**Logotipo

O conteúdo gerado por IA pode estar incorreto.**

**PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro**

**MVP – Minimum Viable Product**

**Pós-graduação em Ciência de Dados e Analytics**

**Sprint: Engenharia de Dados**

Aluno: **Marcos Pasquale Vieira Scofano**

Sumário

[Descrição 3](#_Toc194507021)

[Objetivo 3](#_Toc194507022)

[Plataforma 4](#_Toc194507023)

[Busca 5](#_Toc194507024)

[Coleta 8](#_Toc194507025)

[Modelagem de Dados 10](#_Toc194507026)

[Catálogo de dados 10](#_Toc194507027)

[Carga 14](#_Toc194507028)

[Camada Prata (silver) 14](#_Toc194507029)

[Camada Ouro (gold) 16](#_Toc194507030)

[Análise 19](#_Toc194507031)

[Qualidade de Dados 19](#_Toc194507032)

[Solução do Problema 21](#_Toc194507033)

[Autoavaliação 22](#_Toc194507034)

[Objetivos Conquistados 22](#_Toc194507035)

[Dificuldades Encontradas 22](#_Toc194507036)

[Conclusão 23](#_Toc194507037)

# Descrição

Este documento descreve o trabalho de construção de um MVP que englobe as etapas de busca, coleta, modelagem, carga e análise dos dados.

# Objetivo

O objetivo do trabalho é realizar a análise dos dados obtidos do repositório de informações de filmes e séries IMDB – Internet Movie Database (<https://www.imdb.com>) com o intuito de responder as seguintes questões:

* Quais são os 10 filmes que possuem o maior número de avaliações dos usuários?
* Qual é o ranking dos 20 filmes mais bem avaliados?
* Quais atores e atrizes mais aparecem nos filmes do ranking dos 100 filmes mais bem avaliados?
* Qual é a distribuição de notas médias dos filmes ao longo das décadas?
* Quais diretores que possuem a maior média de avaliação em seus filmes?
* Filmes de orçamento maior tendem a ter avaliações mais altas?
* Quais são os estúdios que mais produziram filmes com notas iguais ou acima de 7,0 ao longo das décadas?
* Quais filmes após 2010 foram encenados apenas por atrizes?
* Quais filmes anteriores ao ano 2000 que foram dirigidos por mulheres?
* Qual filme teve o maior número de atores e atrizes?
* Quantos filmes foram produzidos por década?
* Quais o ranking do número de filmes produzidos por país e por ano?
* Qual a avaliação média dos filmes produzidos por gênero?
* Quais os 3 gêneros mais bem avaliados por década?
* Qual o tempo médio de duração dos filmes e como ele mudou ao longo dos anos?

# Plataforma

A plataforma a ser utilizada para o desenvolvimento do MVP será o Databricks Community (<https://community.cloud.databricks.com/>), conforme recomendação dos professores.

Os notebooks e demais artefatos desenvolvidos para solução serão armazenados no GitHub (<https://github.com/>) em um repositório disponibilizado na conta do aluno ([marcos.scofano@gmail.com](mailto:marcos.scofano@gmail.com)).

# Busca

Conforme informado anteriormente, o IMDB foi escolhido como fonte do conjunto de dados (*datasets*) para o trabalho. Os arquivos de uso não comercial estão disponibilizados no endereço <https://datasets.imdbws.com/>.

Os seguintes arquivos estão disponibilizados na plataforma:

* name.basics.tsv.gz;
* title.akas.tsv.gz;
* title.basics.tsv.gz;
* title.crew.tsv.gz;
* title.episode.tsv.gz;
* title.principals.tsv.gz;
* title.ratings.tsv.gz.

Os arquivos, após a descompactação, têm o formato texto e possuem os valores separados por TAB. A primeira linha apresenta os cabeçalhos das colunas. Os valores nulos são representados por “\N”. O IMDB informa que o conjunto de arquivos é atualizado diariamente.

As seguintes tabelas detalham cada um dos arquivos disponibilizados e seus respectivos campos (informações obtidas de <https://developer.imdb.com/non-commercial-datasets/>):

|  |  |  |
| --- | --- | --- |
| **Arquivo:** name.basics.tsv  **Descrição:** relação de pessoas (atrizes, atores, diretores, produtores, etc.) que trabalham nos filmes e nas séries  **Quantidade de Registros (13/03/2025):** 14.251.048 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| nconst | texto | identificador da pessoa |
| primaryName | texto | nome pessoa |
| birthYear | número inteiro | ano de nascimento |
| deathYear | número inteiro | ano de falecimento |
| primaryProfession | vetor de texto (separados por vírgula) | principais profissões da pessoa |
| knownForTitles | vetor de tconst (referência a obras) | referência às obras pelas quais aquela pessoa é mais conhecida |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.akas.tsv  **Descrição:** relação dos nomes originais e regionais dos filmes e séries  **Quantidade de Registros (13/03/2025):** 51.631.235 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| titleId | texto | identificador da obra |
| ordering | número inteiro | número de ordenação para identificação única |
| title | texto | título ou nome da obra |
| region | texto | região da versão do título |
| language | texto | idioma do título |
| types | vetor de texto (separados por vírgula) | enumeração de tipos que podem assumir os valores:  "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay" |
| attributes | vetor de texto (separados por vírgula) | atributos adicionais da versão do título |
| isOriginalTitle | booleano | indicador se a versão do título é a original |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.basics.tsv  **Descrição:** relação das obras (filmes e séries) com suas informações básicas  **Quantidade de Registros (13/03/2025):** 11.516.411 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| tconst | texto | identificador da obra |
| titleType | texto | O tipo ou formato da obra. Ex.: movie (filme), short (curta), tvseries (série), tvepisode (episódio de TV), vídeo (vídeo), etc. |
| primaryTitle | texto | título da obra |
| originalTitle | texto | título original da obra |
| isAdult | booleano | identificador se a obra tem conteúdo adulto ou não |
| startYear | número inteiro | representa o ano de lançamento de uma obra, ou sua temporada inicial para séries |
| endYear | número inteiro | último ano da série. Nulo para filmes. |
| runtimeMinutes | número inteiro | duração da obra em minutos |
| genres | vetor texto (separados por vírgula) | relação de até três gêneros associados  à obra (animação, comédia, romance, etc.) |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.crew.tsv  **Descrição:** relação com os diretores e roteiristas das obras  **Quantidade de Registros (13/03/2025):** 11.516.411 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| tconst | texto | Identificador da obra |
| directors | vetor de nconsts (referência para pessoas) | lista de diretores da obra |
| writers | vetor de nconsts (referência para pessoas) | lista de roteiristas da obra |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.episode.tsv  **Descrição:** relação entre os episódios e a obra principal (série)  **Quantidade de Registros (13/03/2025):** 8.859.676 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| tconst | texto | identificador da obra (episódio) |
| parentTconst | texto | identificador da obra (referência à obra principal) |
| seasonNumber | número inteiro | número da temporada |
| episodeNumber | número inteiro | número do episódio |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.principals.tsv  **Descrição:** relacionamento entre as obras e as pessoas que trabalharam  **Quantidade de Registros (13/03/2025):** 91.421.753 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| tconst | texto (referência para obra) | identificador da obra |
| ordering | número inteiro | número de ordenação para identificação única |
| nconst | texto (referência para pessoa) | identificador da pessoa |
| category | texto | categoria função da pessoa na obra |
| job | texto | função da pessoa na obra |
| characters | vetor de texto | lista de personagens (em caso de atores e atrizes) |

|  |  |  |
| --- | --- | --- |
| **Arquivo:** title.ratings.tsv  **Descrição:** relação das notas de avaliação dos usuários para a obra  **Quantidade de Registros (13/03/2025):** 1.544.811 | | |
| **Coluna** | **Formato** | **Descrição da coluna** |
| tconst | texto (referência para obra) | identificador da obra |
| averageRating | número de ponto flutuante | nota média atribuída à obra |
| numVotes | número inteiro | número de votos atribuídos à obra |

# Coleta

A coleta dos arquivos é realizada através de um notebook exclusivo para essa finalidade chamado de “ingestao\_dados\_brutos”. O notebook realiza o download direto da página do IMDB (<https://datasets.imdbws.com/>) através de comandos WGET, salvando-os no diretório “/mnt/imdb/” do cluster no Databricks. Veja a captura de tela abaixo da execução desse trecho de código:

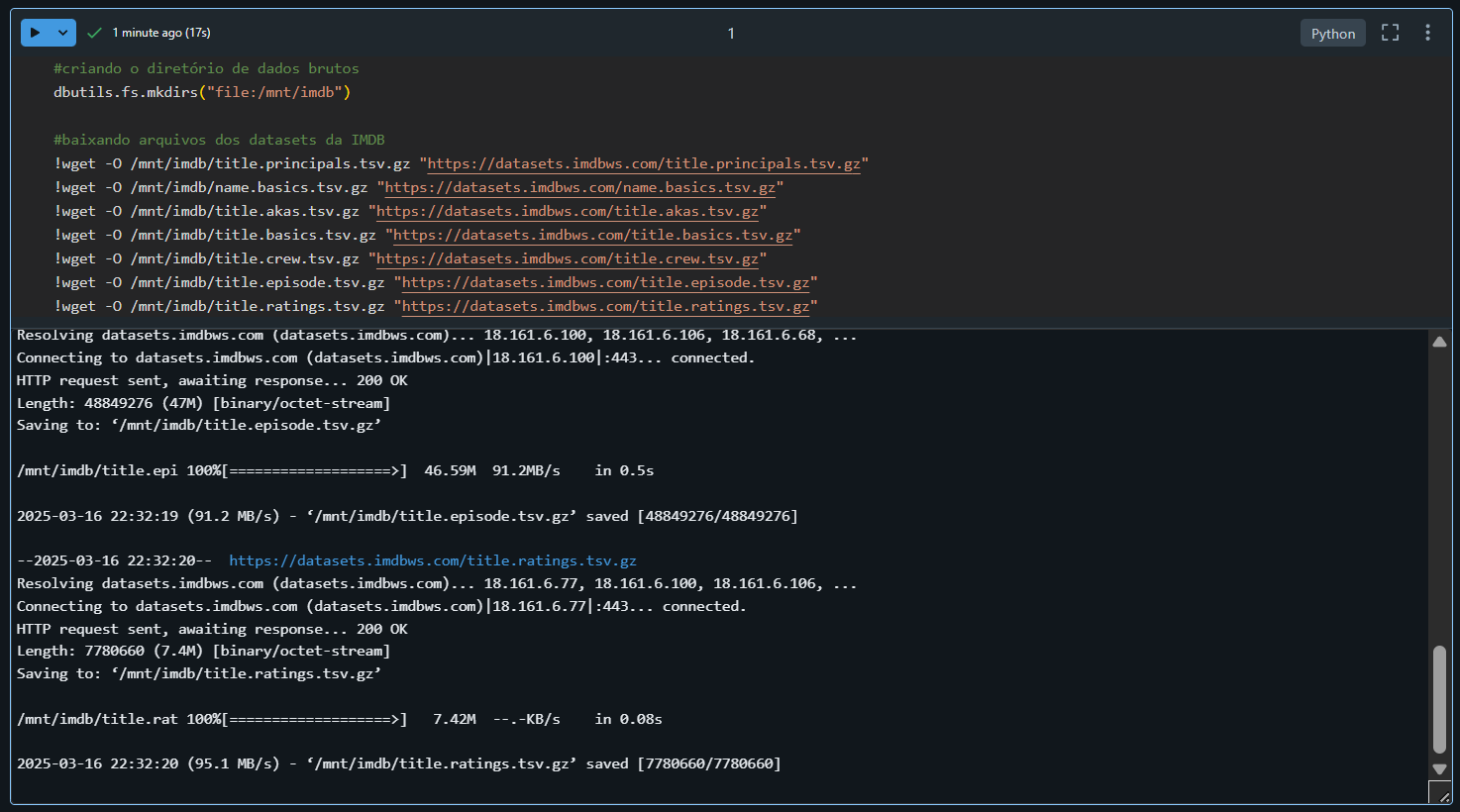


Figura : Download dos arquivos do dataset do IMDB (comandos WGET)

Após a disponibilização dos arquivos no sistema de arquivos (*filesystem*) é realizada a sua descompactação através da execução do comando *gunzip*. Foi utilizada a opção *verbose* (-v), para que seja possível visualizar o processamento da descompactação. A figura abaixo evidencia o processo de descompactação dos arquivos:

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Figura : Processo de descompactação dos arquivos utilizando GUNZIP

Para fins de evidenciação e conferência, foi realizado o comando “ls” (listar arquivos) no diretório criado para receber os arquivos (“/mnt/imdb/”). Veja a figura abaixo:

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Figura : arquivos no diretório criado para receber os dados brutos (raw data)

# Modelagem de Dados

Para procurar atender os objetivos propostos para esse trabalho, foi desenvolvido o seguinte modelo de dados:

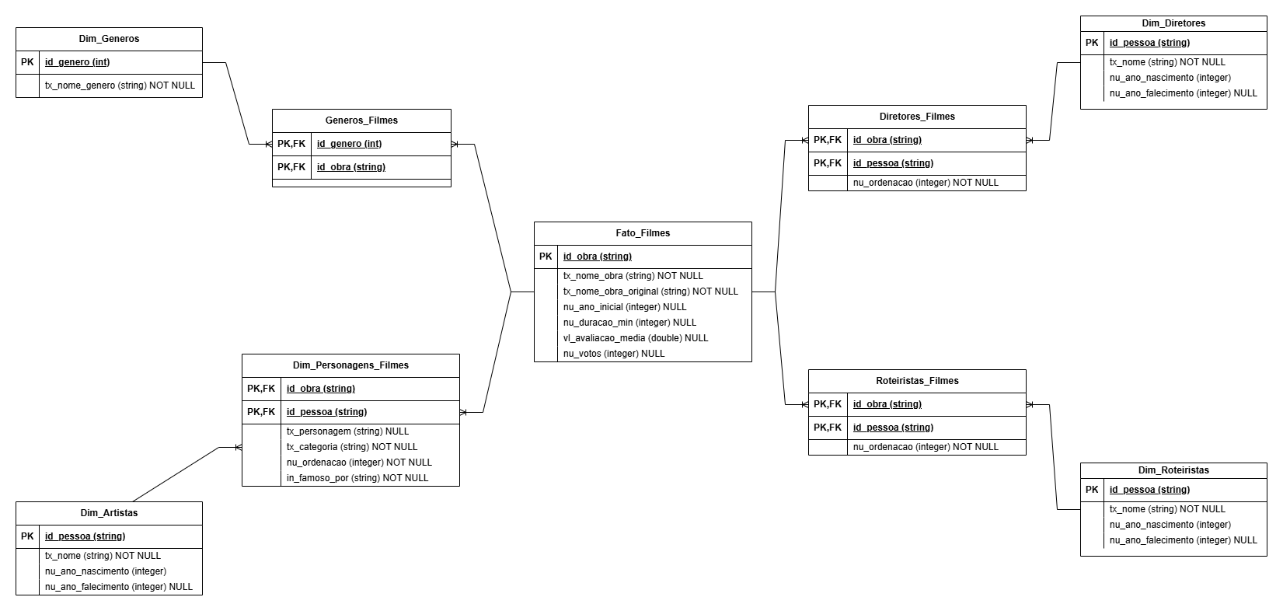


Figura : Modelo de Dados implementado na solução

## Catálogo de dados

A tabela abaixo representa o catálogo dos dados que foram mapeados no modelo acima:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Fato\_Filmes  **Descrição:** Tabela fato com as informações dos filmes (apenas filmes de longa metragem e não adultos).  **Linhagem:** title.basics.tsv e title.ratings.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_obra | string | não | N/A | Identificador único da obra (ou filme) | title.basics.tsv (tconst) 🡪 default.obras (id\_obra) 🡪 curado.fato\_filmes (id\_obra) |
| tx\_nome\_obra | string | não | N/A | Nome do filme | title.basics.tsv (originalTitle)  🡪 default.obras (tx\_nome\_obra) 🡪 curado.fato\_filmes (tx\_nome\_obra) |
| tx\_nome\_obra\_original | string | não | N/A | Nome original do filme | title.basics.tsv (primaryTitle) 🡪 default.obras (tx\_nome\_obra\_original) 🡪 curado.fato\_filmes (tx\_nome\_obra\_original) |
| nu\_ano\_inicial | integer | sim | Entre 1888 e ∞ | Ano de lançamento do filme | title.basics.tsv (startYear)  🡪 default.obras (nu\_ano\_inicial) 🡪 curado.fato\_filmes (nu\_ano\_inicial) |
| nu\_duracao\_min | integer | sim | Entre 1 e ∞ | Duração do filme em minutos | title.basics.tsvr (runtimeMinutes) 🡪 default.obras (nu\_duracao\_min) 🡪 curado.fato\_filmes (nu\_duracao\_min) |
| vl\_avaliacao\_media | double | sim | Entre 0.0 e 10.0 | Média da nota de avaliação | title.ratings.tsv (averageRating) 🡪 default.obras (vl\_avaliacao\_media) 🡪 curado.fato\_filmes (vl\_avaliacao\_media) |
| nu votos | integer | sim | Entre 0 e ∞ | Número de votos | title.ratings.tsv (numVotes) 🡪 default.obras (nu\_votos) 🡪 curado.fato\_filmes (nu\_votos) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Generos\_Filmes  **Descrição:** Tabela de ligação (*bridge table*) entre a tabela fato de filmes e a tabela de dimensão de gêneros.  **Linhagem:** title.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_obra | string | não | N/A | Identificador único da obra (ou filme) | title.basics.tsv (genres) 🡪 default.obras (id\_obra) 🡪 curado.genero\_filmes (id\_obra) |
| id\_genero | int | não | N/A | Identificador único do gênero | Chave única de gênero criada no processo de ETL |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Dim\_Genero  **Descrição:** Tabela dimensão com os gêneros dos filmes (um filme pode ter vários gêneros).  **Linhagem:** title.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_genero | int | não | N/A | Identificador único do gênero | Chave única criada no processo de ETL |
| tx\_nome\_genero | string | não | Action, Adult, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, Game-Show, History, Horror, Music, Musical, Mystery, News, Reality-TV, Romance, Sci-Fi, Short, Sport, Talk-Show, Thriller, War e Western | Nome do gênero | title.basics.tsv (genres)  🡪 default.generos (tx\_nome\_genero) 🡪 curado.dim\_generos (tx\_nome\_genero) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Diretores\_Filmes  **Descrição:** Tabela de ligação (*bridge table*) entre a tabela fato de filmes e a tabela de dimensão de diretores. Foram filtrados apenas as pessoas que trabalharam nos filmes com a categoria (title.principals.tsv (category)) igual a “director"  **Linhagem:** title.principals.tsv e name.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_obra | string | não | N/A | Identificador único da obra (ou filme) | title.principals.tsv (tconst) 🡪 default.pessoas\_obras (id\_obra) 🡪 curado.diretores\_filmes (id\_obra) |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | name.basics.tsv (nconst) 🡪 default.pessoas\_obras (id\_pessoa) 🡪 curado.diretores\_filmes (id\_pessoa) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Dim\_Diretores  **Descrição:** Tabela dimensão com os diretores dos filmes (um filme pode ter vários diretores).  **Linhagem:** title.principals.tsv e name.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | title.principals.tsv (nconst) 🡪 default.pessoas (id\_pessoa) 🡪 curado.dim\_diretores (id\_pessoa) |
| tx\_nome | string | não | N/A | Nome do diretor | name.basics.tsv (primaryName) 🡪 default.pessoas (tx\_nome) 🡪 curado.dim\_diretores (tx\_nome) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Roteiristas\_Filmes  **Descrição:** Tabela de ligação (*bridge table*) entre a tabela fato de filmes e a tabela de dimensão de roteiristas. Foram filtrados apenas as pessoas que trabalharam nos filmes com a categoria (title.principals.tsv (category)) igual a “writer"  **Linhagem:** title.principals.tsv e name.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_obra | string | não | N/A | Identificador único da obra (ou filme) | title.principals.tsv (tconst) 🡪 default.pessoas\_obras (id\_obra) 🡪 curado.roteiristas\_filmes (id\_obra) |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | name.basics.tsv (nconst) 🡪 default.pessoas\_obras (id\_pessoa) 🡪 curado.roteiristas\_filmes (id\_pessoa) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Dim\_Roteiristas  **Descrição:** Tabela dimensão com os roteiristas dos filmes (um filme pode ter vários roteiristas).  **Linhagem:** title.principals.tsv e name.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | title.principals.tsv (nconst) 🡪 default.pessoas (id\_pessoa) 🡪 curado.dim\_roteiristas (id\_pessoa) |
| tx\_nome | string | não | N/A | Nome do roteirista | name.basics.tsv (primaryName) 🡪 default.pessoas (tx\_nome) 🡪 curado.dim\_roteiristas (tx\_nome) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Dim\_Personagens\_Fimes  **Descrição:** Tabela dimensão com os personagens dos filmes (um filme pode ter vários personagens).  **Linhagem:** title.principals.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | title.principals.tsv (nconst) 🡪 default.pessoas\_obras (id\_pessoa) 🡪 curado.dim\_personagens\_filmes (id\_pessoa) |
| id\_obra | string | não | N/A | Identificador único da obra (ou filme) | title.principals.tsv (tconst) 🡪 default.pessoas\_obras (id\_obra) 🡪 curado.dim\_personagens\_filmes (id\_obra) |
| nu\_ordenacao | int | não | Entre 1 e ∞ | Ordenação definida pelo IMDB | title.principals.tsv (ordering) 🡪 default.pessoas\_obras (nu\_ordenacao) 🡪 curado.dim\_personagens\_filmes (nu\_ordenacao) |
| tx\_personagem | string | sim | N/A | Nome do personagem | title.principals.tsv (characters) 🡪 default.pessoas\_obras (tx\_personagem) 🡪 curado.dim\_personagens\_filmes (tx\_personagem) |
| tx\_categoria | string | não | actor e actress | Função desempenhada para atuação do personagem | title.principals.tsv (characters) 🡪 default.pessoas\_obras (tx\_categoria) 🡪 curado.dim\_personagens\_filmes (tx\_categoria) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabela:** Dim\_Artistas  **Descrição:** Tabela dimensão com os artistas (atores e atrizes) que atuaram nos filmes.  **Linhagem:** name.basics.tsv | | | | | |
| **Coluna** | **Tipo do Dado** | **Nulo** | **Domínio ou**  **Valores min./máx.** | **Descrição** | **Linhagem** |
| id\_pessoa | string | não | N/A | Identificador único de pessoa | name.basics.tsv (nconst) 🡪 default.pessoas (id\_pessoa) 🡪 curado.dim\_artistas (id\_pessoa) |
| tx\_nome | string | não | N/A | Nome do artista | name.basics.tsv (primaryName) 🡪 default.pessoas (tx\_nome) 🡪 curado.dim\_artistas (tx\_nome) |
| nu\_ano\_nascimento | int | sim | N/A | Ano de nascimento do artista | name.basics.tsv (birthYear) 🡪 default.pessoas (nu\_ano\_nascimento) 🡪 curado.dim\_artistas (nu\_ano\_nascimento) |
| nu\_ano\_falecimento | Int | sim | N/A | Ano de falecimento do artista | name.basics.tsv (deathYear) 🡪 default.pessoas (nu\_ano\_falecimento) 🡪 curado.dim\_artistas (nu\_ano\_falecimento) |

# Carga

O pipeline da carga de dados para os *schemas* “silver” (*schema* “default” do cluster) e “gold” (*schema* “curado” do cluster) foi desenvolvido pelo código presente no notebook intitulado “curadoria”.

As tabelas do *schema* “default” são intermediárias e fruto das primeiras operações de higienização dos dados. Ao analisar os dados disponibilizados pelo IMDB, é possível inferir que, apesar de serem disponibilizados no formato “tsv” (*tab separated values*), eles têm origem em visões de uma base relacional bem estruturada. Portanto, os dados possuem organização, coerência, tipos e relacionamentos bem definidos e a presença de chaves primárias e estrangeiras.

## Camada Prata (silver)

Podemos elencar as seguintes transformações como as principais operações realizadas nos arquivos originais disponibilizados no *dataset* do IMDB para proporcionar a disponibilização da camada prata:

* **Todos os arquivos:**
  + **Tratamento de nulo:** o IMDB convencionou representar campos nulos com o valor “\N” em todos os arquivos do *dataset*. Portanto, todos os *dataframes Spark* sofrem a operação de substituição desses valores em todas linhas e colunas pelo valor nulo propriamente dito (*None* no Python);
  + **Novos nomes de coluna:** para todos os arquivos foi realizada a troca dos nomes das colunas. Para todos os *dataframes* foi definido um vetor com pares de valores (nome antigo, nome novo) e foi realizada um laço (loop) iterando sobre esse vetor e realizando as operações para renomear as colunas.
* **title.basics.tsv** (dados de obras/filmes)**:**
  + **Conversão de valores:** conversão dos valores numéricos de “ano inicial da obra”, “ano final da obra” e “duração em minutos” para número inteiro;
  + **Tratamento de campo multivalorado de gêneros:** o campo *genres* (gêneros) é multivalorado, ou seja, os múltiplos gêneros associados à obra são elencados nesse campo em valores separados por vírgula (Ex.: “Romance,Animation,Comedy”). Essa situação recebe o tratamento através da criação de uma tabela (generos) com o domínio de todos os gêneros presentes no *dataframe* de filmes e a criação de uma chave primária (id\_genero). Para realizar a associação da nova tabela de gêneros com a tabela de obras/filmes, é criada a tabela generos\_obras relacionando a tabela de obras/filmes com gêneros (relacionamento N para N);
  + **Tabela da camada prata:** os *dataframes* transformados foram salvos nas tabelas **default.obras**, **default.generos\_obras** e **default.generos** na camada prata.
* **title.ratings.tsv** (dados de avaliações)**:** 
  + **Conversão de valores:** conversão dos valores numéricos de “número de votos” para número inteiro e “nota de avaliação média” para ponto flutuante (*double*);
  + **Junção com a tabela de obras/filmes:** como as avaliações se relacionam com obras/filmes com cardinalidade 1 para 1, é realizada a junção dessas tabelas através do identificador do filme/obra presente nos dois conjuntos de dados. Dessa forma, a tabela de obras/filmes passa a incorporar os campos do número de votos e da nota média das avaliações.
  + **Tabela da camada prata:** os dados do *dataframe* foram incorporados na tabela default.obras na camada prata, portanto não há tabela específica para avaliações.
* **name.basics.tsv** (dados de pessoas)**:** 
  + **Conversão de valores:** conversão dos valores numéricos de “ano de nascimento” e “ano de falecimento” para número inteiro;
  + **Tratamento de campo multivalorado que relacionam as pessoas às obras/filmes pelos quais elas são conhecidas (knownForTitles):** o campo *knownForTitles* é multivalorado, ou seja, as múltiplas obras/filmes associados à pessoa são elencados nesse campo em valores (identificadores de obras/filmes) separados por vírgula (Ex.: “tt0051808,tt0089881,tt0080979,tt0057565”). Essa situação recebe o tratamento através da criação de uma tabela (pessoa\_fama\_obra) com o relacionamento entre pessoas e obras/filmes pelos quais elas são notoriamente relacionadas (relacionamento N para N), substituindo a coluna multivalorada por uma tabela de relacionamento com referências ao identificador da pessoa e ao identificador da obra/filme;
  + **Tabela da camada prata:** os *dataframes* transformados foram salvos nas tabelas **default.pessoas** e **default.pessoas\_fama\_obras** na camada prata;
* **title.principals.tsv** (relacionamento de pessoas que trabalharam em obras/filmes):
  + **Conversão de valores:** conversão dos valores numéricos do campo de ordenação do IMDB para número inteiro;
  + **Tratamento do campo de personagens:** caso a relação entre a pessoa e o filme tenha sido como uma função de atuação (*actor* ou *actress*) o campo *characters* (personagens) estará preenchido com um conjunto de nomes do personagem separados por vírgula. Foi verificado que caso um artista atue representando mais de um personagem em um filme, o conjunto de dados separa uma linha no arquivo para cada personagem. Dessa forma, para fins de simplificação, o campo *characters*  recebeu um tratamento para ser substituído pelo primeiro valor do vetor;
  + **Tabela da camada prata:** o *dataframe* transformado foi salvo na tabela **default.pessoas\_obras** na camada prata.
* **title.akas.tsv** (nomes regionais de filmes )**:** este arquivo não recebeu tratamentos além daqueles que foram feitos para todos os arquivos;
  + **Tabela da camada prata:** o *dataframe* transformado foi salvo na tabela **default.nomes\_obras** na camada prata.
* **title.episode.tsv** (episódios de séries/minisséries):
  + **Conversão de valores:** conversão dos valores numéricos de “número da temporada” e “número do episódio” para número inteiro.
  + **Tabela da camada prata:** o *dataframe* transformado foi salvo na tabela **default.episodios** na camada prata.

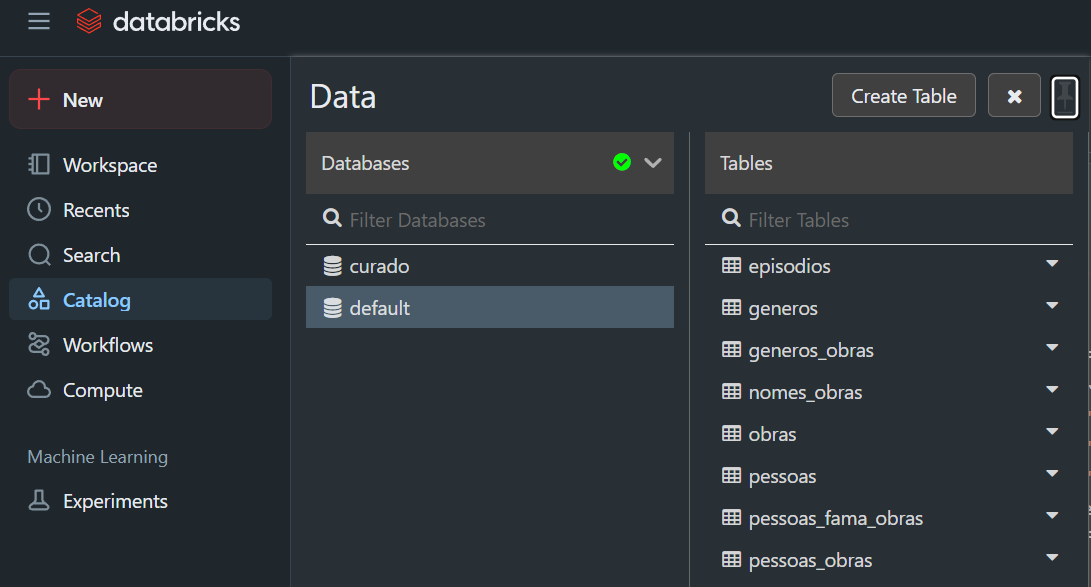


Figura : Tabelas geradas na camada prata

## Camada Ouro (gold)

A estratégia para a criação das tabelas da camada ouro, conforme o modelo da Figura 4, foi escrever *views* em SQL a partir das tabelas da camada prata. Segue abaixo o detalhamento das operações realizadas para a obtenção de cada tabela da camada ouro:

* **curado.fato\_filmes:** a view é resultado da seleção das linhas da tabela default.obras filtradas com *in\_tipo\_obra* igual a “movie”, ou seja, apenas filmes longa-metragem, e *in\_adulto igual* a “0” (zero), que seleciona apenas filmes não adultos;
* **curado.dim\_generos:** seleção simples de todas as linhas da tabela default.generos;
* **curado.generos\_filmes:** seleção das linhas da tabela default.generos\_obras, mas filtrando por apenas aquelas referentes aos filmes presentes na curado.fato\_filmes, a fim de reduzir o volume de dados e manter a consistência no relacionamento;
* **curado.dim\_personagens\_filmes:** seleção das linhas da tabela default.pessoas\_obras, mas filtrando por apenas aquelas referentes aos filmes presentes na curado.fato\_filmes e em que a categoria de função na abro seja de ator (*actor*) ou atriz (*actress*), a fim de reduzir o volume de dados e manter a consistência no relacionamento. Foi feito também um tratamento para sinalizar nessa tabela se um(a) ator(atriz) é conhecido(a) pela participação na obra em questão. Essa sinalização (campo *in\_famoso\_por*) é realizada através da consulta das chaves id\_obra e id\_pessoa na tabela default.pessoas\_fama\_obra. Se a dupla de chaves estiver presente na tabela, o valor “1” é atribuído campo *in\_famoso\_por*, senão é atribuído o valor “0”;
* **curado.dim\_artistas:** seleção das linhas da tabela default.pessoas, mas filtrando por pessoas associadas aos personagens presentes na curado.dim\_personagens, a fim de reduzir o volume de dados e manter a consistência no relacionamento;
* **curado.diretores\_filmes: :** seleção das linhas da tabela default.pessoas\_obras, mas filtrando apenas por aquelas que trabalharam com diretores (*directors*) e que tenham dirigido os filmes selecionados na tabela curado.fato\_filmes, a fim de reduzir o volume de dados e manter a consistência no relacionamento;
* **curado.dim\_diretores:** seleção das linhas da tabela default.pessoas, mas filtrando apenas aqueles que tenham dirigido os filmes selecionados na tabela curado.fato\_filmes, a fim de reduzir o volume de dados e manter a consistência no relacionamento;
* **curado.roteiristas\_filmes: :** seleção das linhas da tabela default.pessoas\_obras, mas filtrando apenas por aquelas que trabalharam com roteiristas (*writers*) e que tenham escrito para os filmes selecionados na tabela curado.fato\_filmes, a fim de reduzir o volume de dados e manter a consistência no relacionamento;
* **curado.dim\_roteiristas:** seleção das linhas da tabela default.pessoas, mas filtrando apenas aqueles que tenham escrito para os filmes selecionados na tabela curado.fato\_filmes, a fim de reduzir o volume de dados e manter a consistência no relacionamento;

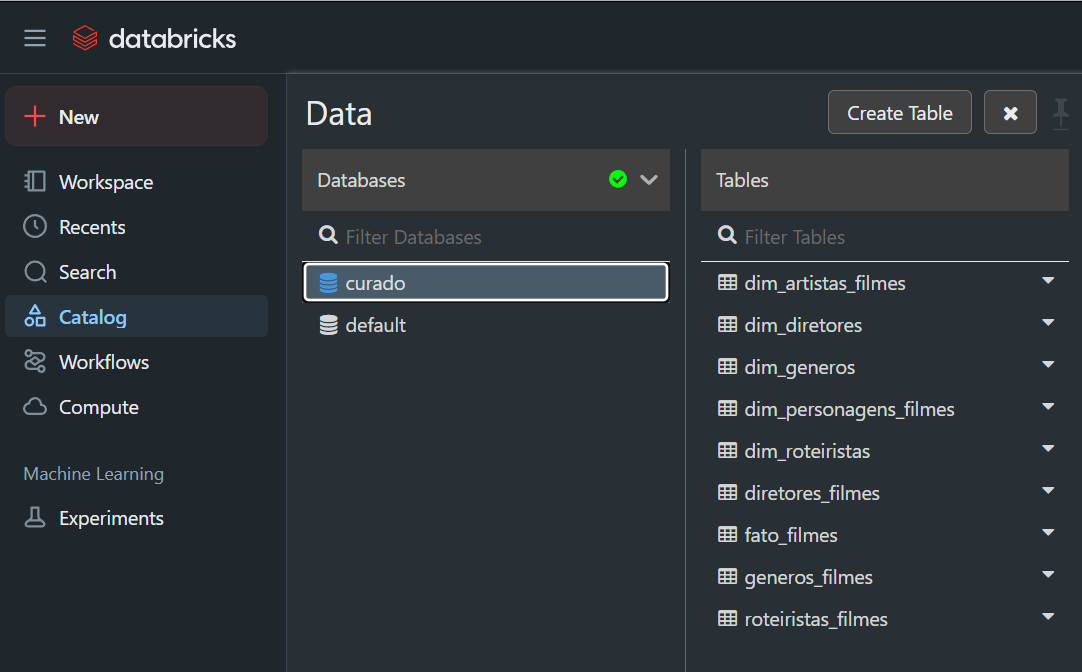
****

Figura : tabelas geradas na camada ouro

# Análise

## Qualidade de Dados

Conforme dito anteriormente, os dados do *dataset* disponibilizado pelo IMDB muito provavelmente são oriundos de uma base de dados bem estruturada, logo não foram encontrados problemas graves de integridade. Apesar disso, seguem algumas observações quanto aos dados:

* **Filmes sem ano de lançamento:** verificou-se na base de dados vários filmes com o campo nu\_ano\_inicial com o valor nulo. Como o *dataset* não possui o status do filme, observou-se diretamente no site do IMDB que tais títulos eram obras ainda não lançadas (em desenvolvimento, em pré-produção, em produção etc.);
* **Filmes com duração muito pequena:** foram observados alguns títulos com duração de apenas 1 minuto. Ao verificar no site do IMDB, encontramos situações que realmente constava a duração como no *dataset*, mas também títulos que apresentavam duração diferente (Ex.: 1h45min), sugerindo algum problema na produção do *dataset*;
* **Filmes com duração muito longa:** foram observados alguns filmes com duração extremamente grande (Ex.: 59.640 minutos ou 991 horas) e, apesar de serem dados reais, podem enviesar a métrica desenvolvida sem que haja um tratamento desses *outliars*. Veja a figura abaixo:

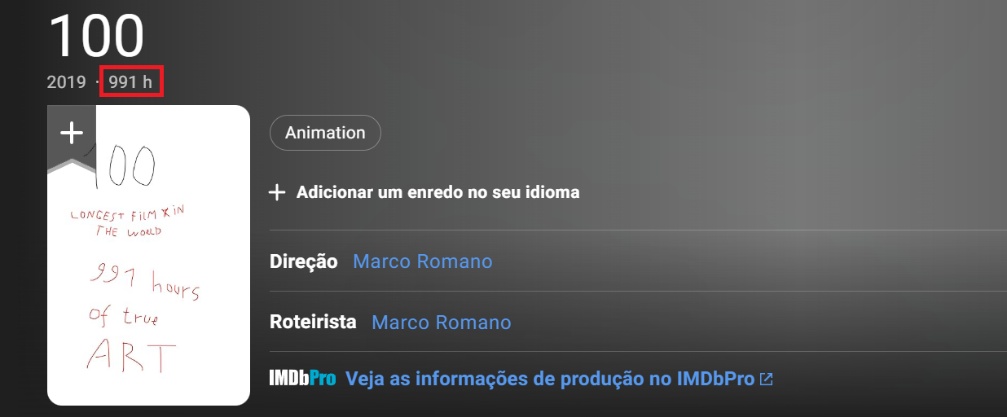


Figura : Filme com 991 horas de duração

* **Pessoas com ano de nascimento e falecimento incoerentes:** foram encontradas pessoas com ano de nascimento posterior ao ano de falecimento:

**Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.**

Figura : pessoas com datas inconsistentes

* **Limitação do conjunto de dados:** durante o processo de respostas aos objetivos do MVP foi encontrada uma limitação do conjunto de dados imposta ao *dataset* pelo IMDB. O número máximo de artistas (atores e atrizes) do elenco de um filme está limitado a 10 pessoas. Essa limitação impactou e até impediu a resposta a algumas questões.

## Solução do Problema

Para a solução dos pontos levantados na seção “Objetivos” desse documento, foi utilizado um notebook específico chamado “analise”. No notebook, foram produzidos os códigos em SQL e Python para responder as questões levantadas.

As evidências das execuções dos códigos e os comentários das respostas estão presentes no notebook “analise” exportado do Databricks Community e disponibilizados para conferência.

# Autoavaliação

Este trabalho teve como objetivo tanto desenvolver um pipeline de engenharia de dados utilizando conjuntos de dados públicos do IMDB (Internet Movie Database) para extrair, transformar e carregar (ETL) informações relevantes sobre filmes, atores e avaliações, como também gerar análises a partir do modelo de dados desenvolvido. Apesar dos desafios enfrentados, os objetivos principais foram cumpridos, ainda que algumas limitações tenham surgido devido à natureza incompleta dos dados disponíveis.

## Objetivos Conquistados

* **Pipeline Funcional:**
  + Foi possível construir um fluxo de trabalho eficiente, desde a extração dos dados brutos até a geração de tabelas analíticas;
  + Utilização de ferramentas como Apache Spark e Python para processamento distribuído, garantindo escalabilidade.
* **Transformações e Análises Realizadas:**
  + Limpeza e padronização de dados (ex: tratamento de valores nulos, normalização de colunas);
  + Criação de métricas para responder as questões levantadas nos objetivos do trabalho;
  + Modelagem dimensional (*Star Schema*) para facilitar consultas analíticas.
* **Superação de Limitações:**
  + Mesmo com conjuntos de dados incompletos (ex: falta de informações detalhadas sobre datas ou orçamentos), foi possível extrair insights válidos, adaptando as perguntas de pesquisa às informações disponíveis.

## Dificuldades Encontradas

* **Plataforma Databricks Community:**
  + **Recursos Limitados**: A versão gratuita impõe restrições de capacidade computacional, o que afetou o desempenho em processamentos mais complexos;
  + **Tempo de Sessão**: As sessões expiravam após certo tempo inativo, exigindo reinício do ambiente e perda de dados temporários;
  + **Armazenamento Efêmero**: Dificuldade para persistir dados entre sessões sem uso de soluções externas.
* **Qualidade dos Dados do IMDB:**
  + **Dados Incompletos**: Algumas tabelas não continham informações essenciais para responder a todas as hipóteses iniciais (ex: detalhes financeiros dos filmes).
  + **Inconsistências**: Campos como data de nascimento e data de falecimento apresentaram situações que não condizem com a realidade.

## Conclusão

O pipeline desenvolvido neste trabalho abrangeu todo o ciclo de vida dos dados, desde a coleta dos *datasets* (dados brutos) do IMDB até a utilização prática do modelo dimensional (Star Schema) criado para as análises. Apesar das limitações do Databricks Community e pelas limitações do conjunto de dados disponibilizado, foi possível estruturar um fluxo de trabalho robusto, que incluiu:

1. **Extração e Carga**: coleta eficiente dos dados brutos, com tratamento de inconsistências e armazenamento em formato apropriado;
2. **Transformação**: limpeza, enriquecimento e modelagem dimensional, garantindo dados prontos para análise;
3. **Análise e Consumo**: exploração do modelo dimensional para gerar as respostas aos objetivos planejados.