PRACTICA ETL

Documentación Practica l SQL

Lucia Poyan

Claudia Gemeno

Olivia Ares

Noviembre 2024

1. **Objetivos del proyecto:**

El objetivo principal de este proyecto es diseñar e implementar un proceso ETL para integrar, limpiar y analizar datos relacionados con películas y programas de televisión provenientes de diversas fuentes, como Netflix e IMDb. Este proceso busca transformar datos no estructurados en información valiosa y accesible, sentando las bases para análisis futuros.

A través del ETL, se pretende identificar tendencias, analizar la popularidad y evaluar el desempeño de contenido en diferentes plataformas, géneros y regiones. Por ejemplo, las puntuaciones y votaciones de IMDb permiten detectar los géneros más apreciados por los usuarios en distintas regiones, mientras que los datos de Netflix ayudan a explorar patrones como la duración promedio de contenido exitoso.

Además, el proyecto tiene como meta estructurar los datos en una base SQL optimizada que permita relacionar métricas clave como géneros, países de producción y puntuaciones de manera eficiente. Esto facilitará consultas avanzadas, como "¿Cuáles son las películas de drama más populares en

EE.UU.?" o "¿Qué géneros tienen las mejores puntuaciones en IMDb en Europa?".

Finalmente, el sistema se diseñó para ser escalable, permitiendo la integración de nuevas fuentes de datos en el futuro y asegurando que el proceso sea adaptable y confiable para análisis continuos.

Justificación de la importancia de Realizar un Proceso ETL

Un proceso ETL es fundamental para transformar datos dispersos y no estructurados en información valiosa. En este proyecto, los datos provienen de diversas fuentes, cada una con su propio formato y características. El ETL permite combinar estas fuentes, unificando formatos y estableciendo relaciones claras entre tablas para lograr una integración coherente y eficiente.

Los datos en bruto suelen presentar problemas como duplicados, valores faltantes y formatos inconsistentes. La etapa de transformación aborda estos desafíos, mejorando la calidad de los datos y garantizando resultados confiables. Por ejemplo, se identificaron títulos faltantes en las tablas de Netflix mediante consultas SQL específicas, lo que permitió generar reportes detallados sobre las discrepancias y tomar acciones correctivas.

Finalmente, un proceso ETL bien diseñado asegura la escalabilidad del sistema, facilitando la integración de nuevas fuentes de datos en el futuro. Esto permite tomar decisiones informadas basadas en datos confiables, como identificar géneros populares en diferentes países para campañas de marketing en plataformas de streaming.

Visualización del Proceso ETL

El proceso ETL implementado se diseñó para garantizar que los datos fueran consistentes, limpios y preparados para su análisis. Este proceso abarcó las fases de extracción, transformación y carga, asegurando un flujo de trabajo eficiente y robusto.

En la fase de extracción, los datos fueron importados a las tablas raw_titles, best_movies_netflix y best_shows_netflix.

Durante esta etapa, se verificó la estructura y el formato de los datos, identificando problemas como valores fuera de rango, duplicados y campos nulos. Esta evaluación inicial permitió definir los pasos necesarios para la limpieza y transformación.

La fase de transformación incluyó diversas tareas críticas para mejorar la calidad de los datos. Se imputaron valores nulos en columnas clave como imdb_score y imdb_votes utilizando promedios calculados, mientras que los valores nulos en age_certification fueron reemplazados con "unknown". Además, se corrigieron duraciones inválidas en la columna runtime y se eliminaron filas con valores nulos en campos esenciales como title. También se normalizaron columnas textuales, como genres y production_countries, eliminando caracteres no deseados y transformando los valores a minúsculas para asegurar uniformidad.

Dentro de la transformación, se definieron métricas de calidad para evaluar completitud, consistencia, precisión y razonabilidad de los datos. Valores fuera de rango fueron corregidos, y los duplicados exactos se eliminaron, conservando aquellos que representaban datos válidos, como películas y series con el mismo título. Estas acciones garantizaron que los datos estuvieran listos para la siguiente fase del proceso.

Finalmente, en la fase de carga, se diseñaron tablas normalizadas, como la tabla de hechos fact_titles y las tablas de dimensiones dim_genres, dim_production_countries, dim_type y dim_age_certification. Estas tablas permitieron organizar los datos de manera eficiente, reduciendo redundancias y optimizando el modelo relacional. Además, se establecieron claves foráneas para conectar las tablas de hechos con sus respectivas dimensiones, asegurando consistencia y facilidad de consulta.

El flujo general del proceso ETL puede representarse de la siguiente forma: [Extracción de Datos] → [Limpieza y Validaciones] → [Normalización] → [Tablas Finales]

Este enfoque organizado no solo nos permitió resolver problemas de calidad, sino que también estableció una base sólida para análisis avanzados y futuros desarrollos.

2. Dataset

Enlace a los Datasets Utilizados:

https://www.kaggle.com/datasets/thedevastator/the-ultimate-netflix-tv-shows-and-movies-dataset?select=raw_titles.csv

Fuentes de Datos y Descripción:

Los datasets utilizados provienen de fuentes públicas confiables, como Kaggle, una plataforma líder en la comunidad de análisis de datos que proporciona acceso a una amplia variedad de conjuntos de datos. Estos datos fueron seleccionados debido a su relevancia para el análisis de películas y programas de televisión, su alta calidad y la diversidad de métricas que ofrecen. Esta variedad de datos permite realizar un análisis profundo y detallado, lo que facilita la identificación de tendencias, patrones y relaciones clave dentro de los datos. Además, el uso de fuentes públicas asegura la transparencia y fiabilidad de los datos, lo que es crucial para la precisión de los resultados obtenidos en el proyecto.

Data Catalog o Data Dictionary

Raw_titles.csv

Campo	Tipo	Descripción
index	INT	Índice numérico del registro.
id	VARCHAR(20)	Identificador único del título en la base de datos.
title	VARCHAR(255)	Título de la película o serie.

type	VARCHAR(50)	Tipo del título, como SHOW o MOVIE.
release_year	INT	Año de lanzamiento del título.
age_certification	VARCHAR(10)	Clasificación por edad, como TV-MA o PG.
runtime	INT	Duración de la película o duración promedio de episodios en series.
genres	VARCHAR (255)	Lista de géneros asociados al título.
Production_countries	VARCHAR (255)	Países donde se produjo el título.
seasons	DECIMAL(4, 1)	Numero de temporada.
Imdb_id	VARCHAR(20)	Identificador único del título en la base de datos IMDb.
Imdb_score	DECIMAL(3, 1)	Puntuación del título en IMDb.
Imdb_votes	DECIMAL(10, 1)	Número de votos que contribuyeron a la puntuación en IMDb.

Best Movies Netflix.csv

Campo	**Tipo**	**Descripción**
index	INT	Índice numérico del registro.
TITLE	VARCHAR(255)	Título de la película.
RELEASE_YEAR	INT	Año de lanzamiento de la película.
SCORE	DECIMAL(3, 1)	Puntuación de la película en Netflix.

NUMBER_OF_VOTES	INT	Número de votos recibidos para esta puntuación.
DURATION	INT	Duración de la película.
MAIN_GENRE	VARCHAR(50)	Genero principal de la película.
MAIN_PRODUCTION	VARCHAR(2)	Código del país de producción principal.

Best Shows Netflix.csv

Campo	**Tipo**	**Descripción**
INDEX	INT	Índice numérico del registro
TITLE	VARCHAR(255)	Título de la serie.
RELEASE_YEAR	INT	Año de lanzamiento del título.
SCORE	DECIMAL(3, 1)	Puntuación de la serie en Netflix.
NUMBER_OF_VOTES	INT	Número de votos recibidos para esta puntuación.
DURATION	INT	Duración de la serie.
NUMBER_OF_SEASONS	INT	Número de temporadas.
MAIN_GENRE	VARCHAR(50)	Genero principal de la serie.
MAIN_PRODUCTION	VARCHAR(2)	Código del país de producción principal.

Frecuencia de Actualización de los Datos

Los tres datasets utilizados en este proyecto, raw_titles, best_movies_netflix y best_shows_netflix, no han sido actualizados desde 2022. No se dispone de información sobre actualizaciones posteriores, lo que significa que los datos disponibles en ellos reflejan la situación de las plataformas hasta ese año. Por lo tanto, las métricas y la información contenida en estos datasets no están actualizadas y no reflejan cambios recientes en los títulos, puntuaciones o interacciones de los usuarios.

3. Características de los Datos:

Los datos utilizados en este proyecto provienen de diversas fuentes y abarcan una variedad de información relacionada con películas y series, incluyendo métricas de plataformas como Netflix e IMDb. Los datasets están organizados en formato tabular, con archivos CSV que contienen columnas y filas. Cada columna representa un atributo específico de las películas o series, como el título, año de lanzamiento, género principal, país de producción, duración, puntuación, entre otros. Las filas corresponden a instancias individuales, es decir, películas o series específicas.

Los tres datasets seleccionados contienen columnas que pueden vincularse entre sí utilizando atributos comunes como title y release_year. El formato tabular facilita la integración con herramientas como DBeaver y bases de datos relacionales, además de ser compatible con otros lenguajes de programación.

Los datos se clasifican en tres categorías principales:

1. Datos Numéricos:
   - Score
   - number_of_votes
   - duration

2. Datos Categóricos:
   - main_genre

- type
- production_countries
- title
- genres

3. Datos Nulos o Faltantes:
   - age_certification
   - seasons

Los datos seleccionados están diseñados para ser combinados y relacionados mediante claves lógicas comunes. El atributo title es el principal nexo entre las tablas de películas y series de Netflix y la tabla general de IMDb (raw_titles). Además, la columna release_year se utiliza como criterio secundario para asegurar que los datos correspondan al mismo título.

Por ejemplo, la tabla best_movies_netflix contiene datos específicos de Netflix, mientras que raw_titles complementa esta información con puntuaciones y votos de IMDb. De manera similar, las series en la tabla best_shows_netflix se vinculan con raw_titles para incluir datos adicionales como géneros más amplios y países de producción.

Calidad de los Datos

En este informe se presenta una evaluación de la calidad de los datos en tres datasets: raw_titles, best_movies_netflix y best_shows_netflix. El análisis se centró en diversas métricas clave como completitud, consistencia, precisión, y identificabilidad, que permiten diagnosticar el estado general de los datos y guiar las transformaciones necesarias.

Para calcular las métricas de calidad, se creó una tabla temporal para cada dataset, en la cual se insertaron los cálculos correspondientes a la media de cada métrica por columna. Primero, se calculó la media de cada métrica por columna dentro de cada dataset. Posteriormente, se calculó la media de todas las métricas para obtener un porcentaje final de calidad de la tabla. Este enfoque permitió evaluar la calidad de los datos de manera

integral, considerando tanto los valores individuales de cada columna como el conjunto total de métricas para cada dataset. La media de cada métrica en todas las columnas se utilizó para determinar el porcentaje final de calidad de los datos en cada tabla.

1. Dataset: raw_titles

Metric_Name	**Average_Metric_Value**
Precisión	82.69
Semántica	92.44
Estructura	65.27
Completitud	93.08
Consistencia	33.32
Razonabilidad	75.77
Identificabilidad	33.37
Media Final	**67,98 %**

El dataset raw_titles presenta un buen nivel de completitud en sus columnas principales, con pocos valores faltantes. Sin embargo, se detectaron varios problemas de consistencia, particularmente en las columnas genres y production_countries, donde los valores no son coherentes debido a inconsistencias en los caracteres y formato, existiendo corchetes y comillas en las celdas. También se observaron problemas de precisión, como valores fuera del rango lógico, especialmente en la columna runtime, donde algunas duraciones eran negativas o excesivamente altas. Además, duplicados en la columna title revelaron que había versiones diferentes de un mismo contenido, lo cual podría generar ambigüedades en futuros análisis. En general, la calidad de este dataset es aceptable, con un 67,98% total, pero se recomienda realizar una limpieza en los valores faltantes, inconsistentes y duplicados para asegurar su fiabilidad.

2. Dataset: best_movies_netflix

Metric_Name	Average_Metric_Value
Precisión	85.49
Semántica	100
Estructura	100
Completitud	100
Consistencia	37.13
Razonabilidad	85.71
Identificabilidad	37.13
Media Final	**77,92 %**

El dataset best_movies_netflix muestra una calidad aceptable en términos de completitud y estructura, sin ausencias en columnas clave. No obstante, se detectaron algunas inconsistencias en valores como release_year y score, donde algunos datos no se ajustan al formato o los rangos esperados. La precisión de los datos en general es alta, pero aún se observan valores atípicos en campos como score, que pueden afectar los análisis si no se corrigen. Además, la identificabilidad de los registros es adecuada, ya que no se encontraron duplicados significativos. En general, el dataset es funcional, con un 77,9% pero los problemas de consistencia y precisión deben ser atendidos para mejorar la calidad de los datos.

3. Dataset: best_shows_netflix

Metric_Name	Average_Metric_Value
Precisión	100
Semántica	100
Estructura	100
Completitud	100
Consistencia	32.31
Razonabilidad	100
Identificabilidad	32.31
Media Final	**80.66 %**

El dataset best_shows_netflix presenta una completitud adecuada en la mayoría de sus columnas, aunque se detectaron valores faltantes en campos como duration y number_of_seasons, que deberían ser corregidos. A pesar de ser completo, el dataset tiene problemas de consistencia y precisión, sobre todo en las columnas release_year y score. Sin embargo, la identificabilidad es buena, con un 80,66%, sin duplicados significativos, lo que facilita el trabajo con los datos. Para mejorar la calidad, es necesario realizar una limpieza de los valores faltantes y revisar los valores erróneos o inconsistentes en los campos mencionados.

No aplicamos directamente las métricas de linaje, puntualidad e integridad en el código SQL debido a las características y el alcance del proyecto:

- **Linaje**: esta métrica se refiere a la capacidad de rastrear el origen y las transformaciones de los datos a lo largo del proceso ETL. No se consideró crítica en este caso porque los datasets de Kaggle no incluyen información sobre su historial de cambios o transformaciones previas.

- **Puntualidad**: evalúa si los datos están actualizados al momento del análisis. Los datasets utilizados provienen de Kaggle y, aunque originalmente podrían haberse actualizado con cierta periodicidad, la última actualización disponible para todos ellos corresponde a 2022. No se dispone de información sobre actualizaciones posteriores, lo que significa que los datos utilizados no reflejan cambios recientes en las métricas o en la información.
  Dado que trabajamos con versiones estáticas de estos datasets, no se puede garantizar que los datos estén actualizados, por lo que no fue necesario verificar la puntualidad de los mismos en este análisis.

- **Integridad**: esta métrica asegura que los datos estén completos y que las relaciones entre las tablas sean consistentes. En este proyecto, asumimos que los datasets cargados eran independientes y que no era necesario integrarlos en un sistema más amplio. El enfoque

estuvo en la limpieza y preparación de los datos para análisis, sin realizar validaciones adicionales sobre relaciones complejas o integridad referencial.

Para abordar estas problemáticas, se implementaron transformaciones específicas. Los valores nulos en imdb_score y imdb_votes fueron imputados con promedios, mientras que en age_certification se asignó "unknown" para mantener la consistencia. Las columnas textuales fueron normalizadas eliminando caracteres innecesarios y transformando los valores a minúsculas. Asimismo, se eliminaron duplicados exactos, pero se conservaron registros válidos con ligeras diferencias.

Después de las transformaciones, se lograron mejoras significativas en la calidad de los datos. Las columnas clave alcanzaron un 100% de completitud, y no se encontraron valores fuera de rango en las columnas numéricas. Además, los títulos y los campos textuales son ahora consistentes, lo que facilita su análisis. Este proceso garantizó un dataset confiable y listo para su uso en análisis avanzados.

Consistencia de los Datos

Durante el proceso de transformación del dataset raw_titles, así como de las tablas de Netflix (best_movies_netflix y best_shows_netflix), se realizaron diversas acciones para garantizar la consistencia y calidad de los datos. Estas acciones son fundamentales para un análisis relacional preciso y para asegurar la integridad de las relaciones entre tablas en el modelo estrella implementado. A continuación, se detallan las principales transformaciones realizadas:

1. **Eliminación de Formatos Inconsistentes en las Columnas genres y production_countries:**
    - Estas columnas contenían datos en forma de listas con corchetes ([]) y comillas adicionales (''). Este formato dificultaba su manipulación y análisis.

- Se implementó la función REPLACE en SQL para eliminar dichos caracteres no deseados, dejando únicamente los valores necesarios.
- Esto permitió que los valores en estas columnas quedaran en un formato plano y consistente, adecuado para la creación de dimensiones en el modelo estrella.

2. **Normalización del Formato de Texto:**
    - Las columnas que contenían texto, como title, genres, y production_countries, presentaban inconsistencias en el uso de mayúsculas/minúsculas y espacios en blanco adicionales.
    - Se aplicaron las funciones LOWER para convertir todo el texto a minúsculas y TRIM para eliminar espacios en blanco no deseados.
    - Esta normalización facilitó la relación entre tablas y mejoró la precisión al unificar los formatos textuales.
3. **Gestión de Valores Nulos e Inconsistentes:**
    - En columnas como age_certification, los valores nulos se reemplazaron con la etiqueta unknown para garantizar que no quedaran huecos en los datos, lo cual es importante para los análisis relacionales y la generación de métricas.
    - Esta acción permitió estandarizar los valores en esta columna y evitar problemas de relación con otras tablas.
4. **Verificación y Resolución de Duplicados:**
    - Se detectaron y eliminaron registros duplicados en la tabla raw_titles. En particular, se observaron duplicados causados por películas y series con el mismo título o películas con el mismo título lanzadas en distintos años.
    - Para resolver esto, se creó una tabla temporal que identificó los registros únicos basados en una combinación de las columnas title, release_year, y type. Posteriormente, los

duplicados fueron eliminados utilizando esta tabla como referencia.

5. **Preparación para la Relación entre Tablas:**
    - Las transformaciones mencionadas garantizaron la consistencia entre los datos de raw_titles y las tablas de Netflix (best_movies_netflix y best_shows_netflix).
    - Estas acciones aseguraron que las relaciones establecidas en el modelo estrella, como la asociación entre los géneros, los países de producción, y las certificaciones por edad, fueran coherentes y estuvieran libres de ambigüedades.

En resumen, las acciones implementadas en esta etapa de transformación fueron críticas para garantizar que los datos fueran consistentes, normalizados y listos para su uso en un modelo relacional robusto. Esto permitió integrar los diferentes datasets de manera efectiva y preparar el sistema para análisis y consultas eficientes.

Precisión de los Datos

Durante el proceso de transformación, se abordaron varios aspectos críticos relacionados con la precisión de los datos en las tablas procesadas, asegurando que las métricas derivadas fueran fiables y representativas de los datos reales.

1. **Validación de Valores en la Columna runtime de raw_titles:**
    - Se identificaron valores atípicos, como duraciones negativas o superiores a 300 minutos, los cuales no son realistas en el contexto de películas o programas de televisión.
    - Para corregir esta discrepancia, se aplicaron validaciones mediante condiciones que establecieron un rango aceptable entre 10 y 300 minutos. Los registros fuera de este rango fueron evaluados individualmente para determinar si debían eliminarse o imputarse con valores más coherentes.

2. **Control de Precisión en la Columna score de best_movies_netflix:**
    - En esta columna, que representa puntuaciones de usuarios, se detectaron valores fuera del rango estándar de 0 a 10.
    - Los valores anómalos fueron reemplazados con la media calculada del conjunto de datos, garantizando que las métricas derivadas no se vieran afectadas por estos errores. Esto también permitió mantener una coherencia en los análisis posteriores.
3. **Inconsistencias entre imdb_score e imdb_votes en raw_titles:**
    - Se evaluaron casos en los que existían diferencias significativas entre las puntuaciones (imdb_score) y el número de votos (imdb_votes), lo que podía indicar datos atípicos o errores de origen.
    - Se documentaron estos casos y se analizaron para comprender su origen y ajustar las reglas de validación aplicadas. En algunos casos, los valores fuera de rango o inconsistentes fueron imputados con valores promedios calculados a partir de datos válidos.
4. **Aplicación de Rangos de Validación:**
    - Los rangos de validación establecidos, como el de runtime entre 10 y 300 minutos y el de score entre 0 y 10, se implementaron para filtrar valores incorrectos antes de utilizarlos en análisis o integrarlos en el modelo estrella.
    - Estas validaciones se aplicaron mediante condiciones SQL y cálculos de promedios que aseguraron la integridad de los datos.
5. **Registro de Valores Atípicos:**
    - Los valores que se encontraron fuera de rango fueron documentados como parte del proceso de validación para comprender su origen y ajustar futuras reglas de validación.

Estas acciones mejoraron significativamente la precisión de los datos en las tablas utilizadas, permitiendo que las métricas y análisis derivados reflejaran comportamientos reales y consistentes, alineados con las expectativas del dominio de estudio.

Razonabilidad de los Datos

Durante la transformación de los datos, se implementaron ajustes para garantizar la razonabilidad y coherencia de las variables en relación con el contexto de cada tipo de contenido. Las medidas adoptadas se detallan a continuación, tomando como referencia el análisis y las correcciones realizadas en el código SQL proporcionado:

1. **Corrección de la Columna seasons en raw_titles:**
    - Se detectaron inconsistencias en la columna seasons para títulos clasificados como películas (type = 'MOVIE'). Dado que las películas no tienen temporadas, este atributo no es relevante en estos casos.
    - Para resolver esta inconsistencia, se estableció el valor de la columna seasons a 0 para todos los registros donde el tipo fuera "MOVIE" y el valor de seasons fuera nulo. Esto aseguró que la columna reflejara correctamente que el atributo no aplicaba en ese contexto.
2. **Ajuste de la Columna duration en best_shows_netflix:**
    - Se observaron discrepancias en los valores de duración (duration) para ciertos títulos en la tabla best_shows_netflix, lo que sugería posibles errores de origen o valores atípicos.
    - Los valores incoherentes en esta columna fueron reemplazados con el promedio calculado de la duración, utilizando únicamente datos válidos. Esto permitió estandarizar los valores y eliminar posibles sesgos en el análisis posterior.

3. **Validación Contextual de los Datos:**
    - Cada ajuste se realizó teniendo en cuenta la semántica y el uso esperado de las variables. Por ejemplo, las temporadas (seasons) solo son relevantes para series de televisión, mientras que la duración (duration) debe ser razonable en función del tipo de contenido.
4. **Impacto en la Calidad de los Datos:**
    - Estas medidas garantizaron que los datos fueran razonables y coherentes en su contexto, eliminando posibles errores que pudieran afectar los análisis relacionales o los resultados derivados de métricas agregadas.

Al implementar estas acciones, se logró un conjunto de datos más limpio y ajustado al propósito de análisis, respetando las particularidades de cada tipo de contenido (películas y series) y eliminando información errónea que podría distorsionar los resultados.

Identificabilidad y Unidad de los Datos:

En el dataset raw_titles, se identificaron registros duplicados debido a diferentes versiones del mismo título, como títulos en distintos años o ediciones remasterizadas. Para abordar este problema, se implementaron estrategias de limpieza que incluyeron:

- **Eliminación de duplicados**: Se retuvo un único registro por cada título basándose en la combinación de columnas clave como title, release_year y type. Adicionalmente, en los casos donde aún persistían duplicados, se seleccionaron los registros más relevantes utilizando métricas como imdb_score y imdb_votes.

En las tablas de Netflix (best_movies_netflix y best_shows_netflix), no se encontraron duplicados evidentes. Sin embargo, se verificó la unicidad mediante combinaciones clave como title y release_year. También se realizó una validación cruzada con la tabla raw_titles para asegurar la consistencia entre los datasets.

La implementación de estas acciones garantizó:

1. **Eliminación de registros redundantes** que podrían sesgar los análisis o complicar las relaciones entre tablas.
2. **Preservación de los datos más relevantes y confiables**, asegurando la calidad de los resultados posteriores.

Además, las métricas de calidad calculadas durante la etapa de transformación permitieron monitorear el impacto de estas acciones correctivas, verificando que el conjunto final de datos mantuviera integridad y consistencia.

Gracias a las medidas adoptadas, el dataset procesado presenta un esquema limpio, sin duplicados ni inconsistencias críticas. Esto asegura que las consultas y análisis derivados del proyecto sean confiables y relevantes para la toma de decisiones. El conjunto de datos final está optimizado para su uso en análisis y reportes posteriores, cumpliendo con los estándares requeridos para un proyecto ETL.

Limpieza de los Datos:

La limpieza de datos fue un paso crucial para garantizar la calidad del conjunto final y la consistencia entre las diferentes tablas utilizadas en el proyecto. Este proceso se llevó a cabo aplicando técnicas específicas en las fases de transformación y carga, teniendo en cuenta los problemas identificados en los datasets iniciales. A continuación, se detallan las principales acciones de limpieza realizadas:

1. **Eliminación de valores nulos y blancos:**
    - En el dataset raw_titles, las columnas imdb_score y imdb_votes presentaban valores nulos. Se imputaron estos valores utilizando la media de las respectivas columnas para no distorsionar las estadísticas.
    - La columna age_certification también contenía valores nulos o cadenas vacías. Se reemplazaron los nulos con el valor

genérico 'unknown' para mantener la consistencia y evitar pérdidas de información en los análisis.

- Las películas (type = 'MOVIE') que tenían valores nulos en la columna seasons fueron actualizadas a 0, ya que este atributo no es aplicable a películas.

2. **Normalización del formato de texto:**
    - Las columnas title, genres, y production_countries en raw_titles fueron limpiadas eliminando caracteres especiales como corchetes y comillas mediante funciones como REPLACE.
    - Para unificar el formato y facilitar la vinculación entre tablas, se utilizaron funciones como LOWER y TRIM, normalizando todas las entradas de texto a minúsculas y eliminando espacios innecesarios.
3. **Gestión de valores fuera de rango:**
    - En la columna runtime, se identificaron y corrigieron valores que estaban fuera del rango esperado (10 a 300 minutos). Los valores atípicos fueron imputados con la media de los valores válidos.
    - En las tablas de Netflix, como best_movies_netflix, se revisaron columnas como score, asegurando que los valores estuvieran dentro del rango aceptable (0 a 10).
4. **Resolución de duplicados:**
    - En raw_titles, se identificaron registros duplicados que representaban películas y series con el mismo título o diferentes versiones del mismo título. Se creó una tabla temporal para retener los registros únicos, priorizando aquellos con los mejores valores de imdb_score y imdb_votes.
    - Este proceso incluyó la validación de combinaciones clave como title, release_year, y type para evitar la pérdida de información crítica.

5. **Creación de datos consistentes para las dimensiones:**
    - Para las columnas genres y production_countries, que contenían listas de valores separados por comas, se implementó una separación y limpieza de los valores. Esto permitió crear tablas de dimensiones (dim_genres y dim_production_countries) con datos únicos y consistentes.
    - Se manejaron las categorías de forma exhaustiva para evitar duplicados o valores inconsistentes en las dimensiones.
6. **Preparación de datos relacionales:**
    - Se aseguraron las relaciones entre tablas mediante claves foráneas y verificaciones de integridad referencial. Las dimensiones (dim_genres, dim_production_countries, dim_type, y dim_age_certification) se vincularon correctamente con la tabla de hechos (fact_titles) tras validar la consistencia de los valores.

Impacto de la Limpieza de los Datos:

Estas acciones permitieron obtener un conjunto de datos limpio, consistente y apto para su análisis. Se corrigieron problemas de formato, se eliminaron valores fuera de rango, se imputaron valores nulos y se resolvieron duplicados. Este trabajo asegura que las métricas y resultados derivados sean confiables, relevantes y representativos del comportamiento real de los datos.

El enfoque adoptado en la limpieza refuerza la solidez del proyecto y garantiza que el dataset final cumpla con los estándares de calidad exigidos para un proceso ETL exitoso.

Relevancia del Catalogado:

El diseño del catalogado en este proyecto ETL se centró en garantizar que los datos transformados sean relevantes y alineados con los objetivos del análisis. Para ello, se seleccionaron cuidadosamente los atributos clave que

reflejan aspectos esenciales del contenido, como title, release_year, runtime, imdb_score, y genres, complementados con variables adicionales como production_countries, age_certification, y type. Esto permite abarcar múltiples perspectivas en los análisis, como géneros más populares, duración promedio o tendencias en certificación por edades.

Se implementó un modelo estrella estructurado en:

- **Dimensiones**: Tablas como dim_genres, dim_production_countries, dim_age_certification, y dim_type, que organizan y normalizan los datos categóricos. Esto asegura que se puedan realizar segmentaciones detalladas en el análisis.
- **Tabla de hechos**: Consolidó métricas esenciales vinculándolas a las dimensiones, permitiendo un acceso eficiente y consultas relacionales consistentes.

El proceso de catalogado incluyó pasos fundamentales como la eliminación de duplicados, la validación de relaciones entre tablas y la normalización de formatos. Esto asegura que los datos no solo sean representativos del contenido disponible, sino que también estén optimizados para análisis más complejos, como la identificación de tendencias de consumo y preferencias regionales.

Este enfoque asegura que los datos son no solo relevantes, sino también fiables y adaptados a los requerimientos del proyecto, maximizando su utilidad para futuras aplicaciones analíticas.

Categorización del Proceso ETL

Con base en el trabajo realizado y la estructura implementada, se puede concluir que este proyecto corresponde a la construcción de un Data Mart. A continuación, se explican las razones que sustentan esta categorización:

1. Enfoque Temático: El proyecto se centra en datos relacionados exclusivamente con películas y series, organizados en dimensiones como géneros, países de producción, clasificaciones por edad y tipos (películas o

series). Este enfoque temático es característico de un Data Mart, diseñado para cubrir necesidades específicas de análisis.

2. Modelo Estrella: Se ha implementado un modelo estrella, estructurado en una tabla de hechos (fact_titles) y múltiples tablas de dimensiones (dim_genres , dim_production_countries , dim_age_certification , y dim_type). Este diseño es común en Data Marts debido a su capacidad para optimizar consultas analíticas.

3. Datos Listos para Consulta: Los datos han sido extraídos, transformados y cargados de manera que están listos para ser utilizados en análisis específicos. Esto incluye limpieza de datos, eliminación de inconsistencias, y estructuración para garantizar su calidad y utilidad.

Por qué no es otro tipo de almacenamiento de datos:

- No es un Data Warehouse: A diferencia de un Data Warehouse, este proyecto no busca integrar datos de múltiples áreas temáticas o departamentos dentro de una organización. Su alcance está limitado a un único dominio: entretenimiento.

- No es un Data Lake: En un Data Lake, los datos se almacenan en su estado bruto y sin procesar, mientras que aquí los datos han sido procesados, estructurados y optimizados para análisis.

- No es un Data Lakehouse: Aunque utiliza un modelo estructurado, no incorpora datos en múltiples formatos (estructurados y no estructurados) ni combina características de Data Lakes y Data Warehouses.