Relatório Trabalho Final - Banco de Dados

Pedro Jorge Oliveira Câmara

1 de dezembro de 2023

1 Introdução

O conjunto de dados foi escolhido a partir da comunidade online *Kaggle*, conhecida, principalmente pela comunidade de ciência de dados e machine learning, pela sua vasta quantidade de datasets públicos dos mais variados temas. Os dados em questão foram aquisitados do *Netflix Movies and TV Shows*, que é constituído por uma grande tabela que conta, no momento de escrita desse relatório, com 8807 linhas, cada uma representando um filme ou um programa de TV (série, anime, etc), e 12 colunas, que descrevem as características da mídia em questão, como título e pessoas que participaram da sua construção. Essa base é uma das mais conhecidas do site, contando com mais de 2 milhões de visualizações e 393 mil downloads para os mais diversos fins de aplicações e estudos. Neste trabalho, visamos migrar de uma única tabela "bruta", que conta com diversas informações, mas que é muito densa e compacta, para um esquema relacional, que tenta, da melhor maneira possível, destrinchar e esquematizar esses dados de forma a podermos consultá-los de forma mais orgânica.

2 Datasets

O dataset escolhido é baixado no formado *.csv*, populado pelos mais diversos conteúdos presentes no serviço de streaming da Netflix. A informação do site Kaggle é que a última atualização dessa base de dados foi em 2021, então os conteúdos mais recentes, que foram adicionados ou removidos nos últimos dois anos, não estão necessariamente nesse presentes catálogo. Cada coluna é descrita e explicada a seguir.

2.1 *show_id*: string

É a coluna identificadora de cada mídia do dataset. Os valores brutos, antes de qualquer manipulação após o download, vêm com "s1", "s2", ..., "s_8807". Como primeira mudança, alteramos essa coluna para ter valores inteiros 1, 2, ..., 8807.

2.2 *type: string*

Descreve se a mídia em questão é um filme ou programa de TV, assumindo os valores "Movie"e "TV Show".

2.3 *title: string*

Descreve o título da mídia em questão, como, por exemplo, "Blood & Water".

2.4 *director: string*

Uma string única que é constituída por diversos nomes de diretores, separados por vírgula, como, por exemplo, "Pedro Jorge, Sidarta". Os valores podem ter somente um nome ou vários nomes.

2.5 *cast: string*

Uma string única que é constituída por diversos nomes de atores e atrizes, separados por vírgula, como, por exemplo, "Pedro Jorge, Sidarta". Os valores normalmente têm vários nomes, mas é possível que, em alguma linha, haja somente um.

2.6 *country: string*

O nome do país do qual a mídia em questão é originária e foi produzida, exemplos são "United States", "South Africa", "India".

2.7 *date_added: string*

Uma string que indica a data quando a mídia em questão foi adicionada à Netflix, no formato "Mês (discursivo) dia, ano", como, por exemplo, "September 25, 2021".

2.8 *release_year: int*

Ano em que a mídia em questão foi produzida, como, por exemplo, 2020.

2.9 *rating: string*

A classificação indicativa da mídia em questão, como, por exemplo, "TV-MA"(para adultos), "PG-13"(acima de 13 anos).

2.10 *duration: string*

Duração da mídia em questão, que pode ser em minutos ou temporadas, como, por exemplo, "90 min"ou "2 Seasons".

2.11 *listed_in: string*

Uma string única que é constituída pelos gêneros aos quais a mídia em questão se encaixa, como, por exemplo, "International TV, Shows, TV Dramas, TV Mysteries"(uma única string). Pode ter apenas um ou vários gêneros em uma mesma string nesta coluna.

2.12 *description: string*

Uma breve descrição ou sinopse da mídia em questão.

A partir dessas informações, construímos nosso modelo relacional, visando uma melhor visualização e entendimento dos dados, ilustrado na figura 1 (final do relatório).

3 Projeto do Banco de Dados

A percepção imediata que tivemos ao analisar a estrutura do dataset é a densidade de informação condensada em pouquíssimas colunas. Apesar de ser uma matriz 8807x12, conseguimos separar as informações em 12 tabelas distintas para relacioná-las melhor. Para a limpeza e transposição do formato "bruto"dos dados para o formato relacional e do MySQL, utilizamos a biblioteca Pandas tanto para análise da composição desse dataset, quanto para a filtragem, manipulação e "dissipação"desses dados em tabelas diferentes. Por exemplo, para as tabelas de diretores e de atores, tivemos que manipular as strings de forma a obter os nomes de cada pessoa separadamente e armazená-las em tabelas apropriadas (para isso, criamos diversmos DataFrames isolados no Pandas, cada um para um dado do nosso banco). Durante essa etapa de limpeza, salvamos cada tabela criada em um .csv, visando passá-las, posteriormente, para o MySQL. Para inserção no banco, utilizamos a biblioteca MySQL Connector do Python, iterando por todos os arquivos .csv criados anteriormente e adicionando no banco no seu devido lugar. O modelo lógico está ilustrado na figura 2 (final do relatório).

3.1 content

A tabela *content* é a principal do nosso banco, ela é a que condensa grande parte das informações pertinentes, como o título e a descrição/sinopse de uma mídia em questão, sendo constituída por:

- id: int - Chave primária
- title: string - Título da mídia
- date_added: date - Data em que foi adicionada à Netflix
- description - text: Descrição/sinopse da mídia

3.2 director

A tabela *director* é a que armazena os identificadores numéricos e os nomes de todos os diretores registrados nesse dataset, sendo constituído por:

- id: int - Chave primária
- name: string - Nome do diretor

3.3 content_director

Tabela que relaciona um *content* ao(s) seu(s) diretor(res) -repare que uma mesma mídia pode ter vários diretores; ela é constituída por:

- content_id: int - Identificador da mídia
- director_id: id - Identificador do diretor

3.4 rating

A tabela *rating* é a que armazena os identificadores numéricos e os nomes de todas as classificações indicativas registradas nesse dataset, sendo constituída por:

- id: int - Chave primária
- name: string - Nome da classificação indicativa

3.5 content_rating

Tabela que relaciona um *content* à sua classificação indicativa; ela é constituída por:

- content_id: int - Identificador da mídia
- rating_id: id - Identificador da classificação indicativa

3.6 type

Tabela que armazena os dois tipos de mídia nesse contexto: *Movie* e *TV Show*. É formada somente por esses dois valores, neste formado:

- id: int - Identificador do tipo (1 ou 2)
- name: string - Tipo (Movie ou TV Show)

3.7 content_type

Tabela que relaciona um *content* ao seu tipo -repare que uma mídia só pode ser apenas Movie ou apenas TV Show; ela é constituída por:

- content_id: int - Identificador da mídia
- type_id: id - Identificador do tipo (1 ou 2)

3.8 genre

A tabela *genre* é a que armazena os identificadores numéricos e os nomes de todos os gêneros registrados nesse dataset, sendo constituído por:

- id: int - Chave primária
- name: string - Nome do gênero

3.9 content_genre

Tabela que relaciona um *content* ao(s) seu(s) gêneros -repare que uma mesma mídia pode ter vários gêneros; ela é constituída por:

- content_id: int - Identificador da mídia
- genre_id: id - Identificador do gênero

3.10 actor

A tabela *actor* é a que armazena os identificadores numéricos e os nomes de todos os atores e atrizes registrados nesse dataset, sendo constituído por:

- id: int - Chave primária
- name: string - Nome do ator ou da atriz

3.11 content_cast

Tabela que relaciona um *content* ao(s) seu(s) ator(res) e atriz(es) - repare que uma mesma mídia pode ter vários atores e atrizes; ela é constituída por:

- content_id: int - Identificador da mídia
- actor_id: id - Identificador do ator ou atriz

3.12 movie_duration

Tabela que relaciona um FILME, especificamente, à sua duração. Ela é constituída por:

- content_id: int - Identificador da mídia que é um filme
- minutes: int - Quantidade de minutos de duração

3.13 tv_show_duration

Tabela que relaciona um PROGRAMA DE TV, especificamente, à sua duração. Ela é constituída por:

- content_id: int - Identificador da mídia que é um filme
- seasons: int - Quantidade de temporadas de duração

3.14 Consultas ao banco

- *SELECT title FROM content LIMIT 5;* Consultamos o título dos 5 primeiros filmes que estão registrados no banco de dados da Netflix.
- *SELECT title, date_added FROM content WHERE title like "{}%";* Com o nome de um filme ou TV show passado como parâmetro, verificamos se ele está disponível na Netflix e em qual data ele foi adicionado ao catálogo.

- *SELECT title FROM content WHERE id in (SELECT content_id FROM director INNER JOIN content_director ON id=director_id WHERE director_name like "%{}%") AND id in (SELECT content_id FROM content_type INNER JOIN type ON id=type_id WHERE type_name like "Movie");* Com o nome de um diretor passado como parâmetro, consultamos quantos filmes ele dirigiu

- *SELECT title FROM content INNER JOIN tv_show_duration ON id=content_id WHERE id in (SELECT content_id FROM content_type INNER JOIN type ON id=type_id WHERE type_name like "Tv Show") AND id in (SELECT content_id FROM content_rating INNER JOIN rating ON id=rating_id AND name like "TV-MA") AND seasons 8;* Consultamos os TV shows que possuem classificação indicativa para adultos e que tenham mais do que 8 temporadas na Netflix.

- *SELECT type_name AS "Tipo de Mídia", COUNT(type_id) AS "Quantidade"FROM content_type INNER JOIN type ON id=type_id GROUP BY(type_id);* Consultamos a quantidade de filmes e a quantidade de TV shows disponíveis no catálogo.

- *SELECT actor_name AS "Ator ou atriz", count(id) AS "Atuações"FROM actor INNER JOIN content_cast ON id=actor_id GROUP BY actor_name HAVING count(id) 25;* Consultamos quais atores possuem participações atuando em mais do que 25 filmes.

- *SELECT AVG(actor_count) AS "Média de atores nos filmes"FROM (SELECT COUNT(actor_id) AS actor_count FROM content_cast WHERE content_id IN (SELECT id FROM content WHERE id IN (SELECT content_id FROM content_type WHERE type_id = 1)) GROUP BY content_id) AS actor_counts;* Consultamos a quantidade média de atores (registrados como atuantes na base de dados) em cada filme.

- *SELECT genre_name AS 'Gêneros de {}' FROM content RIGHT JOIN content_genre on content.id=content_id RIGHT JOIN genre ON genre.id=genre_id WHERE title like '{}';* Passado um filme ou TV Show como parâmetro, consultamos quais os gêneros relacionados com aquela mídia.

- *SELECT title FROM content INNER JOIN content_genre ON id=content_id WHERE genre_id in (SELECT id FROM genre WHERE genre_name like '%{}%');* Dado um gênero de interesse, consultamos todos os filmes e TV shows relacionados àquele gênero.

4 Aplicação

Pelos motivos explicados, as consultas foram visualizadas e feitas em um Jupyter Notebook, utilizando a linguagem Python e as bibliotecas MySQL.Connector e Pandas. O MySQL.Connector é o responsável pela comunicação com o banco, enquanto que o Pandas converte as consultas para uma pequena tabela para visualizarmos.

5 Distribuição do trabalho

- Pedro Jorge: tudo.

6 Considerações finais

- Como manipular dados "brutos": uma das dificuldades iniciais foi em como manipular um arquivo *.csv* de forma a filtrar e separar as informações nele contidas para um esquema relacional. Tanto na parte "lógica", de analisar a estrutura da tabela e separá-la em diversas tabelas que se relacionam entre si, quanto na prática, utilizando o Pandas para essa manipulação.

- Verificação da integridade dos dados: após toda essa filtragem e tradução para um esquema relacional, uma parte importante foi verificar se os dados estão íntegros, se as relações fazem sentido, se conseguimos obter informações das relações tanto quanto da tabela "bruta"original.

- Atenção aos detalhes: quando estamos trabalhando com um banco de dados relativamente robusto e real, alguns detalhes se tornam importantes. Por exemplo, em diversos momentos tive que alterar o tamanho de algumas variáveis do tipo VARCHAR por motivo de, na hora da inserção de dados no MySQL, haver algum dado específico que era diferente do que esperado. Um exemplo desse acontecido era o tamanho VARCHAR(100) para o título de uma mídia mas, no momento da inserção, havia um título que tinha tamanho maior que 100, o que me levou a alterar o tamanho máximo de um título para 500.
- Visualização: manipular uma base de dados realística torna mais compreensível a visualização das relações e distribuição das tabelas, até mesmo as consultas em si tornam-se relativamente intuitivas, principalmente as junções.

Referências

https://www.kaggle.com/datasets/shivamb/netflix-shows

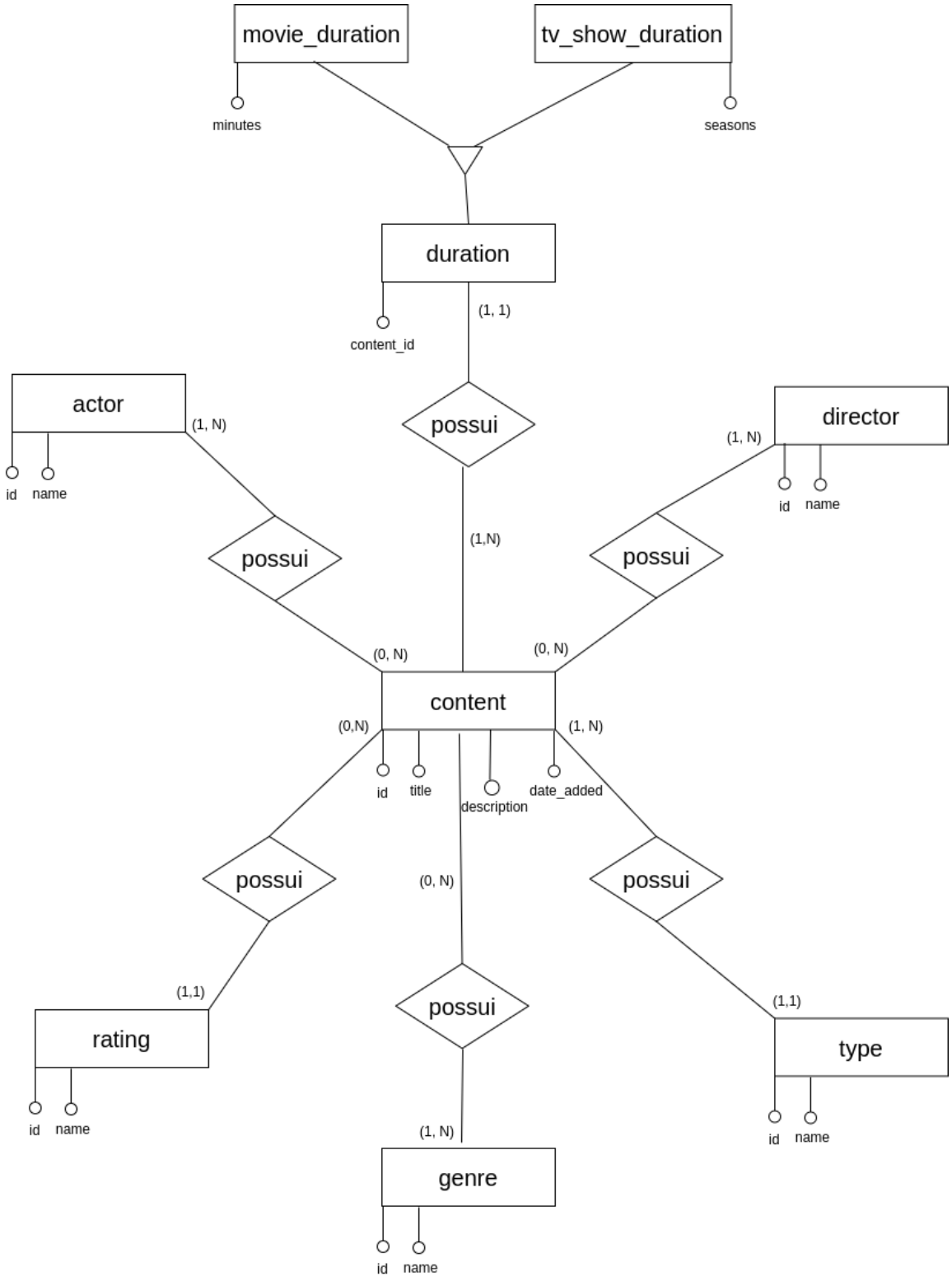

Figura 1: Modelo conceitual

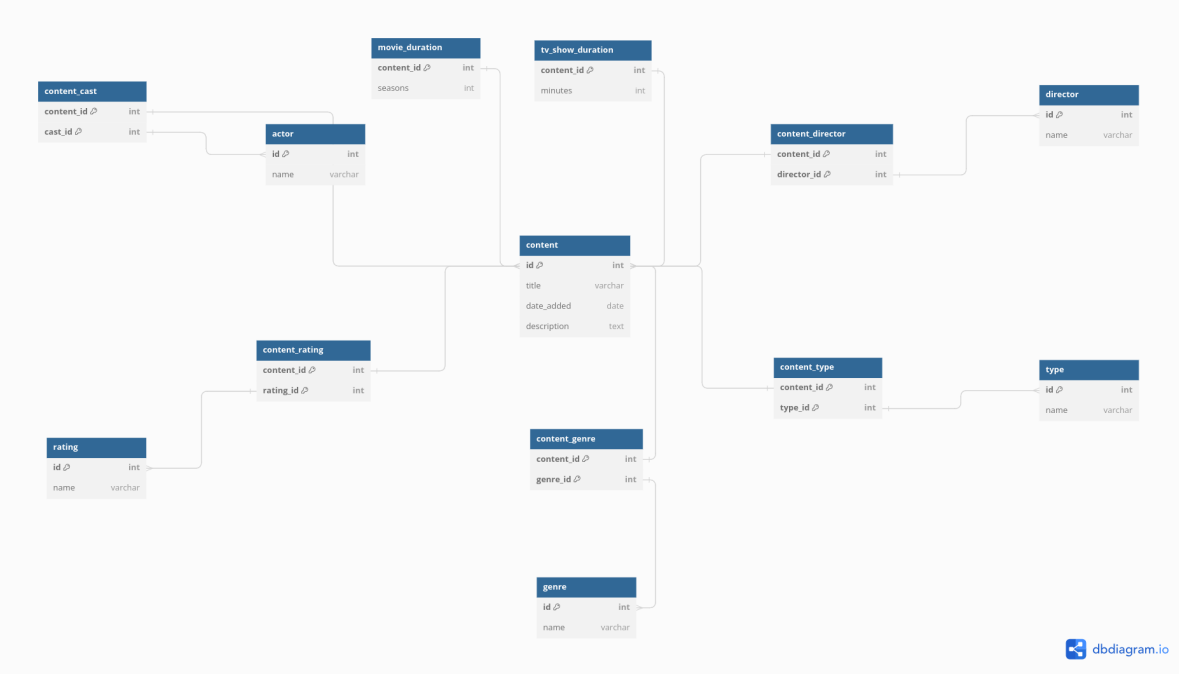

Figura 2: Modelo lógico