiano tienen un mejor promedio en cuanto a la calificación con un 43,97%. Aunque las películas de terror francesas no se quedan atrás con un 40,56%.

- 2020, 2018 y 2019 fueron los 3 mejores años para las películas de terror en la historia, ya que se sacaron una gran cantidad de películas y de alta calificación por encima del 60 en una escala de 100, lo que vendrían siendo películas con más de 3 estrellas en una escala de 5.

Para terminar, podemos concluir que sacar una película el primer día (1) del mes décimo (octubre) del género “Horror y Thriller” en inglés y que dure 99 minutos, son factores claves que podemos considerar si buscamos el éxito de una película de terror.

REFERENCIAS.

1. Kaggle. *Horror movies dataset*. Kaggle. Retrieved August 30, 2024, from https://www.kaggle.com/datasets/sujaykapadnis/horror-movies-dataset
2. The Movie Database (TMDB). *The Movie Database (TMDB)*. Retrieved August 30, 2024, from https://www.themoviedb.org/
3. Google. *Google Colaboratory*. Retrieved September 2, 2024, from https://colab.research.google.com/drive/1H6rF3UgihC26PjC89AIgeFK8VswVQfe7?usp=sharing