

MVP3 - Engenharia de Dados PUC-RIO

Alex Amorim Faia

Definição do Problema

O IMDb (Internet Movie Database) é um banco de dados on-line de informações de filmes e séries de televisão, que podem ser avaliados pelo usuário com notas..

Neste projeto, o objetivo é hipoteticamente produzir combinações para um filme de sucesso.: Após analisar o banco de dados, vamos elencar:

- A. Quais os 10 filmes melhores avaliados da ultima década.

(atualização: Não foi possível avançar nos objetivos seguinte pelas dificuldades apresentadas durante a execução)

- B. Quais atores estiveram nestes filmes
- C. Quais diretores têm o melhor desempenho
- D. Descobrir qual o gênero com a maior nota
- E. E identificar a duração média dos filmes com maior nota

Etapas do projeto

1. Fonte de dados
2. Modelagem de dados
3. Ingestão de dados
4. Extração
5. Transformação de dados
6. Carga de dados
7. Resultados e Visualização de dados

Fonte dos dados:

IMDb Non-Commercial Datasets (disponível em <https://datasets.imdbws.com/>).

Relação dos arquivos:

- 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 são compactados (GZ) com valores separados por tabulação (TSV) formatados no conjunto de caracteres UTF-8.

Modelagem de dados

Para o primeiro objetivo, iremos trabalhar com as tabelas "title.principals" e "title.ratings"

title.basics.tsv.gz

Contém as seguintes informações para filmes:

Coluna	Descrição
tconst (string)	Identificador alfanumérico exclusivo do título.
titleType (string)	O tipo/formato do título (por exemplo, filme, curta, série de TV, episódio de TV, vídeo).
primaryTitle (string)	O título mais popular; o título usado pelos cineastas em materiais promocionais no momento do lançamento.
originalTitle (string)	Título original, no idioma original.
isAdult (booleano)	0: título não adulto; 1: título adulto.
startYear (YYYY)	Representa o ano de lançamento de um título. No caso de séries de tv, é o ano de início da série.
endYear (YYYY)	Representa o ano final da série de TV. "N" para todos os outros tipos de títulos
runtimeMinutes	Tempo de execução principal do título, em minutos.
genres (array de strings)	Inclui até três gêneros associados ao título.

title.ratings.tsv.gz

Contém a classificação da IMDb e informações de votos para títulos:

Coluna	Descrição
tconst (string)	identificador alfanumérico exclusivo do título.
AverageRating	média ponderada de todas as avaliações individuais dos usuários
numVotes	número de votos que o título recebeu.

Ingestão de dados

Trabalhei com o Microsoft Azure pela primeira vez tendo contato com esta ferramenta e com bastante dificuldade na curva de aprendizagem.

Relato aqui dificuldades com conexões, configurações e falta de tutorial dentro do curso e fora para conseguir desempenhar o esperado

Extração

Recursos utilizados no MS Azure:

Resources

Recent Favorite

Name	Type
 axfaiasql (axfaiasql/axfaiasql)	SQL database
 AXFAIA	Resource group
 axfaiasql	SQL server
 axfaiastorage	Storage account
 axfaiadatabricks	Azure Databricks Service
 axfaiafactory	Data factory (V2)
 Basic	Subscription

[See all](#)

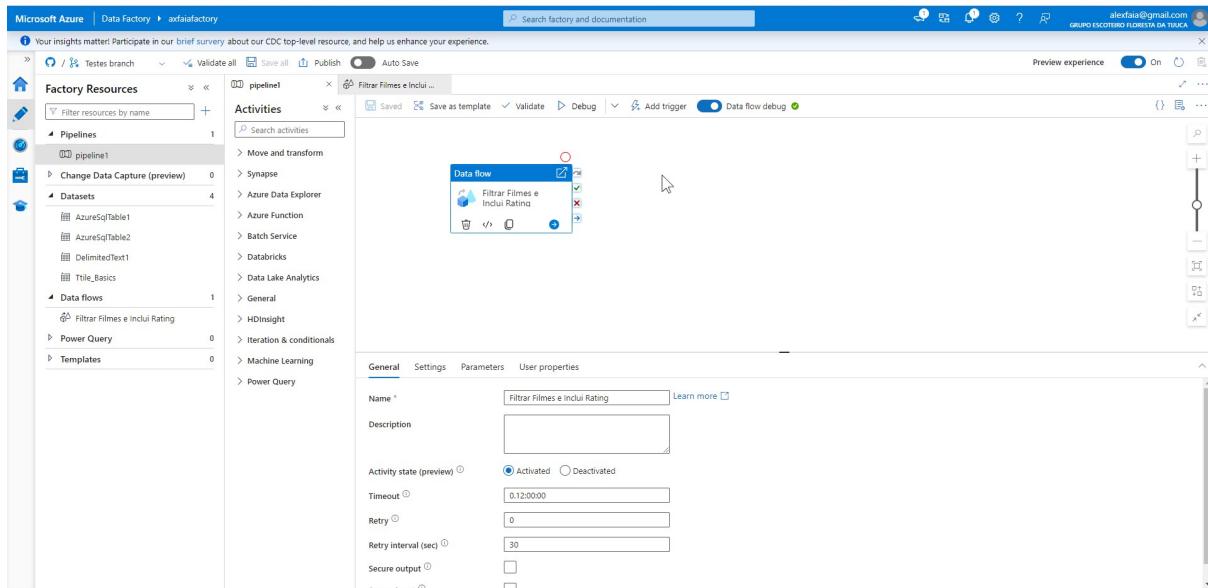
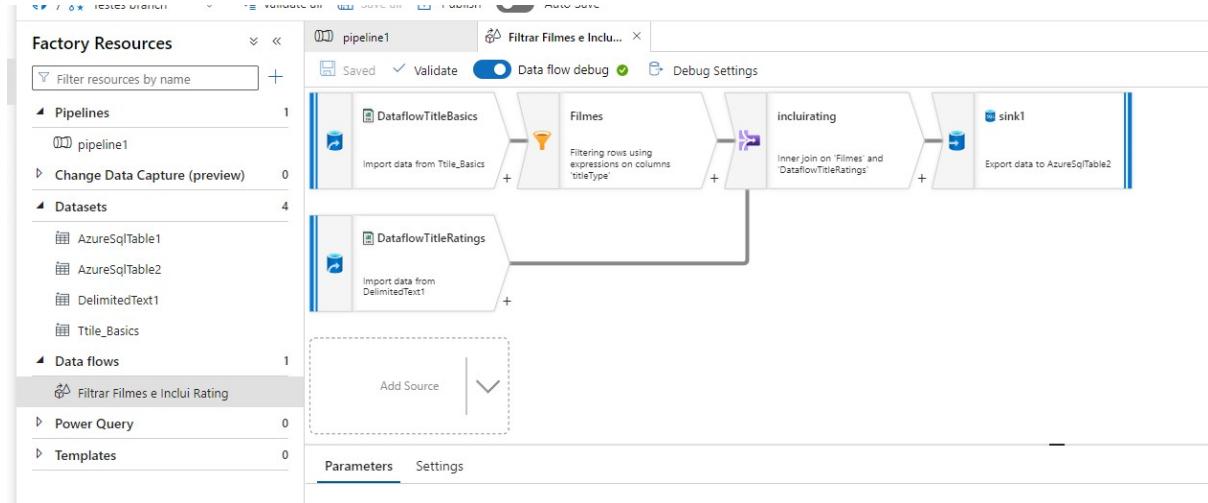
Container utilizado para carga dos arquivos e seus parametros:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
name.basics.tsv.gz	10/1/2023, 10:16:10 AM	Hot (inferred)	Available	Block blob	242.77 MiB	Available
title.akas.tsv.gz	10/1/2023, 10:16:18 AM	Hot (inferred)	Available	Block blob	301.58 MiB	Available
title.basics.tsv.gz	10/1/2023, 10:15:42 AM	Hot (inferred)	Available	Block blob	169.88 MiB	Available
title.crew.tsv.gz	10/1/2023, 10:15:02 AM	Hot (inferred)	Available	Block blob	64.88 MiB	Available
title.episode.tsv.gz	10/1/2023, 10:14:40 AM	Hot (inferred)	Available	Block blob	40.59 MiB	Available
title.principals.tsv.gz	10/1/2023, 10:16:36 AM	Hot (inferred)	Available	Block blob	431.82 MiB	Available
title.ratings.tsv.gz	10/1/2023, 10:15:11 AM	Hot (inferred)	Available	Block blob	6.49 MiB	Available

Transformação de dados

Para transformação, utilizei o Azure Data Fabric:

Dentro do Flow, incluí um filtro para apenas carregar os filmes, descartando séries e curtas.



Carga de dados

Ao final do processo, consegui carregar os dados numa tabela do banco de dados criado no sql serve dentro do azure.

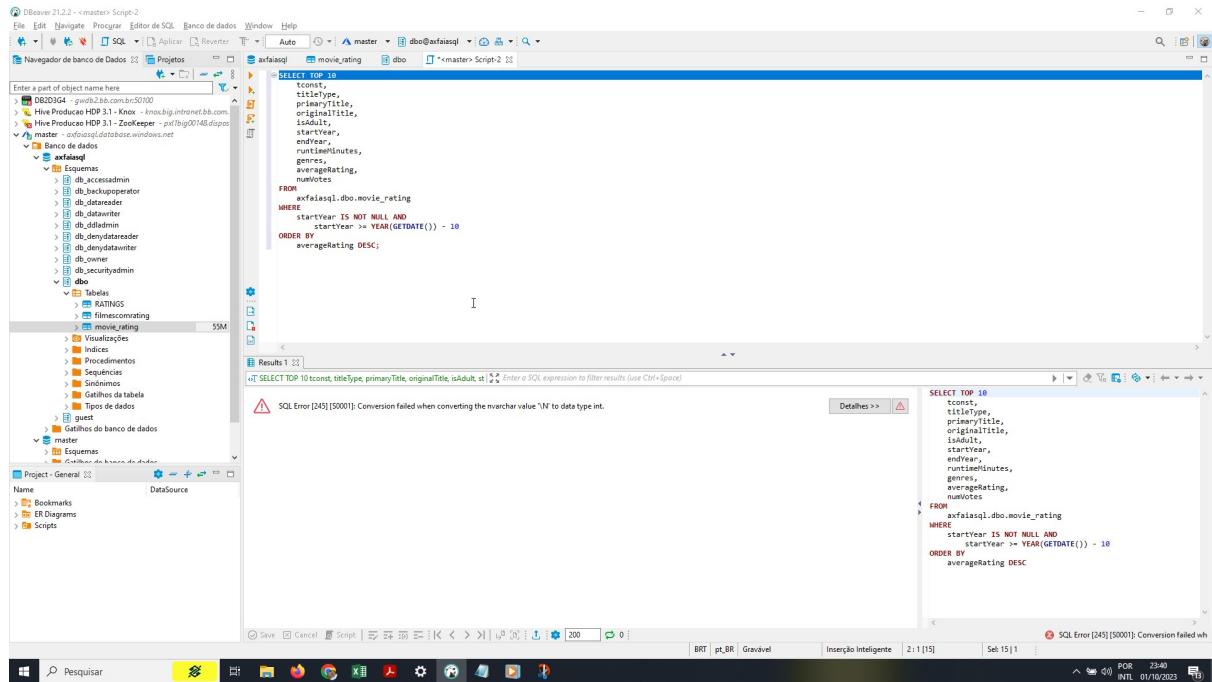
The screenshot shows the Microsoft Azure portal interface for managing a SQL server named 'axfaisql'. The main pane displays the 'Essentials' section, which includes details like Resource group (AXFAIA), Status (Available), Location (East US), and Subscription (Basic). It also shows a notification about a Microsoft Defender for SQL Free Trial expiring in 30 days. Below this, there's a table titled 'Available resources' showing one database named 'axfaisql'. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Quick start, Diagnose and solve problems, Settings (Microsoft Entra ID, SQL databases, SQL elastic pools, DTU quota, Properties, Locks), Data management (Backups, Deleted databases, Failover groups, Import/Export history), Security (Networking, Microsoft Defender for Cloud, Transparent Data Encryption, Identity), and a search bar at the bottom.

Para os SQLs, usei o DBEAVER, conseguindo linkar direto com o servidor e acompanhar a carga dos arquivos em tempo real.

The screenshot shows the DBeaver 21.2.2 interface connected to the 'master' database of the 'axfaisql' SQL server. The central view is the 'movie_rating' table, which has columns such as 'id', 'imdbId', 'siteType', 'recPrimaryTitle', 'recOriginalTitle', 'recIsAdult', 'recStartYear', 'recEndYear', and 'Valor'. The table contains approximately 40 rows of movie ratings. The left sidebar shows the database structure with tables like 'master', 'movie_rating', 'RATINGS', 'filmesrating', 'movie', and 'Visualidades'. The bottom status bar indicates '200 row(s) fetched - 274ms (+132ms)'.

Resultados e Visualização de dados

Logo na primeira tentativa de extrair o TOP 10, identifiquei que não exclui os registros com valores nulos, portanto deveria refazer todo a etapa anterior incluindo um tratamento de dados para estes valores.



The screenshot shows the DBeaver interface with a SQL script editor containing the following query:

```
SELECT TOP 10
    const,
    titleType,
    primaryTitle,
    originalTitle,
    isAdult,
    startYear,
    endYear,
    runtimeMinutes,
    genres,
    averageRating,
    numVotes
FROM
    axfaliasql.dbo.movie_rating
WHERE
    startYear IS NOT NULL AND
    startYear >= YEAR(GETDATE()) - 10
ORDER BY
    averageRating DESC;
```

The results pane shows a single row of data from the movie_rating table. A warning message at the bottom indicates a conversion error: "SQL Error [245] [S0001]: Conversion failed when converting the nvarchar value 'N' to data type int."

Encerro esta versão do relatório, pois estamos em cima do tempo limite para entrega.