Catálogo de Dados

1) gold.dim_title

Descrição

Dimensão que representa o **título** (entidade de conteúdo) do catálogo Disney+. Contém os atributos mais estáveis do conteúdo: identificador natural (show_id), nome do título, ano de lançamento, duração e descrição. É a dimensão central para descrição de "o que é" o item de catálogo.

Campos e domínio

- **title_id** (BIGINT, PK)
  Descrição: chave substituta (surrogate key) do título.
  Domínio esperado: inteiro positivo, >= 1. Não nulo. Único.
- **show_id** (STRING, UK)
  Descrição: identificador natural do dataset (ex.: "s1234").
  Domínio esperado: string não vazia. Único.
  Regra: trim(show_id) <> ''.
- **title** (STRING)
  Descrição: nome do título.
  Domínio esperado: string não vazia.
  Regra: trim(title) <> ''.
- **release_year** (INT)
  Descrição: ano de lançamento do título.
  Domínio esperado: 1900 <= release_year <= ano_atual (ou <= ano de extração do dataset).
  Regras: não nulo para análises de lag; valores fora do intervalo indicam erro de origem.
- **duration** (STRING)
  Descrição: duração textual (ex.: "90 min" para filmes; "1 Season" para séries).
  Domínio esperado: string; pode ser nulo para as análises deste MVP.
  Validação (opcional): regex/parse para padrões "min" ou "Season(s)".
- **description** (STRING)
  Descrição: sinopse do conteúdo.
  Domínio esperado: texto livre; pode ser nulo para as análises desde MVP.

2) gold.dim_date

Descrição

Dimensão de datas (calendário) para suportar análises temporais por dia/mês/ano. Neste MVP utilizada principalmente para **data de adição ao catálogo**.

Campos e domínio

- **date_id** (INT, PK)
  Descrição: chave de data no formato yyyymmdd.
  Domínio esperado: >= 19000101 e <= 21001231 (faixa ampla). Não nulo. Único.
  Regras: deve corresponder a full_date_added.
- **full_date_added** (DATE)
  Descrição: data completa.
  Domínio esperado: data válida. Não nula.
- **year** (INT)
  Descrição: ano extraído de full_date_added.
  Domínio esperado: 1900 <= year <= ano_atual (ou <= ano de extração).
- **month** (INT)
  Descrição: mês extraído.
  Domínio esperado: 1..12.
- **day** (INT)
  Descrição: dia do mês extraído.
  Domínio esperado: 1..31 (consistência real depende do mês/ano).

3) gold.dim_type

Descrição

Dimensão para classificar o tipo de conteúdo (filme ou série). Serve para filtros e segmentação de análises.

Campos e domínio

- **type_id** (INT, PK)
  Descrição: surrogate key do tipo.
  Domínio esperado: inteiro positivo >= 1. Único.

- **type_name** (STRING)
  Descrição: categoria textual do tipo.
  Domínio esperado (dataset Disney+): tipicamente {'Movie', 'TV Show'}.
  Regra: não nulo e não vazio.

4) gold.dim_rating

Descrição

Dimensão de classificação indicativa (rating). Usada para análise de distribuição do catálogo por faixa/classificação.

Campos e domínio

- **rating_id** (INT, PK)
  Descrição: surrogate key do rating.
  Domínio esperado: inteiro positivo >= 1. Único.
- **rating_name** (STRING)
  Descrição: rótulo da classificação indicativa (depende do país/standard do dataset).
  Domínio esperado: conjunto finito de strings específicas (categóricas) - (“G”, “PG”, “PG-13”, “R”, “NC-17”, “TV-Y”, “TV-Y7”, “TV-G”, “TV-PG”, “TV-14”, “TV-MA” ou “Not Rated”). Esta lista válida de domínio esperado é definida pelos órgãos de classificação etária, neste dataset especificamente, temos a Motion Pictures Association (MAPA) que faz a classificação para os filmes (Homepage - MPA Film Ratings) e a TV Parental Guidelines que faz a classificação do conteúdo de séries para a televisão.
  Regra: não nulo e não vazio.
  Observação: os valores exatos dependem do dataset (podem existir variações e valores “Not Rated”).

5) gold.dim_genre

Descrição

Dimensão de gêneros (categorias) do conteúdo, originada do campo multivalorado (listed_in) normalizado na Silver.

Campos e domínio

- **genre_id** (INT, PK)
  Descrição: surrogate key do gênero.
  Domínio esperado: inteiro positivo >= 1. Único.

- **genre_name** (STRING)
  Descrição: nome do gênero/categoria (ex.: “Animation”, “Comedy”, “Family”, etc.).
  Domínio esperado: string não vazia; conjunto finito (variável conforme dataset).
  Regra: trim(genre_name) <> ''.

6) gold.bridge_title_genre

Descrição

Tabela de relacionamento **N:N** entre títulos e gêneros. Um título pode ter vários gêneros e um gênero pode estar em muitos títulos. Esta tabela viabiliza contagens por gênero, diversidade por período e análises de distribuição.

Campos e domínio

- **title_id** (BIGINT, FK → dim_title.title_id)
  Descrição: referência ao título.
  Domínio esperado: >= 1, não nulo; deve existir na dim_title.
- **genre_id** (INT, FK → dim_genre.genre_id)
  Descrição: referência ao gênero.
  Domínio esperado: >= 1, não nulo; deve existir na dim_genre.

Regras de qualidade

- **Unicidade composta:** (title_id, genre_id) deve ser único (sem duplicatas).
- Sem “órfãos”: todos os FKs devem existir nas dimensões.

---

7) gold.fact_catalog_addition

Descrição

Tabela fato que representa o **evento de adição do título ao catálogo** (factless com métricas derivadas). Grão esperado: **1 linha por título adicionado** (por title_id e data). Suporta análises de tempo de entrada (lag), evolução temporal e métricas agregadas.

Campos e domínio

- **fact_id** (BIGINT, PK)
  Descrição: surrogate key da linha fato.
  Domínio esperado: inteiro positivo >= 1. Único. Não nulo.

- **title_id** (BIGINT, FK → dim_title.title_id)
  Descrição: título adicionado.
  Domínio esperado: >= 1, não nulo; deve existir na dimensão.

- **date_id** (INT, FK → dim_date.date_id)
  Descrição: chave da data de adição.
  Domínio esperado: yyyymmdd, não nulo (se a data existir); deve existir na dimensão.

- **type_id** (INT, FK → dim_type.type_id)
  Descrição: tipo do conteúdo no momento da adição.
  Domínio esperado: >= 1, pode ser nulo apenas se não mapeado (não recomendado).

- **rating_id** (INT, FK → dim_rating.rating_id)
  Descrição: classificação indicativa do conteúdo.
  Domínio esperado: >= 1, pode ser nulo se não mapeado (não recomendado).

- **full_date_added** (DATE)
  Descrição: data completa de adição ao catálogo (campo operacional para filtros).
  Domínio esperado: data válida; idealmente não nula para análises temporais.

- **release_date_approx** (DATE)
  Descrição: data aproximada de lançamento (ex.: YYYY-01-01) construída a partir de release_year, usada para calcular lag_days.
  Domínio esperado: data válida; pode ser nula se release_year for nulo.

- **lag_days** (INT)
  Descrição: diferença em dias entre full_date_added e release_date_approx.
  Domínio esperado: >= 0 na maioria dos casos.
  Regras: valores negativos indicam cenário incomum (release_year maior que ano de adição ou erro de origem). Um cenário possível seria um filme ser adicionado à plataforma antes de ser lançado no cinema, caso a base esteja considerando data de lançamento este lançamento no cinema.
  Faixa típica: de 0 até 50.000 (considerando que o primeiro filme do cinema teria sido lançado em 1895-“A Saída dos Trabalhadores da Fábrica Lumière”).

- **title_count** (INT)
  Descrição: métrica contadora (sempre 1) para facilitar agregações.
  Domínio esperado: exatamente 1.

Regras de qualidade

- Sem nulos em: fact_id, title_id, full_date_added e lag_days (para análises de lag).
- title_count = 1.
- Integridade referencial com dimensões.
- Checagem de duplicidade do grão (recomendado): (title_id, date_id) deve ser único se a regra for “um título entra uma vez”.