

Análise dos Dados do Dataset IMDb

Prof. Dr. Sávio Salvarino Teles de
Oliveira

Discentes:

Iury Tavares do Monte – 2025200234

João Pedro José Santos da Silva Guedes – 2025200237

Manuel Lucala Zengo – 2025200249

Owen Alves Lima – 2025200255

Pedro Felipe de Moraes Carrijo – 2025200257

Willgnner Ferreira Santos – 2025200271

1. Introdução	2
1.1 Contextualização do projeto	2
1.2 Fonte dos Dados	2
1.3 Objetivos	2
1.3.1 Objetivo Geral	2
1.3.2 Objetivos Específicos	3
1.4 Metodologia utilizada	3
2. Descrição do Caso de Uso/Domínio de Negócio	4
2.1 Apresentação do Negócio Escolhido	4
2.2 Processos Principais	4
2.3 Stakeholders envolvidos	4
2.4 Requisitos identificados	5
Requisitos Funcionais:	5
Requisitos Não Funcionais:	5
3. Banco de Dados Relacional (OLTP)	6
3.1 Diagrama Entidade-Relacionamento (DER)	6
3.2 Esquema Relacional	7
3.3 Versão Inicial	8
3.4 Versão Normalizada	8
3.5 Justificativas da normalização	10
3.6 Dicionário de Dados Completo	10
Tabela: classificacao_indicativa	10
Tabela: diretores	11
Tabela: generos	11
Tabela: filmes	11
Tabela: filmes_atores (Tabela de Junção N:N)	13
Tabela: filmes_generos (Tabela de Junção N:N)	13
4. Data Warehouse (OLAP)	14
4.1. Processo de negócio analisado	14
4.2. Identificação de Fatos e Dimensões	14
4.2.1 Tabela fato	14
4.2.2 Dimensões (contextos de análise):	14
4.2.3 Granularidade da Tabela Fato	15
4.3. Diagrama do Esquema Estrela	15
5. Decisões de Projeto	16
5.1. Principais escolhas técnicas	16
5.2. Justificativas das decisões	17
6. Conclusões e Aprendizados	17
7. Referências	18

1. Introdução

A disciplina de Introdução a Banco de Dados visa proporcionar aos alunos o conhecimento teórico e prático necessário para a modelagem e implementação de bancos de dados. Este trabalho tem como finalidade aplicar os conhecimentos adquiridos ao longo da disciplina em um projeto prático, promovendo a compreensão dos processos envolvidos na construção de um sistema de banco de dados desde a sua concepção até a análise de dados.

1.1 Contextualização do projeto

O presente trabalho tem como objetivo consolidar e avaliar a capacidade dos alunos na aplicação de conceitos e técnicas de modelagem, projeto, implementação e utilização de bancos de dados relacionais e Data Warehouses. A proposta é desenvolver um projeto prático que simule a construção de um sistema real, abrangendo desde a modelagem conceitual até a criação de um banco de dados relacional normalizado e um Data Warehouse simplificado.

1.2 Fonte dos Dados

Para realização deste projeto, foi utilizado um dataset público disponibilizado na plataforma Kaggle, denominado "IMDb Dataset of Top 1000 Movies and TV Shows". Link de acesso: <https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>. Este dataset contém informações detalhadas sobre os 1000 filmes e séries mais bem avaliados no IMDb, incluindo dados como título, ano de lançamento, classificação indicativa, duração, gêneros, avaliação no IMDb, metaspore, sinopse, receita, número de votos, diretor e atores principais.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver um sistema de banco de dados relacional e um Data Warehouse a partir de um dataset real, aplicando técnicas de modelagem, normalização, implementação e análise de dados, com o intuito de consolidar os conhecimentos adquiridos na disciplina.

1.3.2 Objetivos Específicos

- Realizar a análise exploratória do dataset escolhido;
- Elaborar o modelo conceitual (Diagrama Entidade-Relacionamento);
- Realizar a conversão para modelo lógico e posteriormente para modelo físico;
- Aplicar técnicas de normalização para garantir integridade e evitar redundâncias;
- Implementar o banco de dados em um SGBD relacional;
- Criar um modelo multidimensional para construção de um Data Warehouse simplificado;
- Realizar consultas SQL e análises com base nos dados armazenados;
- Demonstrar a aplicabilidade do sistema através de exemplos práticos.

1.4 Metodologia utilizada

O desenvolvimento deste projeto seguiu uma abordagem prática dividida em etapas sequenciais, que simulam o ciclo de vida de um banco de dados:

- **Coleta e compreensão dos dados:** Estudo do dataset obtido no Kaggle para compreender os atributos disponíveis e possíveis relacionamentos entre as entidades.
- **Modelagem conceitual:** Criação do diagrama ER com base na análise dos dados.
- **Modelagem lógica e física:** Conversão do modelo conceitual para os modelos lógico e físico, incluindo definição de tipos de dados e chaves primárias/estrangeiras.
- **Normalização:** Aplicação das formas normais até a 3ª forma normal para garantir consistência e reduzir redundâncias.
- **Implementação no SGBD:** Criação do banco de dados e suas tabelas em um sistema.
- **Construção do Data Warehouse:** Modelagem estrela com foco em análise, permitindo gerar visões multidimensionais dos dados.

- **Consultas e análise de dados:** Desenvolvimento de consultas SQL e possíveis visualizações para análise e interpretação dos dados.

2. Descrição do Caso de Uso/Domínio de Negócio

2.1 Apresentação do Negócio Escolhido

O domínio de negócio escolhido para este projeto é a indústria cinematográfica, com foco em filmes e séries de televisão mais bem avaliados no IMDb. Esse setor movimenta bilhões de dólares mundialmente e envolve uma ampla cadeia de produção, distribuição e avaliação crítica. A base de dados IMDb representa uma referência reconhecida no segmento audiovisual, sendo amplamente utilizada por profissionais, entusiastas e pesquisadores para análise de tendências, desempenho de produções e preferências do público.

O sistema a ser modelado visa organizar e facilitar o acesso a informações relevantes sobre as obras presentes no dataset, permitindo consultas sobre filmes, diretores, atores, gêneros, classificações indicativas, avaliações e dados financeiros (como receita bruta). A modelagem também permite a expansão do sistema para análises mais aprofundadas, como ranking de filmes por gênero ou avaliação, e cruzamentos entre diretores e número de obras bem avaliadas.

2.2 Processos Principais

- **Cadastro e manutenção de dados:** organização e estruturação das informações de filmes, diretores, atores, gêneros e classificações indicativas;
- **Relacionamento entre entidades:** associação de filmes com seus respectivos diretores, elenco principal, gêneros e classificação indicativa;
- **Consultas e análises:** obtenção de informações específicas para análises de desempenho, como rankings por avaliação ou receita;
- **Construção de relatórios analíticos:** por meio do Data Warehouse, geração de visões agregadas e multidimensionais para análise de tendências e padrões de comportamento do público.

2.3 Stakeholders envolvidos

Os stakeholders envolvidos na aplicação de um sistema baseado nesse domínio de negócio incluem:

- **Estudantes e professores:** para fins acadêmicos, como neste projeto, com foco em ensino e avaliação de conhecimentos em modelagem de dados;
- **Produtoras e estúdios:** interessados em análises de desempenho de filmes e séries, para auxiliar na tomada de decisões;
- **Plataformas de streaming:** que podem utilizar dados para compor catálogos, sugerir conteúdos e avaliar popularidade de produções;
- **Pesquisadores e analistas de dados:** interessados em estudar o comportamento do público e tendências do mercado audiovisual;
- **Fãs e usuários comuns:** que buscam informações detalhadas sobre suas produções favoritas.

2.4 Requisitos identificados

A partir da análise do dataset e da proposta do projeto, foram identificados os seguintes requisitos:

Requisitos Funcionais:

- **RF01:** O sistema deve permitir o armazenamento de dados sobre filmes, incluindo título, ano, duração, sinopse, avaliação e receita;
- **RF02:** O sistema deve manter informações sobre diretores e atores relacionados aos filmes;
- **RF03:** O sistema deve permitir associar múltiplos gêneros e atores a cada filme;
- **RF04:** O sistema deve permitir a classificação indicativa de cada filme;
- **RF05:** O sistema deve possibilitar a realização de consultas por filtros diversos (por gênero, ano, avaliação, etc.);
- **RF06:** O sistema deve gerar visões agregadas no Data Warehouse, como média de avaliação por gênero ou receita total por ano.

Requisitos Não Funcionais:

- **RNF01:** O banco de dados relacional deve estar totalmente normalizado até a 3ª forma normal;
- **RNF02:** A modelagem do Data Warehouse deve seguir um esquema estrela simplificado;
- **RNF03:** O sistema deve garantir integridade referencial entre as entidades envolvidas.

3. Banco de Dados Relacional (OLTP)

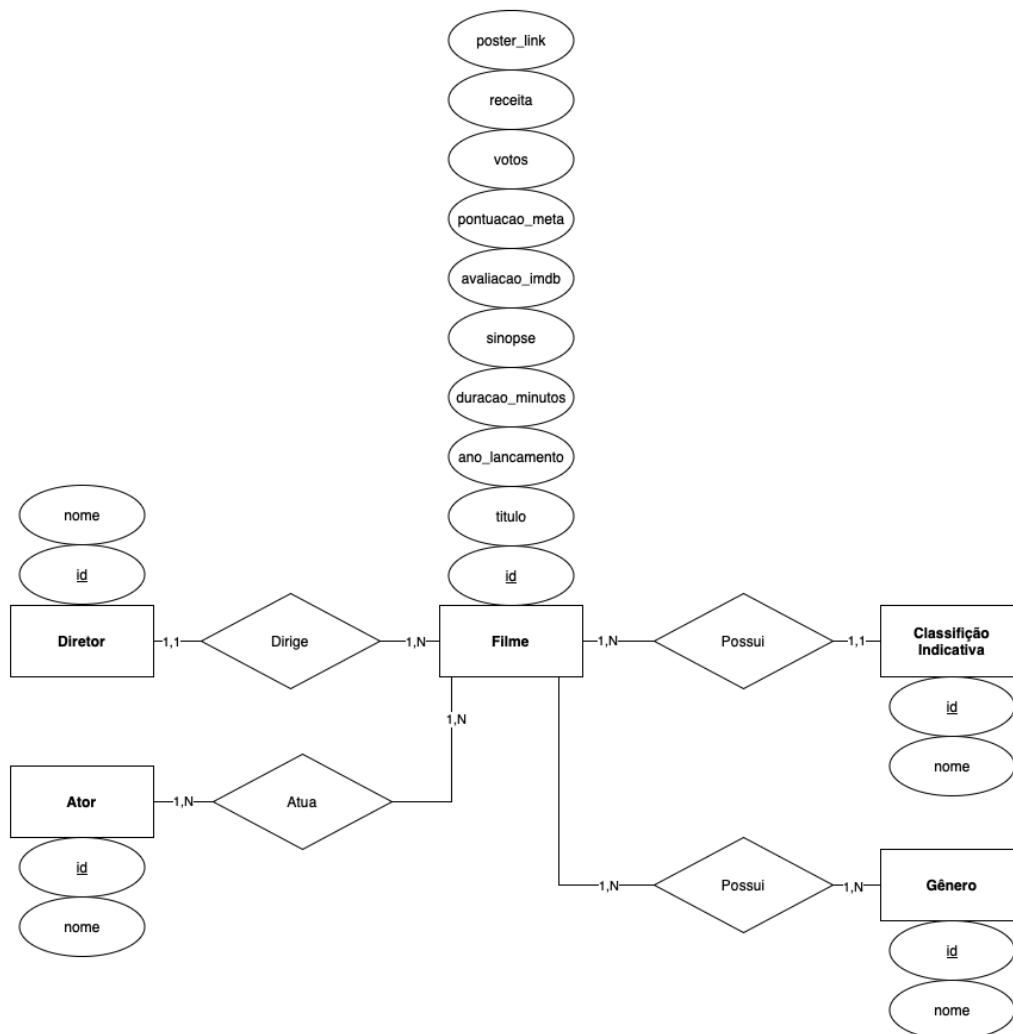
3.1 Diagrama Entidade-Relacionamento (DER)

A modelagem conceitual do banco de dados foi realizada por meio de um Diagrama Entidade-Relacionamento (DER), com o objetivo de representar as principais entidades envolvidas no domínio da indústria cinematográfica, bem como os relacionamentos entre elas.

As entidades identificadas foram: Filme, Diretor, Ator, Gênero e Classificação Indicativa. Para representar os relacionamentos muitos-para-muitos entre filmes e atores, e filmes e gêneros, foram inseridos relacionamentos intermediários, como “Atua” e “Possui”.

A cardinalidade de cada relacionamento foi definida com base na análise do dataset original. Por exemplo, um filme possui um único diretor (1:1), mas pode estar associado a múltiplos gêneros e múltiplos atores (1:N). A classificação indicativa também segue uma relação 1:1 com o filme.

Essa modelagem permitiu compreender a estrutura dos dados e preparar a transição para a modelagem lógica de forma consistente e alinhada com os princípios da normalização.



3.2 Esquema Relacional

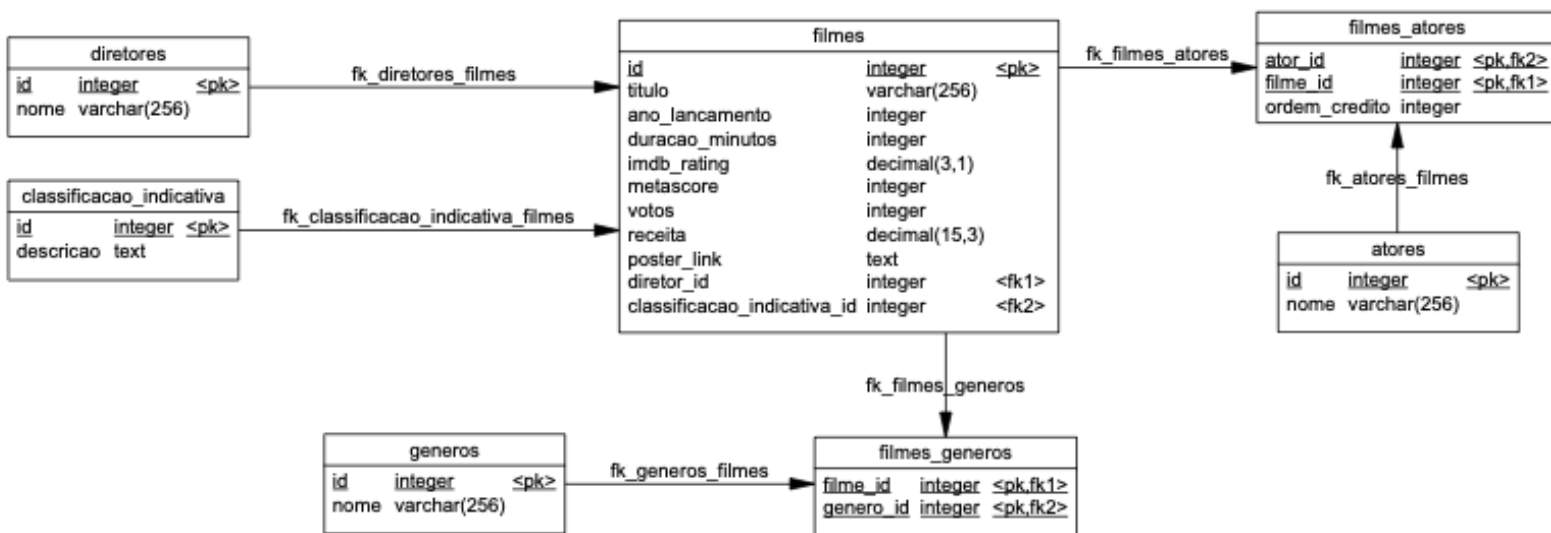
A partir do modelo conceitual, foi elaborado o esquema relacional (modelo lógico) com base nas boas práticas de modelagem de bancos de dados relacionais.

O esquema foi normalizado até a terceira forma normal (3FN), garantindo ausência de redundância e consistência nos dados. As principais tabelas criadas foram:

- filmes – Tabela central, com os atributos descritivos e métricas dos filmes;
- diretores, atores, generos, classificacao_indicativa – Tabelas dimensionais que armazenam dados descritivos associados aos filmes;
- filmes_atores e filmes_generos – Tabelas associativas que representam os relacionamentos muitos-para-muitos.

Todas as relações foram implementadas com chaves primárias e chaves estrangeiras, respeitando a integridade referencial. Além disso, a tipagem dos campos foi definida com base na natureza dos dados (por exemplo, varchar para nomes, decimal para avaliações, integer para identificadores e votos).

A modelagem relacional resultante permitiu realizar operações OLTP eficientes e preparar a base para o processo de extração e transformação dos dados para o Data Warehouse.



3.3 Versão Inicial

Coluna	Tipo
Poster_Link	texto
Series_Title	texto
Released_Year	inteiro
Certificate	texto
Runtime	inteiro
Genre	texto
IMDB_Rating	decimal
Overview	texto
Meta_score	decimal
Director	texto
Star1	texto
Star2	texto
Star3	texto
Star4	texto
No_of_Votes	inteiro
Gross	decimal

3.4 Versão Normalizada

Tabela: classificacao_indicativa

Campo	Tipo
id	PK
descrição	varchar

Tabela: diretores

Campo	Tipo
id	PK
nome	varchar

Tabela: atores

Campo	Tipo
id	PK
nome	varchar

Tabela: generos

Campo	Tipo
id	PK
nome	varchar

Tabela: filmes

Campo	Tipo
id	PK
título	varchar
ano_lancamento	inteiro
duracao_minutos	inteiro
sinopse	texto
avaliacao_imdb	decimal
pontuacao_meta	inteiro
votos	inteiro
receita	decimal
poster_link	texto

Tabela: filmes_atores

Campo	Tipo
filme_id	PK, FK
ator_id	PK, FK
ordem_credito	integer

Tabela: filmes_generos

Campo	Tipo
filme_id	PK, FK
genero_id	PK, FK

3.5 Justificativas da normalização

Forma Normal	Justificativa
1FN	Eliminação de grupos repetitivos (gênero, atores).
2FN	Todos os campos dependem da PK.
3FN	Nenhuma dependência transitiva.

3.6 Dicionário de Dados Completo

Tabela: classificacao_indicativa

- **id**
 - **Tipo de dado:** PK (Primary Key) / SERIAL
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Chave primária de identificação única para cada classificação indicativa.
- **descricao**
 - **Tipo de dado:** VARCHAR(100)
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Descrição da classificação indicativa (ex.: "Livre", "12 anos", "PG-13").

Tabela: diretores

- **id**
 - **Tipo de dado:** PK (Primary Key) / SERIAL
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Chave primária de identificação única para cada diretor.
- **nome**
 - **Tipo de dado:** VARCHAR(256)
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Nome completo do diretor.

Tabela: atores

- **id**
 - **Tipo de dado:** PK (Primary Key) / SERIAL
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Chave primária de identificação única para cada ator/atriz.
- **nome**
 - **Tipo de dado:** VARCHAR(256)
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Nome completo do ator/atriz.

Tabela: generos

- **id**
 - **Tipo de dado:** PK (Primary Key) / SERIAL
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Chave primária de identificação única para cada gênero.
- **nome**
 - **Tipo de dado:** VARCHAR(256)
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Nome do gênero do filme (ex.: "Ação", "Comédia", "Drama").

Tabela: filmes

- **id**
 - **Tipo de dado:** PK (Primary Key) / SERIAL
 - **Restrições:** NOT NULL, UNIQUE
 - **Descrição:** Chave primária de identificação única do filme.
- **titulo**
 - **Tipo de dado:** VARCHAR(256)
 - **Restrições:** NOT NULL
 - **Descrição:** Título original do filme.
- **ano_lancamento**

- **Tipo de dado:** INTEGER
 - **Restrições:** NULL
 - **Descrição:** Ano em que o filme foi lançado.
- **duracao_minutos**
 - **Tipo de dado:** INTEGER
 - **Restrições:** NULL
 - **Descrição:** Duração total do filme em minutos.
- **sinopse**
 - **Tipo de dado:** TEXT
 - **Restrições:** NULL
 - **Descrição:** Pequeno resumo ou descrição da trama do filme.
- **avaliacao_imdb**
 - **Tipo de dado:** DECIMAL(3,1)
 - **Restrições:** NULL
 - **Descrição:** Classificação média do filme no IMDb (Internet Movie Database).
- **pontuacao_meta**
 - **Tipo de dado:** INTEGER
 - **Restrições:** NULL
 - **Descrição:** Pontuação do filme baseada em avaliações de críticos, compilada pelo Metacritic.
- **votos**
 - **Tipo de dado:** INTEGER
 - **Restrições:** NULL
 - **Descrição:** Número de votos ou avaliações que o filme recebeu no IMDb.
- **receita**
 - **Tipo de dado:** DECIMAL(15,2)
 - **Restrições:** NULL
 - **Descrição:** Receita de bilheteria do filme.
- **poster_link**
 - **Tipo de dado:** TEXT
 - **Restrições:** NULL
 - **Descrição:** URL do link para o pôster promocional do filme.
- **diretor_id**
 - **Tipo de dado:** FK (Foreign Key) / INTEGER
 - **Restrições:** NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela diretores, indicando o diretor principal do filme.
- **classificacao_indicativa_id**
 - **Tipo de dado:** FK (Foreign Key) / INTEGER
 - **Restrições:** NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela classificacao_indicativa, indicando a classificação etária do filme.

Tabela: filmes_atores (Tabela de Junção N:N)

- **filme_id**
 - **Tipo de dado:** PK (Parte da Chave Primária Composta), FK (Foreign Key) / INTEGER
 - **Restrições:** NOT NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela filmes, ligando o filme ao ator.
- **ator_id**
 - **Tipo de dado:** PK (Parte da Chave Primária Composta), FK (Foreign Key) / INTEGER
 - **Restrições:** NOT NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela atores, ligando o ator ao filme.
- **ordem_credito**
 - **Tipo de dado:** INTEGER
 - **Restrições:** NULL
 - **Descrição:** Indica a ordem em que o ator aparece nos créditos do filme (ex.: 1 para ator principal).

Tabela: filmes_generos (Tabela de Junção N:N)

- **filme_id**
 - **Tipo de dado:** PK (Parte da Chave Primária Composta), FK (Foreign Key) / INTEGER
 - **Restrições:** NOT NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela filmes, ligando o filme ao gênero.
- **genero_id**
 - **Tipo de dado:** PK (Parte da Chave Primária Composta), FK (Foreign Key) / INTEGER
 - **Restrições:** NOT NULL
 - **Descrição:** Chave estrangeira que referencia o id da tabela generos, ligando o gênero ao filme.

4. Data Warehouse (OLAP)

4.1. Processo de negócio analisado

O processo de negócio analisado no contexto deste projeto foi definido como análise de popularidade e desempenho de obras audiovisuais. Este processo visa mensurar, organizar e disponibilizar dados analíticos relacionados a filmes e séries de televisão, permitindo a avaliação do impacto das produções a partir de indicadores como avaliações do público, crítica especializada, número de votos e receita de bilheteria.

Diferente de processos típicos de vendas ou empréstimos, o foco está em compreender tendências de consumo, identificar os filmes mais bem avaliados por gênero ou diretor, e oferecer uma base consistente para exploração de dados históricos da indústria audiovisual. Essa análise pode servir como base para decisões estratégicas em áreas como produção, distribuição e marketing.

4.2. Identificação de Fatos e Dimensões

A estrutura do Data Warehouse segue um modelo estrela, onde foi definida uma tabela fato central e um conjunto de dimensões relacionadas diretamente a ela.

4.2.1 Tabela fato

A tabela fato principal foi nomeada `fato_avaliacao_filme`, e armazena as métricas quantitativas relacionadas ao desempenho de cada filme. As principais métricas (fatos) consideradas são:

- `avaliacao_imdb`: Nota média atribuída pelos usuários no IMDb;
- `metascore`: Pontuação atribuída pela crítica especializada (Metacritic);
- `quantidade_de_votos`: Número total de votos recebidos no IMDb;
- `receita`: Receita bruta estimada do filme.

4.2.2 Dimensões (contextos de análise):

- `dim_filme`: Armazena os dados principais de identificação dos filmes, como título, duração, sinopse e link do pôster.
- `dim_tempo`: Utilizada para representar o tempo de lançamento da obra, começando pelo nível anual (ano) e podendo ser expandida com mês e dia, se necessário.
- `dim_diretor`: Representa os diretores responsáveis pelas obras, permitindo análise de desempenho por direção.
- `dim_classificacao_indicativa`: Define a faixa etária recomendada de cada filme (como PG, R, etc.), sendo útil para análises demográficas.

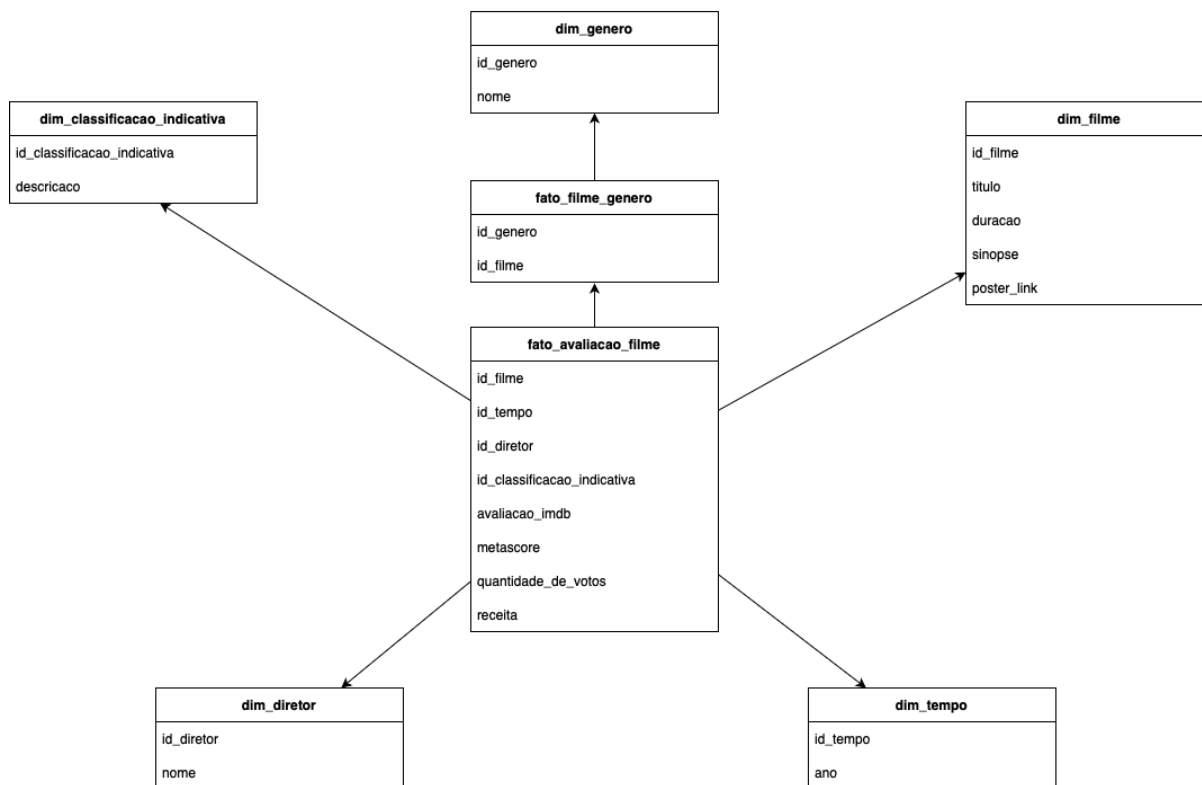
- **dim_genero:** Representa os gêneros dos filmes (ação, drama, comédia, etc.). Como um filme pode ter múltiplos gêneros, foi criada uma tabela ponte chamada **filme_genero_dw**, responsável por armazenar os relacionamentos N:N entre filmes e gêneros.

4.2.3 Granularidade da Tabela Fato

O nível de granularidade adotado para a tabela fato é um registro por filme e ano de lançamento, ou seja, cada linha da tabela representa as métricas consolidadas de um filme em um determinado ponto no tempo.

4.3. Diagrama do Esquema Estrela

O modelo estrela foi desenhado com a **fato_avaliacao_filme** ao centro e suas dimensões conectadas diretamente. A exceção é o relacionamento entre filmes e gêneros, que é intermediado pela tabela de junção **filme_genero_dw**, responsável por associar múltiplos gêneros a cada filme.



Esse modelo permite realizar consultas analíticas com alto desempenho e flexibilidade, respondendo a perguntas como:

- Qual a média de avaliações por gênero?
- Quais diretores possuem os filmes mais bem avaliados?
- Quais anos registraram maior receita de bilheteria?
- Qual a classificação indicativa com maior número de filmes populares?

5. Decisões de Projeto

5.1. Principais escolhas técnicas

Durante o desenvolvimento do projeto, foram realizadas diversas escolhas técnicas com o objetivo de garantir a clareza da modelagem, a integridade dos dados, a escalabilidade das consultas e a aderência às boas práticas de banco de dados. As principais decisões adotadas foram:

- **Escolha do domínio de negócio (IMDb):** A seleção do dataset do IMDb foi motivada pela riqueza de informações e variedade de atributos presentes, permitindo a exploração de um cenário realista com dados suficientemente complexos para representar diferentes entidades e relacionamentos. O dataset contém dados como avaliações, sinopses, classificações, diretores, elenco, receita, entre outros.
- **Separação de dados em entidades especializadas:** Com base na análise do dataset, optou-se por desmembrar atributos compostos e repetir valores em entidades específicas. Foram criadas tabelas dedicadas para atores, diretores, gêneros e classificações indicativas, o que promoveu a reutilização de dados e eliminou redundâncias.
- **Modelagem de relacionamentos N:N com tabelas associativas:** Para lidar com os relacionamentos muitos-para-muitos entre filmes e gêneros e entre filmes e atores, foram criadas as tabelas associativas filmes_generos e filmes_atores. A tabela filmes_atores também incorporou o campo ordem_credito, preservando a ordem de aparecimento dos atores, uma informação semântica relevante para o contexto de análise de elenco.
- **Geração automatizada de scripts SQL:** Todos os comandos de criação (DDL) e inserção (DML) foram gerados com auxílio de ferramentas automatizadas, código Python e utilização da ferramenta ChatGPT, garantindo consistência e economia de tempo no processo de importação de dados reais para o banco.

5.2. Justificativas das decisões

As decisões tomadas ao longo do projeto foram orientadas por princípios de engenharia de dados, boas práticas de modelagem relacional e dimensional, e experiências simuladas de cenários reais de sistemas de informação. Entre as principais justificativas:

- A separação das entidades em tabelas independentes facilita a manutenção dos dados, reduz duplicações e garante escalabilidade do sistema.
- A normalização melhora a qualidade dos dados e reduz inconsistências sem comprometer a capacidade analítica, já que o DW foi desenhado paralelamente.
- A modelagem estrela simplificada atende bem ao escopo da disciplina, promovendo aprendizado de conceitos fundamentais de OLAP e facilitando a construção de consultas agregadas.
- O uso do dataset IMDb favoreceu o engajamento do grupo por tratar de um tema atual, relevante e de fácil compreensão, além de conter dados públicos e prontos para exploração.
- A geração automatizada de scripts foi uma decisão pragmática diante da quantidade de dados a serem inseridos, otimizando o tempo do grupo e reduzindo a probabilidade de erro manual.

Essas decisões em conjunto permitiram o desenvolvimento de um sistema robusto, didático e próximo da realidade de aplicações reais de bancos de dados em ambientes corporativos e acadêmicos.

6. Conclusões e Aprendizados

A realização deste projeto permitiu aos discentes vivenciarem o ciclo de desenvolvimento de um banco de dados, desde a coleta e análise de dados reais até a construção de um Data Warehouse voltado à análise multidimensional. A escolha do domínio, utilizando o dataset do IMDb, contribuiu para tornar a experiência mais aplicável a contextos reais de mercado.

Entre os principais aprendizados, destacam-se:

- Compreensão e aplicação das etapas de modelagem de dados, incluindo a criação do modelo conceitual, lógico e físico, além da normalização até a terceira forma normal, assegurando consistência e eliminação de redundâncias.
- Estruturação de dados complexos em um modelo relacional, enfrentando desafios como separação de dados compostos, como múltiplos gêneros e atores, identificação de chaves primárias e estrangeiras, e organização de tabelas associativas.
- Desenvolvimento de habilidades em SQL, tanto na definição da estrutura do banco (DDL) quanto na inserção e manipulação de dados (DML).
- Construção de um modelo de Data Warehouse com esquema estrela, com definição clara de fatos e dimensões, adequado a análises de negócio.

- Utilização consciente de ferramentas de apoio à produtividade, como o ChatGPT para geração de scripts e estruturação da documentação, reforçando a importância da IA como facilitadora no contexto acadêmico e profissional.

Além disso, um aprendizado importante foi adquirido ao trabalhar com o Amazon Redshift, especialmente no que se refere às diferenças no suporte a tipos de dados em comparação com outros SGBDs.

Durante a criação do Data Warehouse, identificamos que o tipo SERIAL, amplamente utilizado para auto incremento em outros bancos, não é compatível com o Redshift. Para contornar essa limitação, utilizamos o tipo IDENTITY, que oferece funcionalidade equivalente para geração automática de chaves primárias.

Outro ponto relevante foi a descoberta de que o tipo TEXT também não é suportado diretamente no Redshift; ao utilizá-lo, o sistema o converte implicitamente para VARCHAR(256). Essa limitação causou erro ao tentar inserir sinopses de filmes com mais de 256 caracteres. A situação exigiu atenção especial à definição dos campos de texto, destacando a importância de validar as especificidades do SGBD escolhido para garantir compatibilidade e integridade dos dados.

7. Referências

IMDb Dataset of Top 1000 Movies and TV Shows. Kaggle. Disponível em:

<https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>. Acesso em: maio de 2025.

OpenAI. **ChatGPT** – Assistente de geração de código, estruturação de dados e auxílio na documentação. Modelo utilizado: GPT-4, acesso via <https://chat.openai.com>. Acesso em: maio de 2025.