Sumário

Github

Link com os processos realizados nos Notebooks criados para cada etapa do MVP.

Github pessoal (Thomas Abrantes):
https://github.com/hannemanbr/PUC-RIO-MVP/tree/main

Google Drive Pessoa com acesso público com mesmo contéúdo do GitHub:
https://drive.google.com/drive/folders/1zZjQfyDJLOGY6TKNUHqNqmVgkAdTChjK?usp=drive_link

Notebooks criados no Databricks

Link dos notebooks com cada processo da MVP no GitHub.

Camada Bronze

https://github.com/hannemanbr/PUC-RIO-MVP/blob/main/Thomas-Notebook-PUC-RIO%20-%20Camada%20Bronze.ipynb

Camada Silver

https://github.com/hannemanbr/PUC-RIO-MVP/blob/main/Thomas-Notebook-PUC-RIO%20-%20Camada%20Silver.ipynb

Camada Gold

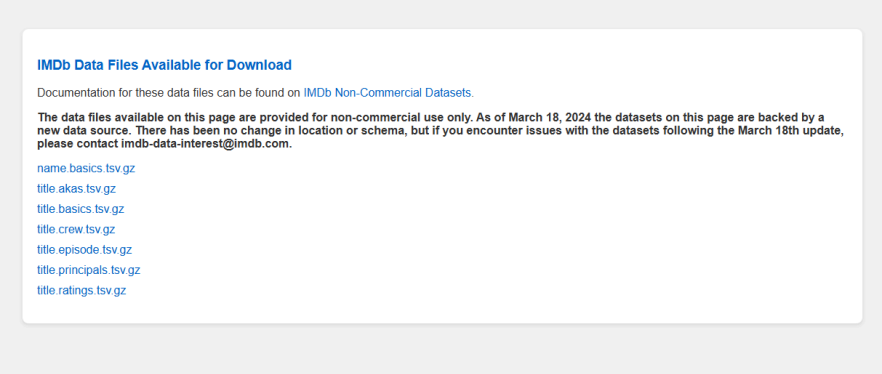
https://github.com/hannemanbr/PUC-RIO-MVP/blob/main/Thomas-Notebook-PUC-RIO%20-%20Camada%20Gold.ipynb

Relatórios

https://github.com/hannemanbr/PUC-RIO-MVP/blob/main/Thomas-Notebook-PUC-RIO%20-%20Relatorios.ipynb

Coleta

O IMDb (*Internet Movie Database*) é uma Base de Dados de Filmes na Internet, contendo informações de séries, documentários, videogames e outras mídias. O IMDb disponibiliza os datasets públicos no link https://datasets.imdbws.com/, o qual foi utilizado neste trabalho.

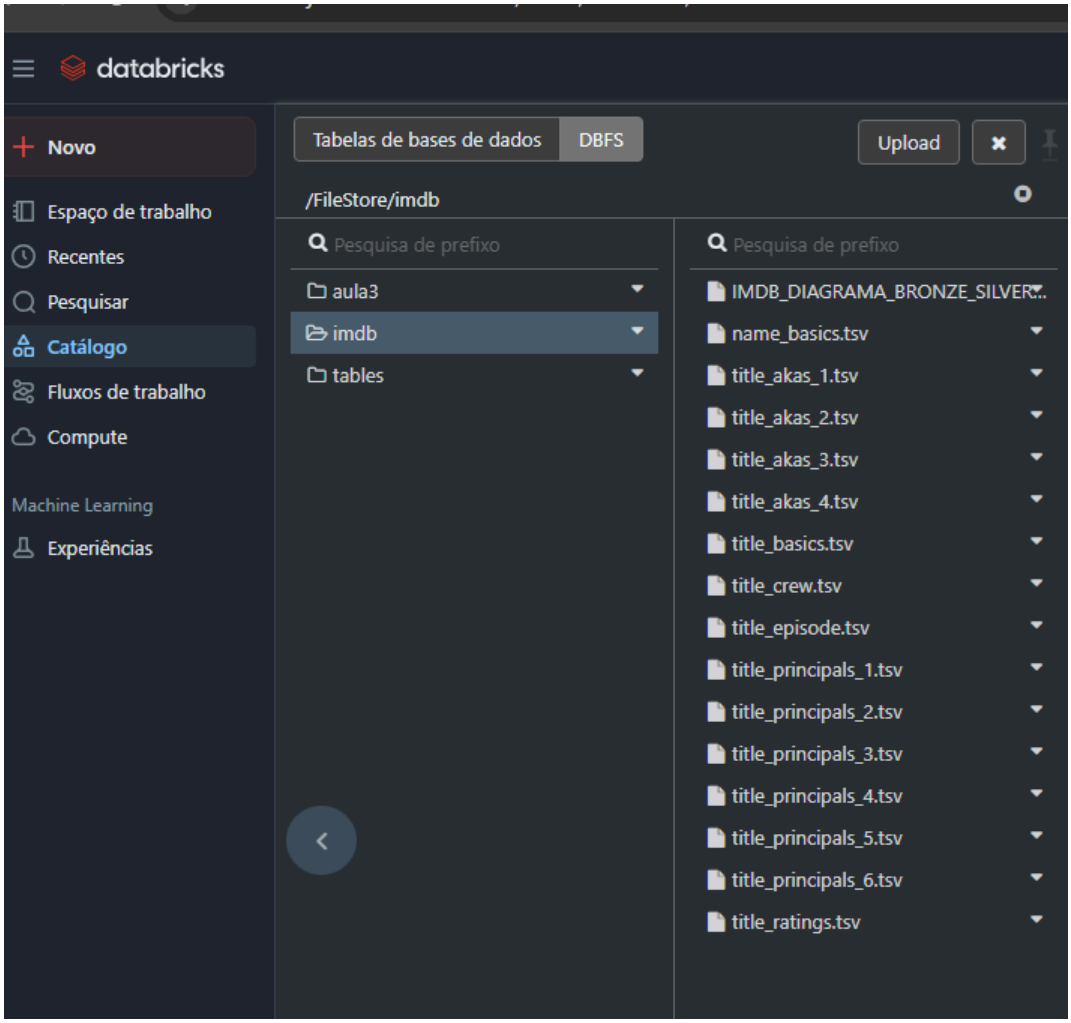


Página com os arquivos utilizados na coleta de dados em https://datasets.imdbws.com

Ao acessar essa página podemos fazer download dos arquivos no formato TSV e compactado no padrão GZIP (formato Gnu Zip).

- title.principals.tsv.gz
- title.basics.tsv.gz
- name.basics.tsv.gz
- title.akas.tsv.gz
- title.crew.tsv.gz
- title.episode.tsv
- title.ratings.tsv.gz

Após a descompactação dos arquivos estes foram enviados para o Catálogo no Databricks em uma pasta IMDB por meio do DBFS no Databricks.



Catalogo no Databricks após upload dos arquivos TSV

Objetivo

Nesse MVP optei por analisar os dados públicos disponibilizados pelo IMDb (*Internet Movie Database*) por se tratar de um conteúdo que tenho familiaridade.

Através da análise dos arquivos e da documentação disponivel no link https://developer.imdb.com/non-commercial-datasets/, procurei responder às seguintes perguntas:

- Quantidade de títulos lançados no Brasil por categoria e década (80, 90, etc.)?
- Qual categoria teve mais lançamentos no Brasil por década?

- Todos os títulos que possuem nome exclusivo no Brasil (com nome diferente do nome original ou comercial).
- Qual a porcentagem de pessoas que atuaram como Ator ou Atriz em Títulos do tipo Filmes, Curtas ou Vídeo, lançados no Brasil?
- Porcentagem de atores do sexo feminino e masculino que atuaram nos títulos do tipo Filme em cada década existente na base.
- Filmes lançados no Brasil, onde o Diretor(a) atuou como Ator/Atriz.
- Ranking dos diretores com títulos lançados no Brasil.
- Filmes brasileiros distribuídos no exterior.

Dicionário de dados (Camada Bronze / Silver)

Conforme pude análise nos arquivos, a primeira linha em cada arquivo contém os cabeçalhos (header) que descrevem o tipo de informação em cada coluna.
Todo valor '\N' (não aplicável) que possa ser encontrado em cada registro é uma indicação que um campo específico está faltando informação ou é nulo.

A seguir as tabelas criadas na camada bronze (e na camada silver com mesmo nome) e seus respectivos arquivos de origem.

TB_TITLE_AKAS (TITLE.AKAS.TSV)

Observação: Devido ao seu tamanho, o arquivo foi particionado em 4 partes.
Então, os arquivos foram carregados em um Dataframe específico e posteriormente, unificado em um único dataframe para a carga dos dados e criação da tabela.

Arquivo contendo os títulos AKAs, que são Títulos Alternativos.
São os nomes diferentes que um filme pode ter, em diferentes idiomas e países.
O arquivo consiste nas colunas:

- **titleId** (string): Identificador no formato alfanumérico exclusivo do título.
- **ordering** (inteiro) – É um número para identificar exclusivamente as linhas para um determinado **titleId**. Neste arquivo há mais de uma linha com mesmo titleId.
- **title** (string) – Nome do título.

- **region** (string) - Descreve a região, com dois caracteres.
  Exemplo: US para United States, JP para Japão na versão do título.
- **language** (string) - o idioma do título.
- **types** (string) -Conjunto enumerado em forma de string de atributos para este título alternativo.
  Na documentação é indicado como matriz mas *na importação esses valores vão como uma string contendo todos os valores da coluna e caso seja usado será tratado como uma lista.*
  De acordo com a documentação pode conter os valores abaixo mas **destaca que** pode haver novos valores podem ser adicionados no futuro::
  - alternative
  - dvd
  - festival
  - tv
  - video
  - working
  - original
  - imdbDisplay
- **attributes** (string) - Termos adicionais para descrever este título alternativo.
  Na documentação é indicado como matriz mas *na importação esses valores vão como uma string contendo todos os valores da coluna e caso seja usado será tratado como uma lista.*
- **isOriginalTitle** (inteiro) – O Valor 0 (false) indica que o título não original e 1 (true) que o título é original.

TB_TITLE_BASICS (TITLE.BASICS.TSV)

Arquivo contendo as informações básicas dos títulos.
O arquivo consiste nas colunas:

- **tconst** (string) - Identificador no formato alfanumérico exclusivo do título.
- **titleType** (string) – Informa o tipo/formato do título.
  Exemplo: filme, curta, série de TV, episódio de TV, vídeo, etc.
  No arquivo esses valores estão registrados em inglês como movie, short, tvseries. tvepisode, video.
- **primaryTitle** (string) – Nome do título popularmente conhecido ou usado pelos cineastas em materiais promocionais no seu lançamento.
- **originalTitle** (string) - Titulo original, no idioma original.

- **isAdult** (boolean) - O valor 0(false) informa que o título não é somente para o público adulto e 1(true) informa que título é destinado ao público adulto.
- startYear (inteiro com 4 caracteres) - Ano de lançamento do título. Caso seja uma série de TV o valor é o ano da primeira temporada da ´serie.
- **endYear** (inteiro com 4 caracteres) - Ano de término para série de TV. Para outros títulos esse valor vem preenchido com “\N”.
- **runtimeMinutes** - Tempo, em minutos, de execução do título principal.
- **genres** (matriz de strings) – Pode conter até três gêneros associados ao título. Esse campos quando houver mais de uma gênero, sera exibido separado por uma vírgula.
  Exemplo: Documentary,Short.

TB_TITLE_CREW (TITLE.CREW.TSV)

Arquivo que contém informações da equipe de filmagem de um filme ou programa de TV. O arquivo consiste nas colunas:

- **tconst** (string) - Identificador no formato alfanumérico exclusivo do título.
- **directors** (matriz de **nconsts** no formato string) - Identificador no formato alfanumérico identificando o(s) diretore(s) do título.
- **writers** (matriz de **nconsts** no formato string) - Identificador no formato alfanumérico identificando o(s) escritore(s) do título.

TB_TITLE_EPISODE.TSV (TITLE.EPISODE.TSV)

Arquivo que contém informações de um episódio de uma série de um programa de TV. O arquivo consiste nas colunas:

- **tconst** (string) - identificador alfanumérico do episódio
- **parentTconst** (string) - identificador alfanumérico da série de TV de Origem (pai).
- **seasonNumber** (inteiro) – Número da temporada do episódio.
- **episodeNumber** (inteiro) – Número do episódio com valor **tconst** em uma série de TV.

TITLE.PRINCIPALS.TSV

Observação: Devido ao seu tamanho, o arquivo foi particionado em 6 partes.
Então, os arquivos foram carregados em um Dataframe específico e posteriormente, unificado em um único dataframe para a carga dos dados e criação da tabela.

Arquivo que contém informações do título original de uma obra, na sua língua original.
O arquivo consiste nas colunas:

- **tconst** (string) - Identificador alfanumérico exclusivo do título
- **ordering** (inteiro) – É um número para identificar exclusivamente as linhas para um determinado **titleId**.
- **nconst** (string) - Identificador alfanumérico exclusivo do nome/pessoa
- **category** (string) - Categoria do trabalho em que a pessoa executou na produção.
- **job** (string) - Cargo específico, se aplicável, caso contrário '\N'
- **characters** (string) - Nome do personagem interpretado, se aplicável, caso contrário '\N'

TB_TITLE)RATINGS (TITLE.RATINGS.TSV)

Arquivo que contém informações sobre as classificações e o número de votos para cada título no site. O arquivo consiste nas colunas:

- **tconst** (string) - Identificador alfanumérico exclusivo do título
- **averageRating** – Valor da média ponderada de todas as avaliações individuais dos usuários.
- **numVotes** (inteiro) - Número de votos que o título recebeu.

TB_NAME.BASICS.TSV (NAME.BASICS.TSV)

Arquivo que contém informações básicas de uma pessoa.
O arquivo consiste nas colunas:

- **nconst** (string) - identificador alfanumérico exclusivo do nome/pessoa
- **primaryName** (string) – nome pelo qual a pessoa é mais frequentemente creditada

- **birthYear** (inteiro) – Ano de nascimento no formato AAAA (4 dígitos).
- **deathYear** (inteiro) – Ano de falecimento no formato AAAA (4 dígitos). se aplicável, caso contrário '\N'.
- **primaryProfession** (matriz de strings) – Valor das três principais profissões da pessoa.
- **knownForTitles** (matriz de **tconsts**) – Valores dos títulos pelos quais a pessoa é conhecida.

Camada BRONZE

Este processo consiste em:

- Obter os dados brutos dos sete arquivos disponiveis em https://datasets.imdbws.com/
- Importação dos dados sem alteração, na forma como estão dispostos nos arquivos de origem.
- Carga dos dados dos arquivos e criação das tabelas no banco de dados (*schema*) IMDB_DB_BRONZE.
- Garantir a qualidade dos dados validando o volume de dados importados de acordo com a quantidade de registros dos arquivos.

Banco de Dados: IMDB_DB_BRONZE

Observação: Todas as colunas das tabelas criadas com a importação dos dados dos arquivos, são do tipo string.

Com os arquivos disponíveis no datasets do IMDB, podemos realizar a carga para o banco de dados (*schema*) IMDB_DB_BRONZE, importando os dados dos arquivos para tabelas com nomes contendo prefixo "tb_" e nome do arquivo de origem nos seus respectivos DataFrames.

Processo de importação

- Upload dos arquivos para a pasta IMDB pelo DBFS no Databricks
- Leitura e conversão em dataframe de cada arquivo TSV com o comando **spark.read.csv('CAMINHO DO ARQUIVO NO DBFS', sep='\t', header=True)** da biblioteca pyspark.sql.session.sparksession.

  O parâmetro **header=true** foi utilizado para que o PySpark considere a primeira linha como header do arquivo (representando as colunas)

- Para cada Dataframe, criação da tabela com comando **SQL** com nome das colunas iguais ao disponível na primeira linha de cada arquivo (header).

- Inserção dos dados na tabela criada com o comando **df_title_basics.write.format("delta").mode("overwrite").saveAsTable([schema].[NOME DA TABELA])**

  O parâmetro “*overwrite*” indica que sempre que ele for executado vai sobrescrever os dados na tabela de destino.

Após a importação termos o seguinte diagrama da base IMDB_DB_BRONZE:

:

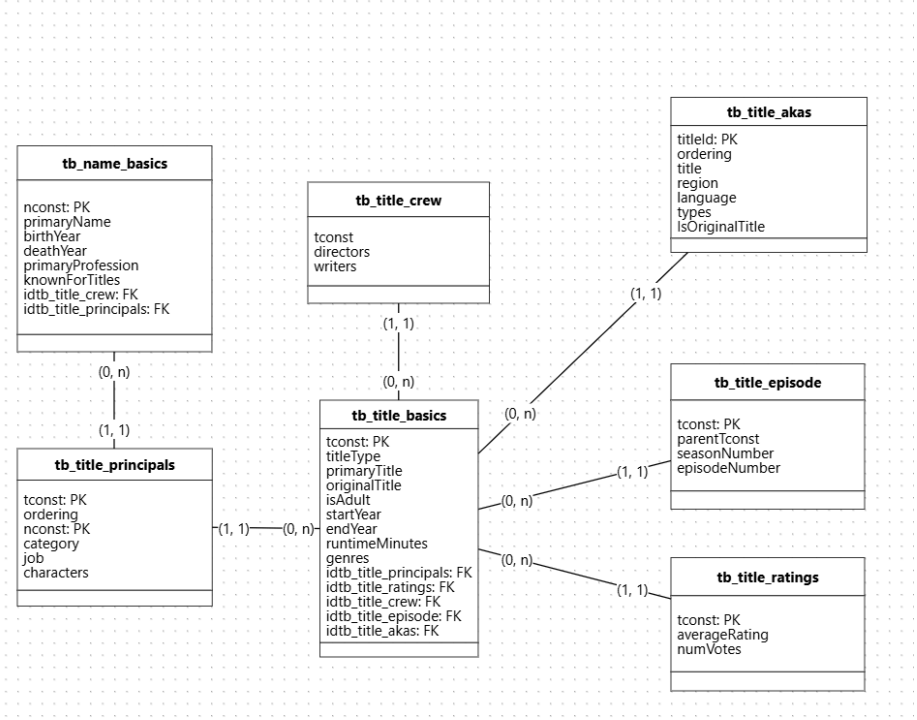


Diagrama de Entidade e Relacionamento da base de dados Bronze e Silver

Observação:

- Todas as informações das tabelas e entidades desta base de dados estão na guia “Dicionário de Dados”.

Camada SILVER

Banco de Dados: IMDB_DB_SILVER

Neste processo, ocorre a importação dos dados existentes nas tabelas do *schema* IMDB_DB_BRONZE, mantendo o mesmo nome das tabelas no *schema* IMDB_DB_SILVER.

Durante a importação ocorre a limpeza dos dados conforme os seguintes passos:

- Criação do banco de dados (*schema*) IMDB_DB_SILVER
- Criação de todas as tabelas, com mesmo nome e nome das colunas, na IMDB_DB_BRONZE (camada bronze).
- As tabelas criadas na camada Silver, tabelas seguindo o tipo das colunas de acordo com a documentação.
- Remover a valor da legenda "\N" em algumas colunas que indica que o valor “não é aplicável” e dados nulos ou inadequados..
- Tratamentos específicos, se for necessário, nas tabelas (maiores detalhes estão especificados no comentário do Notebook). Verificação na qualidade dos dados em busca de chaves duplicadas ou duplicidade, valores indevidos e nulos.
- Conferência na quantidade de dados importados
    - Quando não há condicional que remova registros, verificar se mantém a mesma quantidade de registros na camada SILVER.
    - Quando há condicional que remova registros, verificar se mantém a quantidade igual ou menor de registros na camada SILVER.

Observação:

- Todas as informações das tabelas e entidades desta base de dados estão na guia “Dicionário de Dados”.
- Na versão Community do databricks, ele não permite a criação de chave primária e estrangeira, então para garantir a qualidade dos dados, na camada SILVER os campos CHAVE da tabela são do tipo NOT NULL evitando a carga de valor nulo.

Após a importação termos o seguinte diagrama da base IMDB_DB_SILVER:

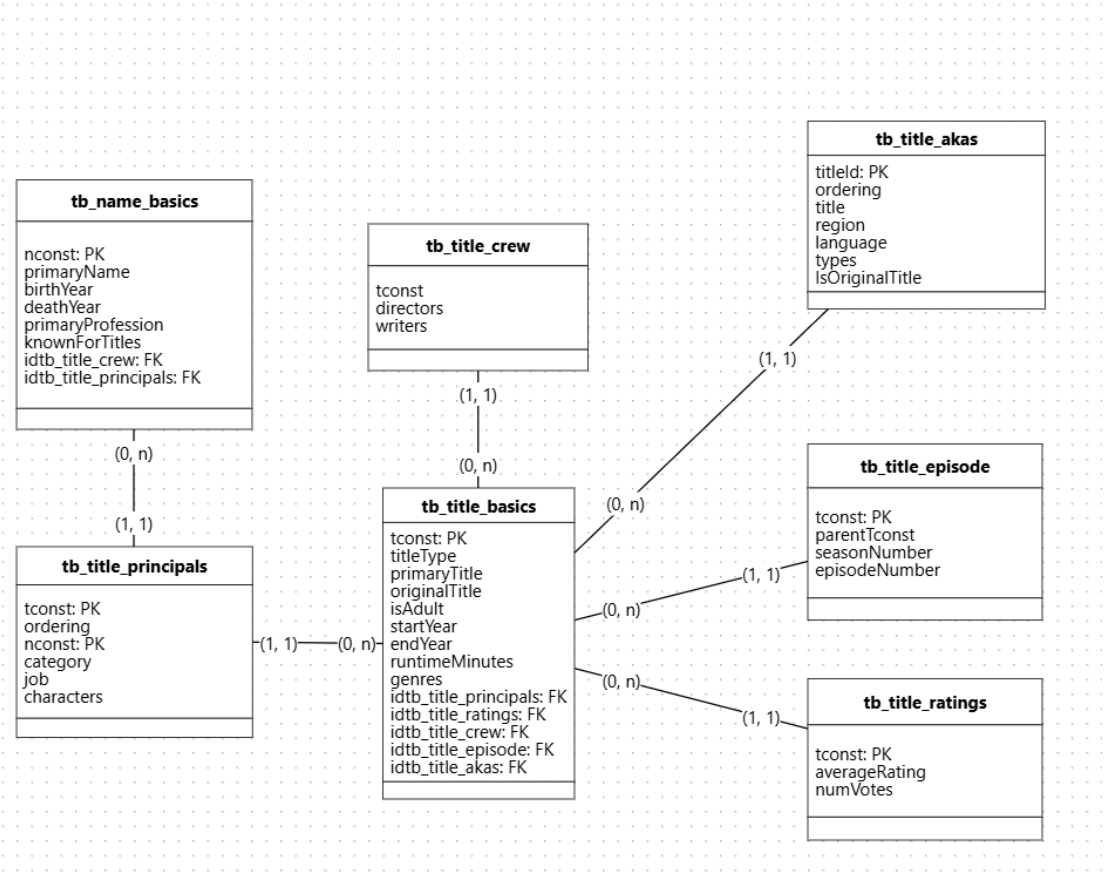


Diagrama de Entidade e Relacionamento da base de dados Bronze e Silver

Camada Gold

Banco de Dados: IMDB_DB_SILVER

Todos os dados contidos na camada Gold são de origem da camada Silver.

Na versão Community do databricks, ele não permite a criação de chave primária e estrangeira, então para garantir a qualidade dos dados, na camada GOLD os campos CHAVE da tabela são do tipo NOT NULL evitando a carga de valor nulo. A camada GOLD consiste em:

- Tabelas e suas colunas com nomes mais amigáveis.
- Tabelas com origem dos relacionamentos de tabelas na camada SILVER, que atendem ao resultado que será consultado no Notebook Relatórios.
- Tabelas com origem dos relacionamentos de tabelas na camada GOLD, que atendem ao resultado que será consultado no Notebook Relatórios.

No diagrama abaixo podemos ver as origens de cada tabela criada na Camada Gold.

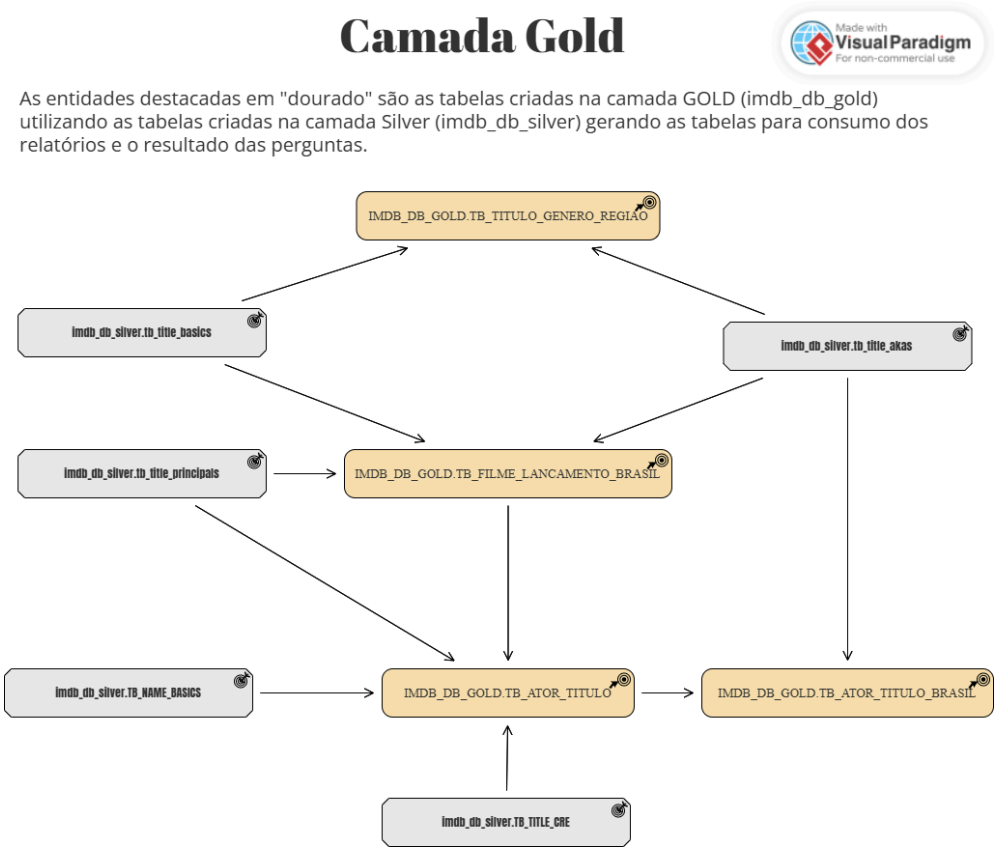


Diagrama das Entidades criadas na camada Gold e as origens na camada Silver

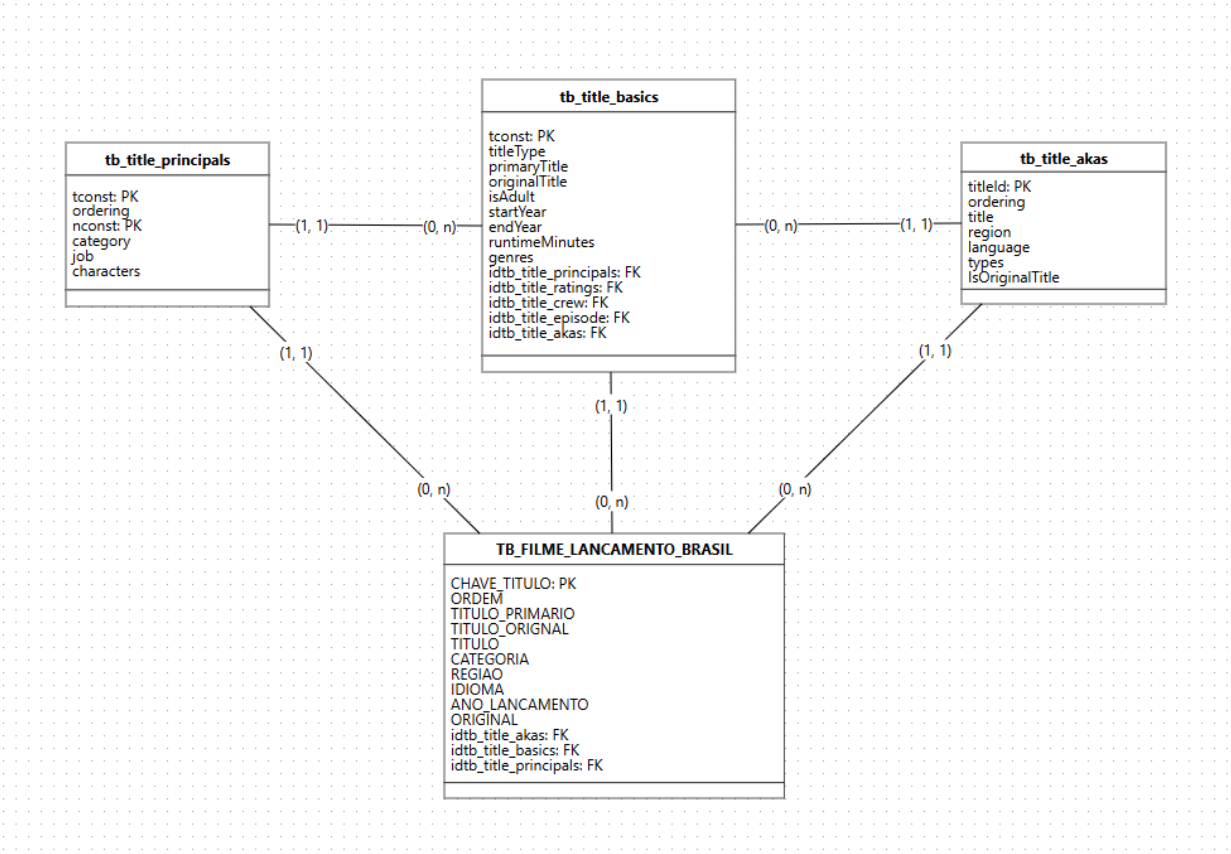
Dicionário de Dados (Camada Gold)

TB_TITULO_LANCAMENTO_BRASIL

Tabela com todos os título que sejam do tipo Movies, Short e Vídeo com lançamento no Brasil contendo as colunas:

- **CHAVE_TITULO** (CHAR(10) NOT NULL, dados da coluna tconst da tabela imdb_db_silver.tb_title_basics): Identificador exclusivo do título.

- **ORDEM** (INT NOT NULL, dados da coluna ordering da tabela imdb_db_silver.tb_title_akas): Número para identificar exclusivamente as linhas para um determinado titleId na tabela imdb_db_silver.tb_title_akas.
- **TITULO_PRIMARIO** (VARCHAR(1000), dados da coluna primaryTitle da tabela imdb_db_silver.tb_title_basics): Descrição do título popularmente conhecido ou usado pelos cineastas em materiais promocionais no seu lançamento.
- **TITULO_ORIGINAL** (VARCHAR(1000), dados da coluna primaryTitle da tabela imdb_db_silver.tb_title_basics): Descrição do título original, no idioma original.
- **TITULO** (VARCHAR(1000), dados da coluna title na tabela imdb_db_silver.tb_title_akas): Descrição do título no Brasil.
- **CATEGORIA** (VARCHAR(200), dados da coluna titleType na tabela imdb_db_silver.tb_title_basics):
  Formato ou tipo do título, se é um filme (movie), curta (short) ou vídeo (video).
- **REGIAO** (CHAR(2) NOT NULL, dados da coluna region na tabela imdb_db_silver.tb_title_akas): Contendo sigla com 2 caracteres da região de lançamento do título
- **IDIOMA** (CHAR(2), dados da coluna language na tabela imdb_db_silver.tb_title_akas): Contendo sigla com dois caracteres do idioma.
- ANO_LANCAMENTO (INT, dados da coluna startYear na tabela imdb_db_silver.tb_title_basics)
- ORIGINAL (INT, dados da coluna isOriginalTitle na tabela imdb_db_silver.tb_title_akas): 1 indica que o título é original e 0 indica que o título não é original.



Relacionamento da tabela IMDB_DB_GOLD.TB_LANCAMENTO_BRASIL

TB_ATOR_TITULO

OBSERVAÇÃO

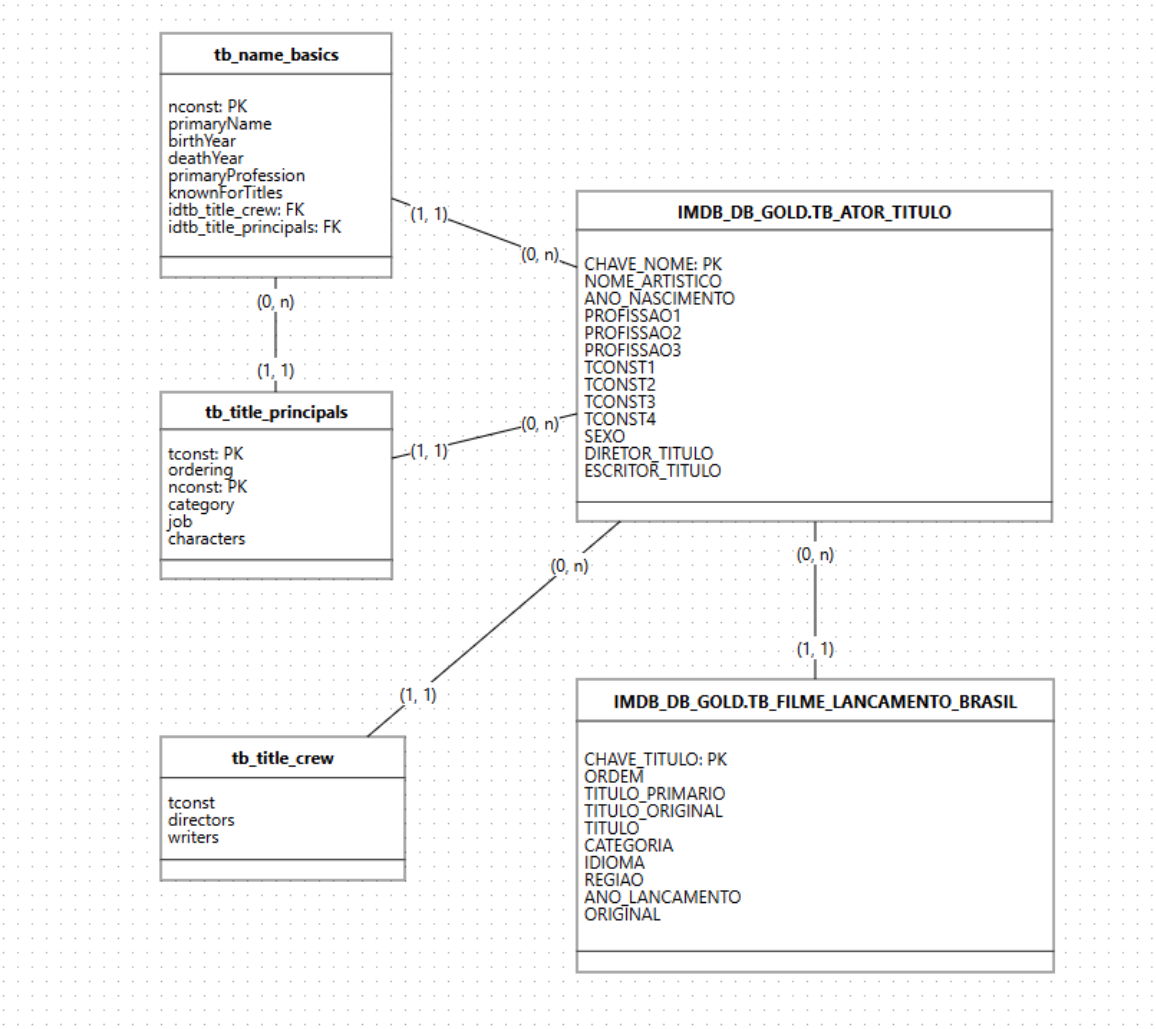
- Tabela criada com a finalidade de diminuir a carga para listagem de títulos ou profissões vinculadas aos atores, pois as colunas primaryProfession e knowForTitles é um array em formato string separado por vírgula.
- Desta forma a busca direta sem "separar" as informações dos arrays sobrecarrega a consulta sem necessidade.

Tabela com todos informações de pessoas que tenham participado de algum título, que sejam do tipo Movies, Short e Vídeo contendo as colunas:

- **CHAVE_NOME** (CHAR(10) NOT NULL, dados da coluna tconst da tabela imdb_db_silver.): Identificador exclusivo do título.
- **NOME_ARTISTICO** (VARCHAR(200) NOT NULL, dados da coluna primaryName da tabela imdb_db_silver.tb_name_basics): Descrição do título popularmente conhecido ou usado pelos cineastas em materiais promocionais no seu lançamento.
- **ANO_NASCIMENTO** (INT NOT NULL, dados da coluna birthYear da tabela imdb_db_silver.tb_name_basiscs):
  Ano de nascimento do Ator/Atriz.
- **PROFISSAO[NÚMERO]** (VARCHAR(50), dados da coluna primaryProfession da tabela imdb_db_silver.tb_name_basiscs): Convertida em 3 colunas, onde cada coluna contém a descrição de uma profissão que o ator exerce.
- **TCONST[NÚMERO]** (CHAR(10), dados da coluna knowForTitles da tabela imdb_db_silver.tb_name_basiscs): Convertida em 4 colunas, onde cada coluna possui a tconst (CHAVE_TITULO) de um título que o ator colaborou.
- **SEXO** (CHAR(1), preenchido de acordo com o CASE SQL da consulta que gera os dados): Informa com valor "M" se for Masculino, "F" se for Feminino ou “-” se não houver informação..
- **DIRETOR_TITULO** (CHAR(1) NOT NULL, preenchido de acordo com o CASE SQL dados da coluna directors da tabela imdb_db_silver.tb_title_crew): Informa com "S" se a pessoa foi um dos diretores no Título.
- **ESCRITOR_TITULO** (CHAR(1) NOT NULL, preenchido de acordo com o CASE SQL dados da coluna writters da tabela imdb_db_silver.tb_title_crew): Informa com "S" se a pessoa foi um dos escritores do Título.

OBSERVAÇÃO

- As colunas primaryProfession consistem em apenas 3 elementos.
- Foi realizada a consulta com a imdb_db_silver.tb_name_basics verificando a quantidade máxima de títulos por Ator/Atriz.
- Desta forma a lista foi separada em colunas com a finalidade de facilitar consultas com valor do campo "cheio" (completo).



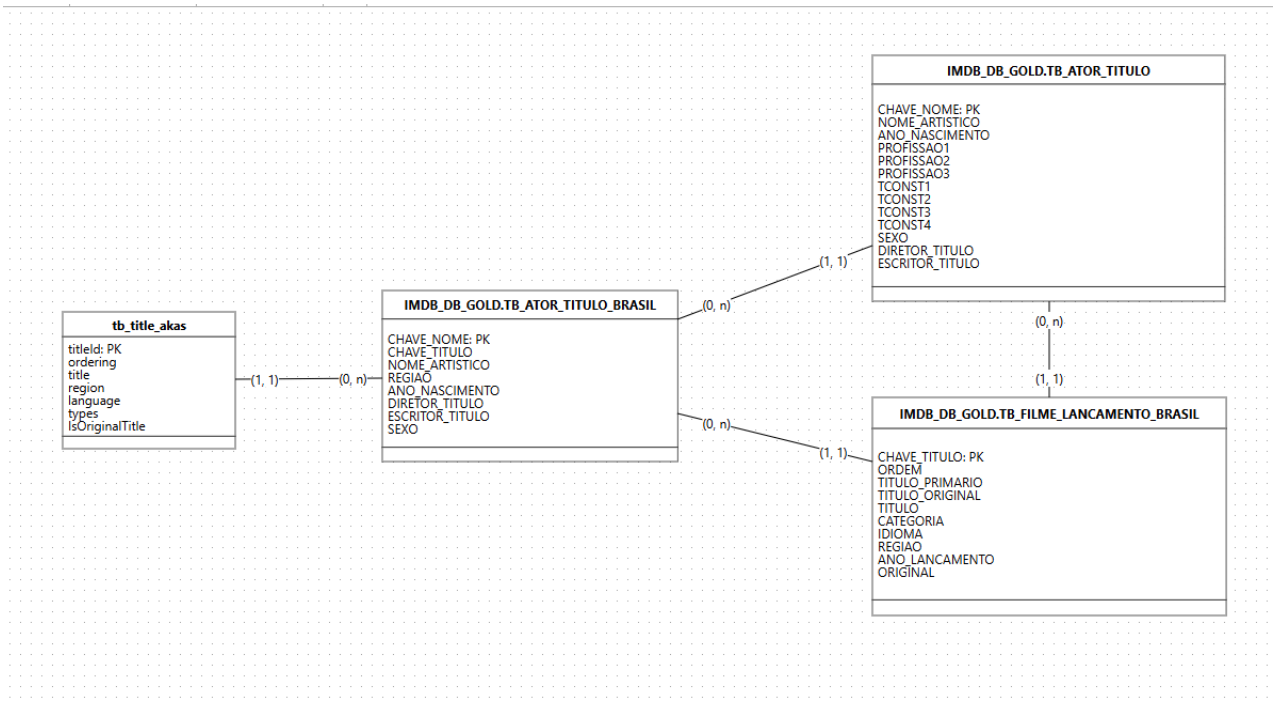
Relacionamento da tabela IMDB_DB_GOLD.TB_ATOR_TITULO

TB_ATOR_TITULO_BRASIL

Lista dos atores que atuaram em um título lançado no Brasil do tipo Movie, Short ou Vídeo, contendo as colunas:

- **CHAVE_NOME** (CHAR(10) NOT NULL, dados da coluna nconst da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Identificador exclusivo do ator/atriz.
- **CHAVE_TITULO** (CHAR(10) NOT NULL, coluna tconst da tabela IMDB_DB_GOLD.TB_FILME_LANCAMENTO_BRASIL): Identificador exclusivo do título que o ator/atriz participou.
- **NOME_ARTISTICO** (VARCHAR(20) NOT NULL, coluna NOME_ARTISTICO da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Nome artístico do ator/atriz.

- **REGIAO** (CHAR(2) NOT NULL, coluna region da tabela imdb_db_silver.tb_title_akas): Região (país) com dois caracteres em que o título foi lançado.
- **ANO_NASCIMENTO** (coluna ANO_NASCIMENTO da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Ano de nascimento do ator/atriz.
- **DIRETOR_TITULO** (coluna DIRETOR_TITULO da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Informa com "S" se a pessoa foi um dos diretores do Título.
- **ESCRITOR_TITULO** (coluna ESCRITOR_TITULO da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Informa com "S" se a pessoa foi um dos escritores do Título.
- **SEXO** (coluna SEXO da tabela IMDB_DB_GOLD.TB_ATOR_TITULO): Informa "F" para Feminino, "M" para Masculino ou "-" quando não há informação.



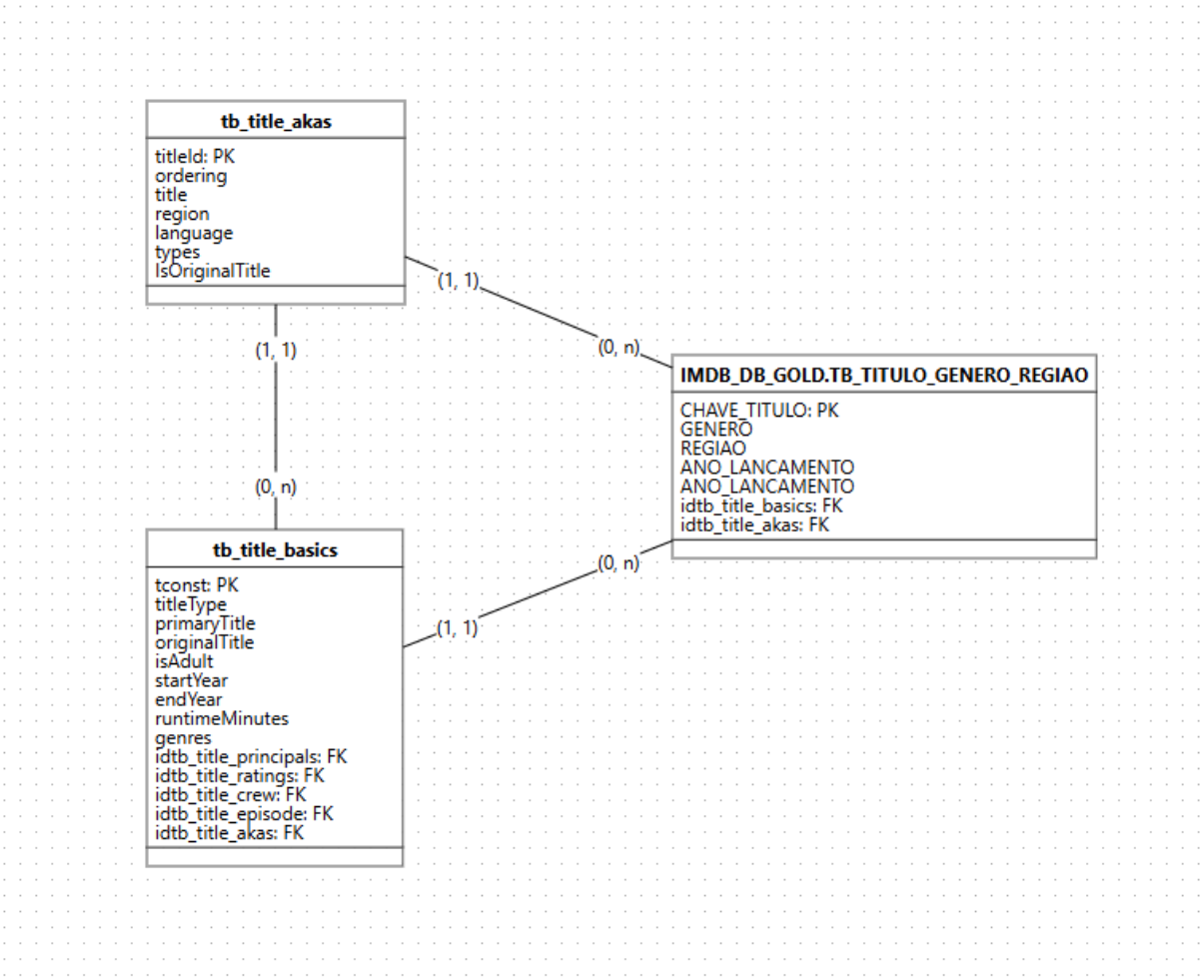
Relacionamento da tabela IMDB_DB_GOLD.TB_ATOR_TITULO_BRASIL

TB_TITULO_GENERO_REGIAO

Lista de todos os Gêneros de Títulos, disponíveis na base fornecida pelo IMDB por título e região de lançamento, contendo as seguintes colunas:

- **CHAVE_TITULO** (CHAR(10) NOT NULL, dados da coluna tconst da tabela imdb_db_silver.tb_title_basics): Identificador exclusivo do título.
- **GENERO** (VARCHAR(50) NOT NULL, dados da coluna genres da tabela imdb_db_silver.tb_title_basics): Informa o nome do gênero do título.

- **REGIAO** (VARCHAR(50) NOT NULL, dados da coluna region da tabela imdb_db_silver.tb_title_akas): Informa a região que o titulo foi lançado.
- **ANO_LANCAMENTO** (VARCHAR(50) NOT NULL, dados da coluna startYear da tabela imdb_db_silver.tb_title_basics): Informa ano de lançamento do título.



Relacionamento da tabela IMDB_DB_GOLD.TB_TITULO_GENERO_REGIAO

Análises e Resultados

Banco de Dados: IMDB_RELATORIO

Resposta às perguntas

Conforme mencionado no tópico objeto, com o tratamento dos dados na Camada Silver e na Camada Gold foi possível responder às perguntas.

A fonte de dados de todas as respostas são visões (views) materializadas em IMDB_RELATORIO com as tabelas da Camada Gold.

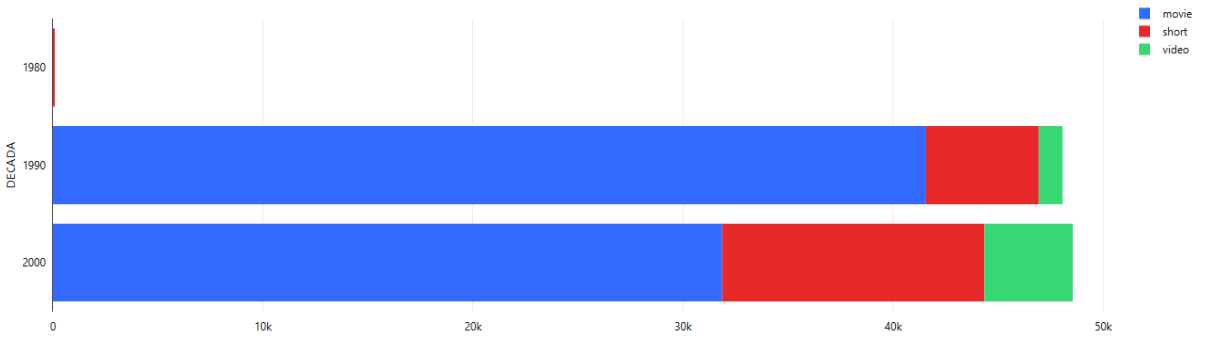
1 - Quantidade de títulos lançados no Brasil por categoria e década (80, 90, etc.)?

Todas as informações já estavam consolidadas na tabela tb_filme_lancamento_brasil, pois seus registros possuem a data de lançamento do título e já está “filtrada” em títulos do tipo Movie, Short e Vídeo.

Foi criada a view materializada para exibição do resultado foi necessário agrupar a coluna “ANO_LANCAMENTO” no formato (YYYY) por década, truncando a coluna com o valor do terceiro dígito, pois os dados históricos partem de 1914.

Resultado: view **VW_LANCAMENTO_POR_TIPO_DECADA**

DÉCADA	CATEGORIA	TÍTULOS LANÇADOS
1980	short	76
1990	movie	41553
1990	short	5383
1990	video	1127
2000	movie	31863
2000	short	12476
2000	video	4212



Resultado: view **VW_LANCAMENTO_POR_TIPO_DECADA**

2 - Qual categoria teve mais lançamentos no Brasil por década?

Com a uma view criada para a pergunta de Quantidade de títulos por década e categoria (pergunta 1, pude especializar a consulta em mais uma pergunta.

Criei a visão que faz a junção da view VW_LANCAMENTO_POR_TIPO_DECADA e VW_LANCAMENTO_POR_TIPO_DECADA no schema IMDB_RELATORIO, que permitiu o agrupamento de categorias e década de lançamento.

Resultado: view **VW_CATEGORIAS_COM_MAIS_LANCAMENTO_BRASIL**

DÉCADA	CATEGORIA COM MAIOR QUANTIDADE DE LANCAMENTO	QUANTIDADE
1980	Curta metragem	76
1990	Filme	41553
2000	Filme	31863

3 - Os últimos 10 títulos que possuem nome exclusivo no Brasil (com nome diferente do nome original ou comercial).

A resposta veio da consulta na tb_filme_lancamento_brasil que permitiu o agrupamento por ano de lançamento, até 2025, pois a base possui registros de lançamentos em 2026 que não é o foco da pergunta.

Resultado: view **VW_TITULOS_COM_NOME_EXCLUSIVO_NO_BRASIL (contendo todos os titulos até 2025)**

TÍTULO NO BRASIL	TÍTULO ORIGINAL	TÍTULO COMERCIAL	ANO LANÇAMENTO
366	UN AÑO Y UN DÍA	UN AÑO Y UN DÍA	2025
A ARMADILHA DO COELHO	RABBIT TRAP	RABBIT TRAP	2025
A BATALHA DO MAR CHINÊS	JIAO LONG XING DONG	OPERATION HADAL	2025
A CASA DO MICKEY: PARQUE DA MORTE	MICKEY'S SLAYHOUSE	MICKEY'S SLAYHOUSE	2025
A EMPREGADA	THE HOUSEMAID	THE HOUSEMAID	2025
A FAZENDA DOS ANIMAIS	ANIMAL FARM	ANIMAL FARM	2025
A FÚRIA DO LEÃO	LION FIST	LION FIST	2025
A GAROTA MAIS LINDA DO MUNDO	THE MOST BEAUTIFUL GIRL IN THE WORLD	THE MOST BEAUTIFUL GIRL IN THE WORLD	2025
A GUERRA DAS PALAVRAS	WORDS OF WAR	WORDS OF WAR	2025
A LISTA DO PERIGO: QUEM VAI ESCAPAR?	THE HIT-LIST	THE HIT-LIST	2025

4 - Qual a porcentagem de pessoas que atuaram como Ator ou Atriz em Títulos do tipo Filmes, Curtas ou Vídeo, lançados no Brasil?

Resultado: view **VW_TOTAL_POR_DECADA_SEXO_ATOR**

Neste resultado, optei por realizar em dois estágios.

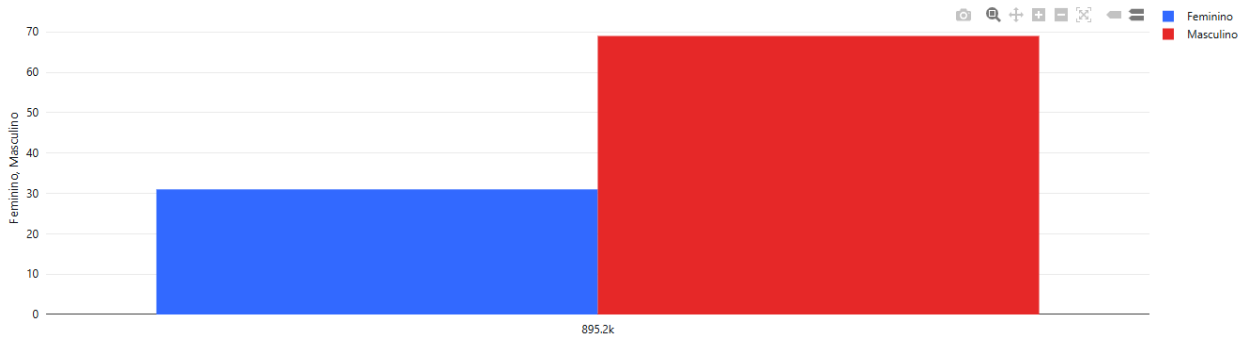
Criação de uma visão VW_TOTAL_POR_DECADA_SEXO_ATOR, consultando as tabelas TB_ATOR_TITULO_BRASIL e com a tabela TB_FILME_LANCAMENTO_BRASIL da camada Gold. Depois o resultado foi agrupado por SEXO (Masculino, Feminino e Não informado), década do lançamento do título e o Total de títulos em que participaram.

Com resultado da visão VW_TOTAL_POR_DECADA_SEXO_ATOR, criei uma visão que possui 4 *subqueries* na mesma view onde cada consulta contabiliza o total por SEXO e o total de todos os SEXOS, considerando que cada consulta seja uma tabela virtual

Observação: Como temos valores “Não informado” este resultado não pode ser considerado de boa qualidade, pois se analisarmos o resultado há uma porcentagem grande (33%) para a decada de 90 e 2000, mesmo que o resultado esperado seja próximo a realidade, a quantidade de Homens nas produções costuma ser maior.

Com o resultado das consultas foi possível realizar a junção e calcular a porcentagem.

TOTAL	FEMININO	MASCULINO	NÃO INFORMADO	DECADA
99	26%	71%	3%	1980
5113199	17%	50%	33%	1990
2584759	17%	36%	46%	2000



Resultado: view **VW_TITULOS_COM_NOME_EXCLUSIVO_NO_BRASIL**

5 - Porcentagem de atores do sexo feminino e masculino que atuaram nos títulos do tipo Filme em cada década existente na base.

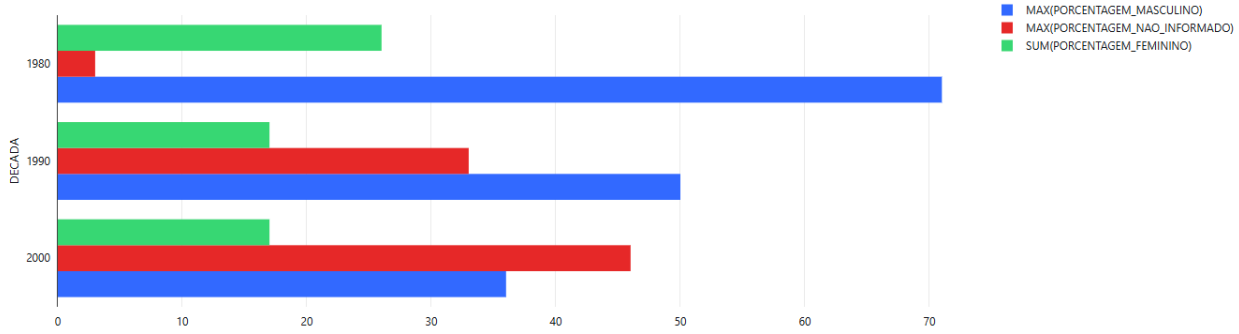
Para responder essa questão, foi criada a visão VW_TOTAL_POR_DECADA_SEXO_ATOR que tem como fonte as tabelas do schema IMDB_DB_GOLD TB_ATOR_TITULO_BRASIL e a tabela TB_FILME_LANCAMENTO_BRASIL.

O relacionamento entre elas ocorre com a junção utilizando a CHAVE_TITULO existente em ambas tabelas

Observação: Por estar respondendo outra pergunta que utiliza o SEXO como índice do resultado, contiuamos com os dados de baixa qualidade por ter valor do sexo “Não informado”, mesmo que o resultado esperado seja próximo a realidade, a quantidade de Homens nas produções costuma ser maior.

Resultado: view **VW_TOTAL_POR_DECADA_SEXO_ATOR**

TOTAL	FEMININO	MASCULINO	NÃO INFORMADO	DÉCADA
99	26	71	3	1980
5113199	17	50	33	1990
2584759	17	36	46	2000



Resultado: view **VW_TITULOS_COM_NOME_EXCLUSIVO_NO_BRASIL**

6 - Os primeiros 20 títulos lançados no Brasil, onde o Diretor(a) atuou como Ator/Atriz.

Resultado: **Consulta SQL**

Consulta realizada com a junção da tabela TB_ATOR_TITULO na camada Gold e a tabela TB_FILME_LANCAMENTO_BRASIL na camada Gold (IMDB_DB_GOLD) em que ordena pelo ano de lançamento do título e verifica se o campo DIRETOR_TITULO na tabela TB_ATOR_TITULO esta com valor 'S'.

Ao analisar os dados, não encontrei informações que possam contaminar a qualidade dos dados, desta forma podemos afirmar que dentro da massa de dados contida nos arquivos o resultado é íntegro.

NOME ARTÍSTICO	TÍTULO ORIGINAL	ANO LANÇAMENTO	REGIÃO	PROFISSÃO 1	PROFISSÃO 2	PROFISSÃO 3
Georges Méliès	LE MANOIR DU DIABLE	1896	BR	director	actor	producer
Georges Méliès	L'ÉCLIPSE DU SOLEIL EN PLEINE LUNE	1907	BR	director	actor	producer
Benjamin Oliveira	OS GUARANIS	1908	BR	actor	director	writer

Francisco Marzullo	OS ESTRANGULADORES	1908	BR	actor	director	cinematographer
Francisco Marzullo	ROUBO DOS 1,400 CENTOS	1908	BR	actor	director	cinematographer
Antônio Serra	LUCRÉCIA BORGIA	1909	BR	director	actor	null
Antônio Serra	PEGA NA CHALEIRA	1909	BR	director	actor	null
Antônio Serra	UM CAVALHEIRO DEVERAS OBSEQUIOSO	1909	BR	director	actor	null
Antônio Serra	UMA LICAO DE MAXIXE	1909	BR	director	actor	null
Eduardo Leite	A VIÚVA ALEGRE	1909	BR	actor	director	writer
Eduardo Leite	DONA INÊS DE CASTRO	1909	BR	actor	director	writer
Eduardo Leite	JOÃO JOSÉ	1909	BR	actor	director	writer
Eduardo Leite	O REMORSO VIVO	1909	BR	actor	director	writer
Emílio Silva	A VIÚVA ALEGRE	1909	BR	art_department	actor	cinematographer
Emílio Silva	AVENTURAS DE ZÉ CAIPORA	1909	BR	art_department	actor	cinematographer
Emílio Silva	PEGA NA CHALEIRA	1909	BR	art_department	actor	cinematographer
José Gonçalves Leonardo	SONHO DE VALSA	1909	BR	actor	director	null
João Colas	JOÃO JOSÉ	1909	BR	actor	director	null
João Colas	LA CHICANERA	1909	BR	actor	director	null
Emílio Silva	O RIO POR UM ÓCULO	1910	BR	art_department	actor	cinematographer

7 - Ranking dos diretores com mais de 7 títulos lançados no Brasil

Resultado: **Consulta SQL**

Consulta nas tabelas da camada Gold (IMDB_DB_GOLD) TB_ATOR_TITULO_BRASIL e TB_FILME_LANCAMENTO_BRASIL com agrupamento por título e diretor.
Para atender a pergunta esse agrupamento foi filtrado onde a contagem tenha valor maior e igual a 7.

Foi realizada a análise de títulos na tabela tb_title_crew com a tb_title_basics, indicando que há títulos sem referência de diretores e escritores.

Podemos afirmar que o resultado ilustra somente os títulos que possuem registros dos diretores nos arquivos.

NOME ARTÍSTICO	TOTAL
Pedro Murad	9
John Waters	8
Luciano Mello	8
Bruno de Oliveira	8
Bruno Costa	8
Marcelo Leme	8
Leonardo Martinelli	8
Anna Azevedo	7
George Miller	7
Joaquim Haickel	7
Maria Ribeiro	7
Paulo Miranda	7
Robert Gordon	7
Christian Caselli	7

8 - Filmes brasileiros distribuídos no exterior.

Verifiquei que a consulta de filmes com a flag IsOriginalTytle no arquivo title.akas.tsv não trás resultados com a coluna região contendo o valor “BR” (Brasil).

Com a fonte de dados do IMDb não posso gerar um resultado, mesmo que a obra tenha título em português, mesmo que seja de um diretor e atores conhecidos, por que não posso afirmar que foi produzida, dirigida e com atuação de atores brasileiros.

Análise

A documentação não detalha alguns valores que encontrei nos arquivos, como o caso de Types no arquivo title.akas, pois o atributo tem valores imdbDisplay e Alternative, em que posso

apenas especular uma definição mas não poderia usar como resultado de um relatório por não conhecer a finalidade desse valor.

Na Camada Gold listei a falta de referência nas tabelas que contêm os atores e os títulos (tb_name_basics e tb_title_basics) e na tabela tb_title_principals foi verificada a falta de referência em tb_title_basics. Então não posso afirmar que o resultado das minhas perguntas se refere a todos os títulos cadastrados em title_principals. Por ser uma dos principais arquivos da base, essa ineficiência compromete qualquer informação que seja relacionada a qualquer obra.

A base de dados conseguiu responder 7 das 8 perguntas que listei ao fazer a primeira validação do que cada arquivo.

Afirmo que para um relatório com base em percentual ou indicadores, essa base atende, mas para resultados com números absolutos pode não refletir a realidade.

Autoavaliação

Por ter bastante familiaridade com SQL e Python, foquei muito em testes de novas consultas, para conhecer melhor o Databricks, mas acho que deveria ter focado em uma versão V0 e depois ganhar tempo para utilizar o Power BI o Mongo e outras ferramentas similares.

Confesso que essa Sprint superou as expectativas, me “obrigando” conhecer novos ambientes de tecnologia (Databricks, Collab, Mongo) e por esse motivo acho que perdi tempo tentando aprender a usar esses ambientes.
.
Mas nas aulas de dúvidas, vi que estava sem direcionamento e me ajudou a focar nesse trabalho com mais qualidade, apesar de ter consumido um tempo desnecessário.