Relatório do Projeto ETL e Análise de Conteúdo Disney

Este relatório descreve o processo de Elaboração, Transformação e Carga (ETL) e a subsequente Análise Exploratória de Dados (EDA) realizada sobre o catálogo de filmes e séries da Disney Plus.

1. Dados

1.1. Descrição e Fonte dos Dados

No projeto utilizei o dataset "Disney Movies and TV Shows". Este conjunto de dados contém um catálogo de títulos disponíveis na plataforma de streaming Disney+, incluindo informações cruciais sobre o conteúdo como data que foi adicionado a plataforma, classificação etária, ano de lançamento, duração em minutos ou temporadas, gênero e etc.

- Link para a Fonte de Dados: https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows/data ("shivamb/disney-movies-and-tv-shows")

1.2. Motivação da Escolha

A escolha deste dataset se deu pelo interesse em explorar as estratégias de conteúdo de uma das maiores produtoras de mídia do mundo. É interessante analisar como a Disney tem utilizado seu catálogo (Filmes vs. Séries de TV), quais gêneros estão em alta (como Animação) e como a duração dos filmes se relaciona com as classificações indicativas. A análise pode revelar insights sobre a produção de conteúdo na última década.

2. Extração e Transformação (ETL)

2.1. Extração

A extração dos dados foi realizada diretamente da plataforma KaggleHub utilizando a biblioteca kagglehub do Python. O arquivo disney_plus_titles.csv foi baixado e carregado para um DataFrame do pandas (df_disney_raw).

2.2. Tratamento e Limpeza (Transformação)

A fase de transformação foi essencial para estruturar os dados para a análise e o carregamento no banco de dados:

- Criação de Chave Primária: Foi criada uma coluna id_pk sequencial, essencial para garantir a unicidade de cada registro no banco de dados.
- Tratamento de Nulos: Valores ausentes nas colunas críticas (director, cast, country, rating) foram substituídos pela string 'Unknown' ou 'Unrated' para evitar valores vazios na base de dados.
- Engenharia de Features (duration): A coluna duration (que misturava valores como "102 min", "5 Seasons") foi dividida em duas colunas

numéricas e categóricas: duration_value (o número) e duration_unit (min, Season, Seasons), facilitando cálculos e filtros (como a média de runtime).

- Conversão de Tipos: A coluna date_added foi convertida para o formato datetime e duration_value para float.
- Separação por Tipo: O dataset foi logicamente separado em dois DataFrames auxiliares: df_disney_movie_sql e df_disney_serie_sql, embora o carregamento final tenha usado o dataset completo e limpo.

2.3. Banco de Dados Utilizado

No projeto utilizei o SQLite para o carregamento dos dados no SQLite por ser um banco de dados serverless e leve, ideal para a prototipagem e análise local de dados estruturados e tabelares. A tabela final criada foi denominada conteudo_disney.

3. Consultas SQL (Análise Exploratória)

As consultas foram projetadas para extrair insights específicos sobre o desempenho do catálogo Disney.

Consulta	Descrição do que Faz	Motivação e Insight Esperado
Consulta 1	**Filmes Longa-metragem com Diretores Populares:** Identifica diretores com pelo menos 5 filmes no catálogo e cuja duração média dos filmes excede 90 minutos. Em seguida, lista os 10 filmes mais longos desses diretores.	Avalia quais diretores a Disney confia para produzir consistentemente filmes mais longos (e presumivelmente de maior orçamento/popularidade), revelando um padrão de investimento em talentos.
Consulta 2	**Tendência de Lançamento por Gênero (Pós-2010):** Calcula o total de títulos lançados por ano (após 2010) e o percentual específico de títulos classificados como 'Animation'.	Analisa a mudança na prioridade de conteúdo da Disney após a última década (época de crescimento do *streaming*). Espera-se que o percentual de animação e séries tenha crescido.
Consulta 3	**Duração Média de Filmes por Classificação Indicativa (Pós-2010):** Calcula a duração média, mínima e máxima de filmes, agrupados pela classificação indicativa (rating).	Investiga a relação entre o *runtime* e a audiência. Filmes com classificações mais maduras (ex: R) tendem a ser mais longos, ou o público infantil (ex: G) tem maior tendência a filmes curtos?
Consulta 4	**Programas de TV com MAIOR Número de Temporadas:** Lista os 5 *TV Shows* com o maior valor na coluna duration_value (o maior número de temporadas).	Identifica os produtos de maior longevidade no catálogo, que são cruciais para reter a audiência e garantir o valor da assinatura.

Consulta	Descrição do que Faz	Motivação e Insight Esperado
Consulta 5	**Programas de TV com MENOR Número de Temporadas:** Lista 5 *TV Shows* com o número mínimo de temporadas (geralmente 1).	Identifica produções mais recentes que podem ser minisséries ou que não foram renovadas, ajudando a entender o ciclo de vida e a taxa de sucesso de novos programas.