Apresentação
Prévia do Projeto Final
MC536

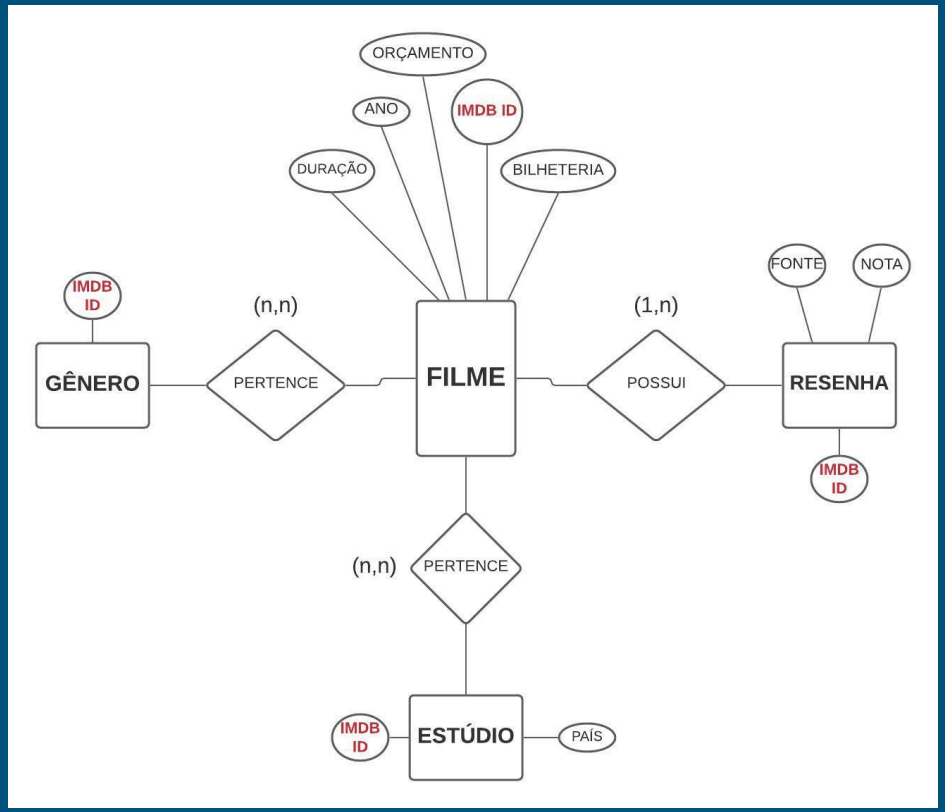
Daniel Credico de Coimbra – 155077
Gabriel Bonfim Silva de Moraes – 216111
Victor Durço Gomes Bijos – 206508

Resumo do Projeto

Dataset que integra informações de diversas fontes sobre os 25 filmes com maior bilheteria de cada um dos últimos 50 anos.

Cada filme está relacionado a gêneros e estúdios, além de ser informado características como ano, país de origem, avaliação crítica (IMDb e Metacritic), bilheteria, número de bilhetes vendidos, e código IMDb.

Modelo Conceitual Preliminar



Modelos Lógicos Preliminares

MODELO RELACIONAL

- FILMES: (IMDb ID, filme, ano, orçamento, bilheteria, duração)
- RESENHA-FILME: (IMDb ID, fonte, nota)
- ESTÚDIO-FILME: (IMDb ID, estúdio, país)
- GÊNERO-FILME: (IMDb ID, gênero)

MODELO DE GRAFOS

Nódulo: filme (imdb_id: int, título: str, ano: int, bilheteria: int, número_ingressos_vendidos: int)
Nódulo: resenha (fonte: str, nota: float).
Nódulo: estúdio (nome: str, país: str).
Nódulo: gênero (nome: str).
Relação: possui (resenha × filme).
Relação: pertence (estúdio × filme).
Relação: pertence (gênero × filme).

MODELO HIERÁRQUICO

```
{
  imdb_id,
  título,
  ano,
  bilheteria,
  número_ingressos_vendidos,
  resenhas: {
        fonte,
         nota
   },
  estúdios: {
         estúdio,
         país
  },
  gêneros: {
        nome_gênero
  }
}
```

Dataset Preliminar

Tabela única contendo uma linha por cada filme no nosso recorte, informando: código IMDb, título, ano, bilheteria, número de ingressos vendidos, avaliação IMDb, avaliação Metacritic, e seus gêneros.

	imdb_id	imdb_rating	title	year	genres	boxOffice	numTickets
1	tt1074638	7.8	Skyfall	2012	Action,Adventure,Thriller	296804366	37286980
2	tt1078912	6	Night at the Museum: Battle of the Smithsonian	2009	Adventure,Comedy,Family	177243721	23632496
3	tt10954652	5.8	Old	2021	Drama,Horror,Mystery	48242510	5266649
4	tt1104001	6.8	TRON: Legacy	2010	Action,Adventure,Sci-Fi	131304844	16641932
5	tt1080016	6.9	Ice Age: Dawn of the Dinosaurs	2009	Adventure,Animation,Comedy	196573705	26209827
6	tt1099212	5.2	Twilight	2008	Drama,Fantasy,Romance	176922850	24641065
7	tt1133985	5.5	Green Lantern	2011	Action,Adventure,Sci-Fi	116601172	14703805
8	tt1155076	6.2	The Karate Kid	2010	Action,Drama,Family	176591618	22381701
9	tt1170358	7.8	The Hobbit: The Desolation of Smaug	2013	Adventure,Fantasy	228934309	28159201
10	tt1192628	7.2	Rango	2011	Adventure,Animation,Comedy	123477607	15570947
11	tt1194173	6.6	The Bourne Legacy	2012	Action,Adventure,Thriller	113203870	14221591
12	tt1259571	4.8	The Twilight Saga: New Moon	2009	Adventure,Drama,Fantasy	287954655	38393954

Bases de Dados

Título de Base	*Link*	*Breve Descrição*
The Movie Database	https://www.themoviedb.org/ (API)	O Movie Database (TMDB) é um banco de dados popular e editável pelo usuário para filmes e programas de TV.
IMDB Database	https://www.imdb.com/ (API)	API oficial do IMDB (com grandes restrições de uso)
Metacritic (site)	https://www.metacritic.com/ (Web scraping)	Popular site para reviews de jogos, filmes e séries
The Numbers (site)	https://www.the-numbers.com/ (Web scraping)	Site com útil serviço de dados financeiros sobre filmes.

Operações Realizadas

- Os scripts *TheNumbers.py* e *Metacritic.py* foram utilizados para realizar webscraping e obter dados, sendo o *TheNumbers.py* utilziado para as bilheterias e o *Metacritic.py* para as avaliações.
- Utilizamos um adaptador Python de SQLite3 para converter arquivos .tsv (tab-separated values) obtidos de uma *API limitada* do IMDB em um arquivo SQL em formato .db.
- O software SQL Server foi utilizado para importar arquivos e transformá-los em tabelas SQL. Através de operações JOIN, criamos novas tabelas que consolida os dados encontrados nos arquivos CSV e TSV

Pergunta 1 - Modelo de Grafos

- Quais estúdios mais presentes na produção de filmes de alta bilheteria nos últimos 50 anos?

Através do modelo de grafos, podemos lidar com relacionamentos entre estúdios e filmes. Em específico, a relação (estúdio) -[:produz]-> filme pode nos mostrar a quantidade de filmes produzidos por um estúdio. Assim, podemos contar a quantidade de relações de produção em um nódulo (estúdio) e descobrir os estúdios mais presentes.

```
MATCH (a)-[:produz]->(b)
RETURN a, COLLECT(a) as productors
ORDER BY SIZE(productors) DESC LIMIT
10
```

Exemplo de Cypher Query

Pergunta 2 - Modelo Hierárquico

- Quais filmes possuem maior colaboração internacional?

Na constituição dos dados do nosso dataset cruzamos informações vindas da api do "The Movie Database", do qual obtemos json com os estúdios que participaram na produção de cada filme e o país de origem de cada estúdio. Dessa forma será possível percorrer os objetos dos filmes e quantificar quais filmes possuem uma maior quantidade de países diferentes dentro da propriedade estúdios de produção.

Pergunta 3 - Modelo Relacional

- Qual a evolução temporal da bilheteria nominal média dos filmes de maior sucesso ao longo dos anos?

Com o modelo relacional é fácil obter dados para essa questão. Em primeiro lugar, basta adquirir uma média da bilheteria nominal, depois agrupar todos os filmes conforme o ano em que foram lançados. Enfim, ordená-los pelo ano para observar a mudança no valor da média.

```
SELECT AVG(boxOffice), year
FROM movie_table
GROUP BY YEAR
ORDER BY year  DESC;
```

Exemplo de SQL Query