

Pipeline de dados - PUC

Objetivo

A representação feminina nos bastidores de hollywood, embora em tendência ascendente, também não tem sido maravilhosa. Segundo a Forbes, “Entre 2008 e 2022, houve 325 nomeações para diretores em grandes eventos de premiação – apenas 8,9% dessas nomeações foram dadas a diretoras. Num estudo patrocinado pela Universidade Estadual de San Diego com foco nos 250 filmes de maior bilheteria de 2021 nos EUA, apenas 25% dos papéis nos bastidores foram preenchidos por mulheres. A porcentagem de editoras era de 22% e ainda menor, de 17% para diretores e roteiristas. O número de diretoras de fotografia era péssimo, seis em cada 100. Quando se trata de cargos de alto nível em mídia e entretenimento, apenas 27% são ocupados por mulheres.”

Atualmente, diante do sucesso do filme “Barbie” e das grandes produções realizadas por diretoras neste ano, o questionamento a seguir surgiu como objetivo de investigação:

Vale a pena investir no trabalho realizado por mulheres em filmes?

Detalhamento

1. Busca pelos dados

Os dados foram encontrados, após uma pesquisa utilizando o google, nos sites abaixo, representados nas figuras a seguir:

U.S. movies with gender-disambiguated actors, directors, and producers

[https://figshare.com/articles/dataset/U S movies with gender-disambiguated actors directors and producers/4967876](https://figshare.com/articles/dataset/U_S_movies_with_gender-disambiguated_actors_directors_and_producers/4967876)

figshare Browse Search on figshare... Log in Sign up

1/1

actors.json (24.42 MB) directors.json (2.37 MB) movies.json (42.4 MB) producers.json (10.35 MB)

Switch View | 4 files

U.S. movies with gender-disambiguated actors, directors, and producers

[Cite](#) [Download all \(79.54 MB\)](#) [Share](#) [Embed](#) [+ Collect](#)

Dataset posted on 2017-05-04, 12:40 authored by [Amaral Lab](#)

These datasets contain complete genre, cast, director, and producer information about 15,425 U.S.-produced movies released between 1894 and 2011.

The initial movie year, title, and genre information was obtained by Wasserman et al. (Cross-evaluation of metrics to estimate the significance of creative works, PNAS, 2015) from IMDb.com That dataset was expanded by Moreira et al. (forthcoming, 2017) to include movie budget, gender composition, cast, director, and producer information.

USAGE METRICS

1343 views	298 downloads	0 citations
------------	---------------	-------------

CATEGORIES

- Sociology not elsewhere classified
- Gender studies not elsewhere classified

Site contendo as informações de dados – US movies with gender

O tipo de formato dos dados encontrados para a criação do banco de dados são em .json, divididos em 4 arquivos: actors.json, directors.json, movies.json, producers.json.

Apenas o arquivo movies.json foi utilizado para a criação do modelo de dados.

Conjuntos de dados não comerciais da IMDb

<https://datasets.imdbws.com/>

IMDb data files available for download

Documentation for these data files can be found on <http://www.imdb.com/interfaces/>

- [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](#)

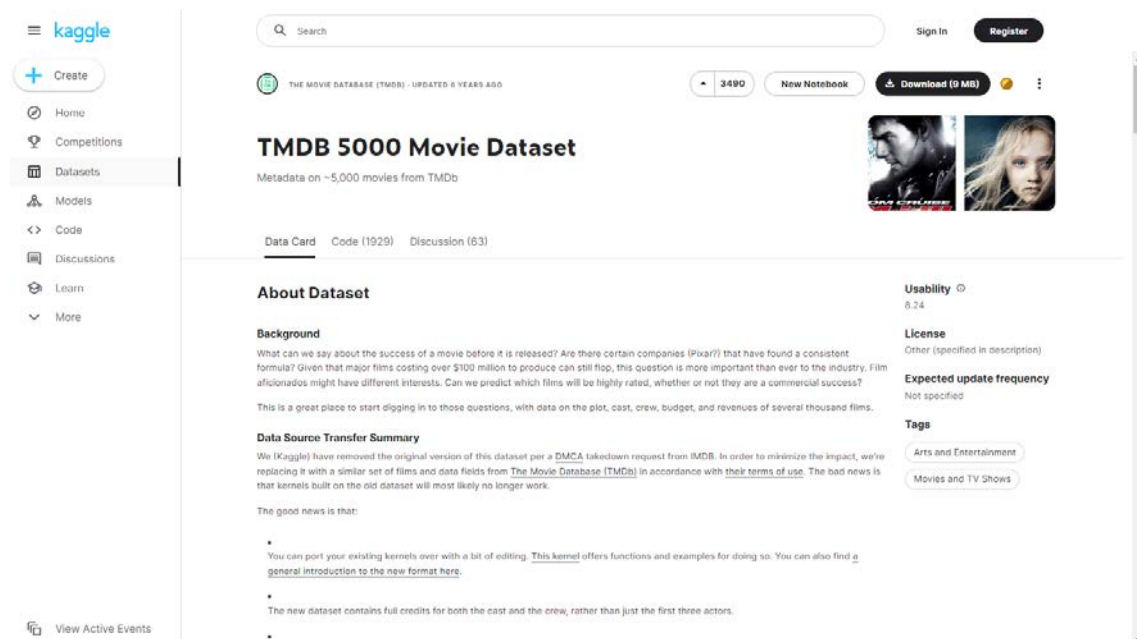
Site contendo as informações de dados – IMDb

O tipo de formato dos dados encontrados para a criação do banco de dados são em .tsv, divididos em 7 arquivos: name.basics.tsv, title.akas.tsv, title.basics.tsv, title.crew.tsv, title.episode.tsv, title.principals.tsv e title.ratings.tsv

Apenas o arquivo title.ratings.tsv. foi utilizado para criação do modelo de dados, renomeado como “ratings”.

Conjunto de dados de filmes TMDB 5000

<https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/>

The image is a screenshot of the Kaggle website showing the 'TMDB 5000 Movie Dataset' page. On the left is a navigation sidebar with links like 'Create', 'Home', 'Competitions', 'Datasets', 'Models', 'Code', 'Discussions', 'Learn', and 'More'. The main content area has a search bar at the top, followed by the dataset title 'TMDB 5000 Movie Dataset' and a subtitle 'Metadata on ~5,000 movies from TMDB'. Below this are tabs for 'Data Card', 'Code (1929)', and 'Discussion (63)'. The 'Data Card' is selected, showing an 'About Dataset' section with a 'Background' paragraph, a 'Data Source Transfer Summary' paragraph, and a list of bullet points. On the right side of the 'About Dataset' section, there are sections for 'Usability' (8.24), 'License' (Other), 'Expected update frequency' (Not specified), and 'Tags' (Arts and Entertainment, Movies and TV Shows). At the bottom left of the sidebar, there is a 'View Active Events' link.

Site contendo as informações de dados - filmes TMDB 5000

O tipo de formato dos dados encontrados para a criação do banco de dados são em .csv, divididos em 2 arquivos: tmdb_5000_credits.csv e tmdb_5000_movies.csv.

Apenas o arquivo tmdb_5000_movies.csv foi utilizado para criação do modelo de dados.

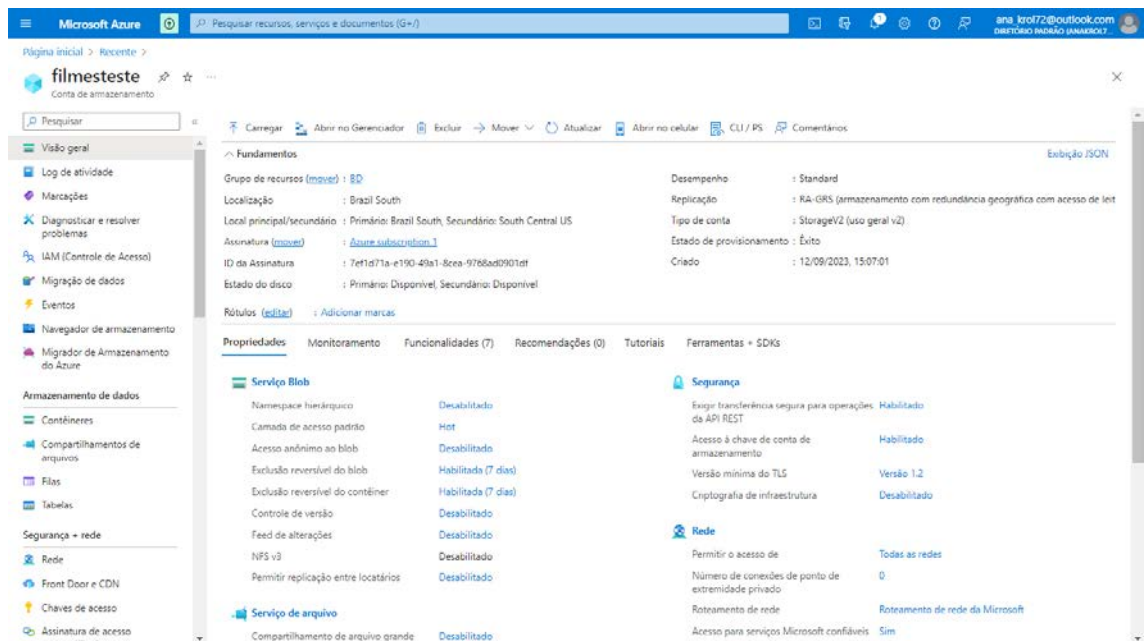
2. Coleta

Os dados descritos na etapa anterior foram baixados manualmente e armazenados localmente em meu notebook. Em seguida, foi realizada a escolha da plataforma de nuvem para armazenamento em nuvem: Azure.

O Azure é uma plataforma de nuvem de grande importância no mercado com uma variedade de recursos, que pode ser uma plataforma preferida para clientes que já estão usando produtos da Microsoft.

Pela integração com Visual Code Studio, PowerBI, ferramentas da Microsoft que utilizo, e pela capacidade de armazenamento oferecida na conta gratuita, optei pela escolha desta plataforma para o trabalho.

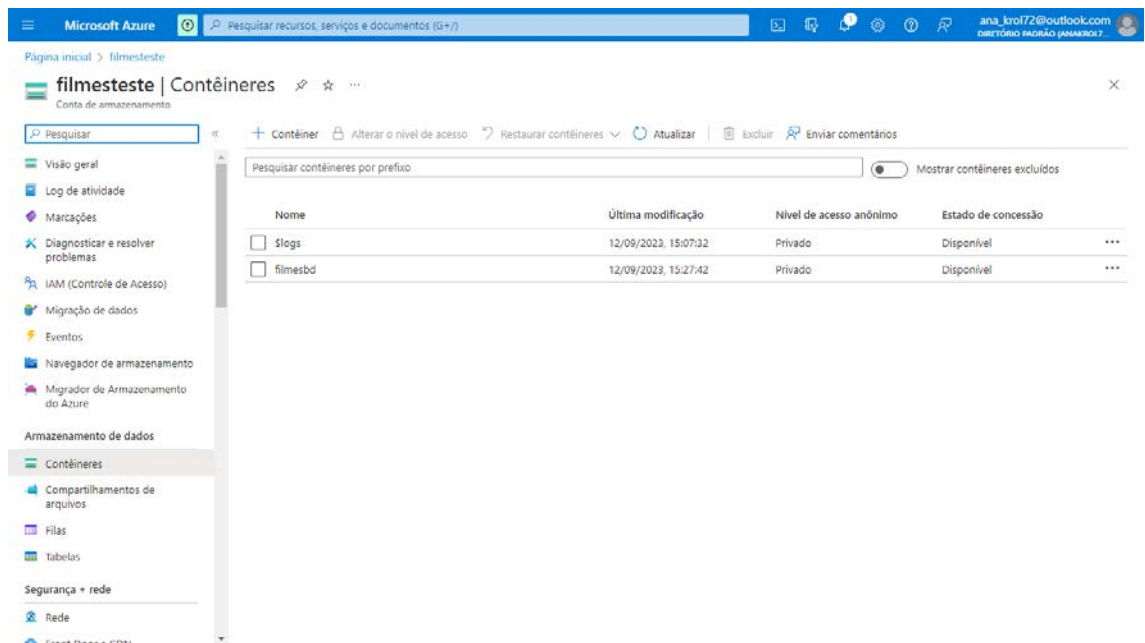
Para a adição dos dados na nuvem foi criada uma conta de armazenamento, filmeteste, que é o tipo de conta Storage Gen2, referente ao Azure Data Lake Storage Gen2.



Conta de armazenamento – filmesteste

O Azure Data Lake Storage é uma solução de data lake corporativa baseada em nuvem. Ele foi projetado para armazenar grandes quantidades de dados em qualquer formato e facilitar cargas de trabalho analíticas de Big Data. O Azure Data Lake Storage Gen2 refere-se à implementação atual da solução do Data Lake Storage do Azure.

Para utilização da conta de armazenamento é necessário a criação de um contêiner. O contêiner criado e utilizado foi o filmesbd, apresentado na figura abaixo.



Criação e utilização do contêiner filmesbd

Para facilitar a conexão localmente, via python, com o Data Lake em questão, que foi realizada através das ferramentas da biblioteca `azure.storage.blob`, os formatos de arquivo `.tsv` e `.csv` foram transformados para `.json` através do código python abaixo:

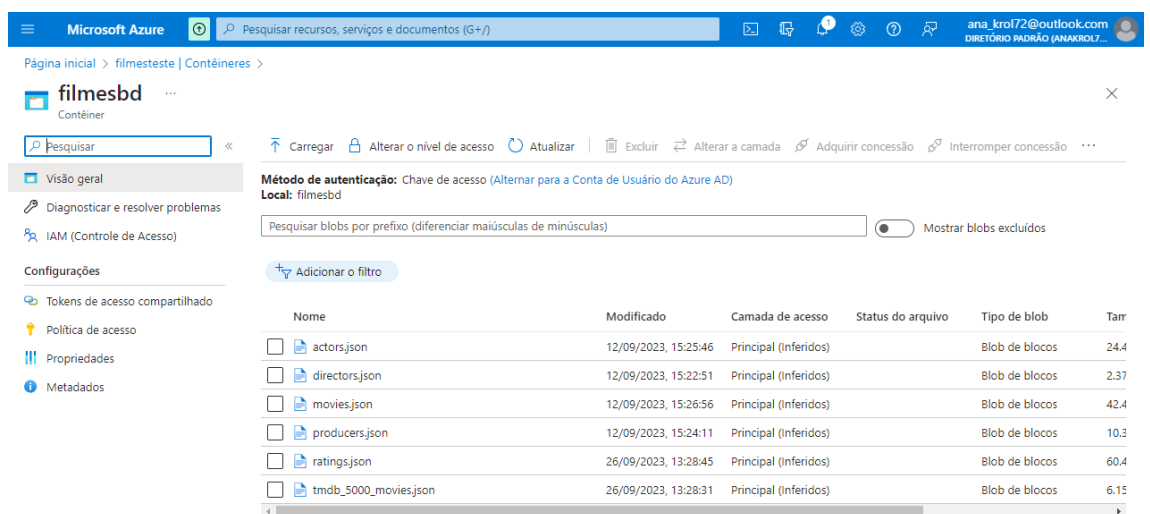
```
[1] import pandas as pd

[2] df_ratings = pd.read_table('ratings.tsv')
    df_ratings.to_json('ratings.json')

[3] df_box_office = pd.read_csv('tmdb_5000_movies.csv')
    df_box_office.to_json('tmdb_5000_movies.json')
```

Código python para transformação de arquivos .json

Assim, neste contêiner, os arquivos de dados `actors.json`, `directors.json`, `movies.json`, `producers.json`, `ratings.json` e `tmdb_5000_movies.json` foram carregados manualmente para a nuvem.



Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tam
<input type="checkbox"/> actors.json	12/09/2023, 15:25:46	Principal (inferidos)		Blob de blocos	24.4
<input type="checkbox"/> directors.json	12/09/2023, 15:22:51	Principal (inferidos)		Blob de blocos	2.37
<input type="checkbox"/> movies.json	12/09/2023, 15:26:56	Principal (inferidos)		Blob de blocos	42.4
<input type="checkbox"/> producers.json	12/09/2023, 15:24:11	Principal (inferidos)		Blob de blocos	10.3
<input type="checkbox"/> ratings.json	26/09/2023, 13:28:45	Principal (inferidos)		Blob de blocos	60.4
<input type="checkbox"/> tmdb_5000_movies.json	26/09/2023, 13:28:31	Principal (inferidos)		Blob de blocos	6.15

Arquivos carregados no contêiner

A partir dos arquivos carregados o processo de modelagem do modelo de dados será realizado.

3. Modelagem

a. Data Lake flat

O Data Lake flat foi o modelo de dados utilizado para o trabalho. Foi construído um Data Lake que continha todas as informações dos dados encontrados, “filmesdb”, e outro para análise pós ETL, “database”.

O modelo foi criado a partir da premissa que determina que este armazena dados em uma única tabela, sem relacionamento entre as linhas. Com isso os dados são organizados em colunas, com cada coluna representando um campo de dados.

Algumas regras de negócios foram avaliadas para

Regras de negócio:

Qual a relação do faturamento diante das categorias de gêneros sociais?

Qual a diferença entre produtores, atores e diretores em bilheteria se avaliado por gêneros sociais?

b. Catálogo de dados

O arquivo encontra-se separado, em anexo com o trabalho, descrevendo as informações que compõem os dados para o Data Lake.

c. Linhagem dos dados

Nesta parte, tem-se uma explicação sobre a origem dos dados encontrados para a formação Data Lake.

U.S. movies with gender-disambiguated actors, directors, and producers

Esses conjuntos de dados contêm informações completas sobre gênero, elenco, diretor e produtor sobre 15.425 filmes produzidos nos EUA lançados entre 1894 e 2011.

As informações iniciais sobre o ano do filme, título e gênero foram obtidas por Wasserman et al. (Avaliação cruzada de métricas para estimar a importância de trabalhos criativos, PNAS, 2015) do IMDb.com. Esse conjunto de dados foi ampliado por Moreira et al. (a ser publicado em 2017) para incluir informações sobre orçamento do filme, composição de gênero, elenco, diretor e produtor.

Dados para download –

U.S. movies with gender-disambiguated actors, directors, and producers -
[https://figshare.com/articles/dataset/U S movies with gender-disambiguated actors directors and producers/4967876](https://figshare.com/articles/dataset/U_S_movies_with_gender-disambiguated_actors_directors_and_producers/4967876)

Conjuntos de dados não comerciais da IMDb

Cada conjunto de dados está contido em um arquivo compactado e formatado com valores separados por tabulação (TSV) no conjunto de caracteres UTF-8. A primeira linha de cada arquivo contém cabeçalhos que descrevem o que há em cada coluna. Um '\N' é usado para indicar que um campo específico está ausente ou é nulo para esse título/nome. Os dados são atualizados diariamente.

Dados para download –

IMDb Non-Commercial Datasets - <https://developer.imdb.com/non-commercial-datasets/>

Conjunto de dados de filmes TMDB 5000

A Kaggle removeu a versão original deste conjunto de dados por uma solicitação de remoção DMCA do IMDb. Para minimizar o impacto, o mesmo foi substituído por um conjunto semelhante de filmes e campos de dados do The Movie Database (TMDB) de acordo com seus termos de uso.

Dados para download –

TMDB 5000 Movie Dataset - https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/?select=tmdb_5000_movies.csv

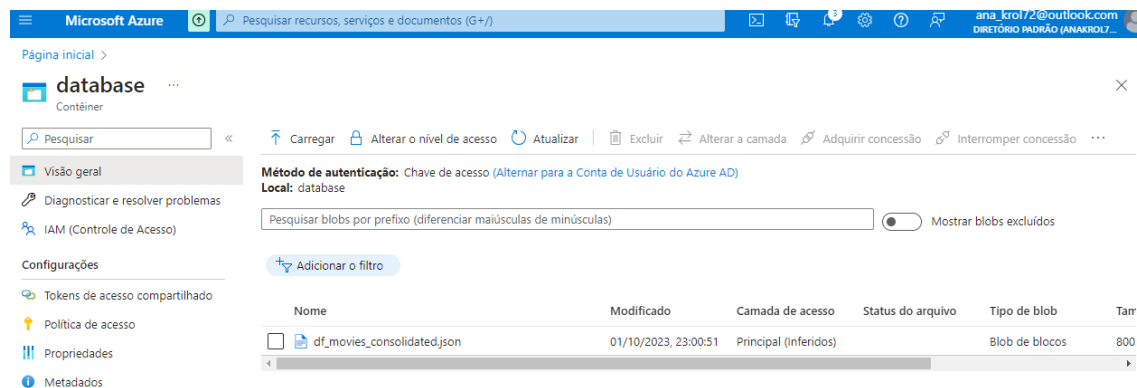
4. Carga

a. ETL

Esta parte encontra-se em anexo ao trabalho.

b. Carga

A carga foi realizada manualmente do notebook para a nuvem através da plataforma da Azure.



Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tam
df_movies_consolidated.json	01/10/2023, 23:00:51	Principal (Inferidos)		Blob de blocos	800

Data lake usado para análise

5. Análise

a. Qualidade de dados

Acerca da qualidade dos dados, verificou-se alguns pontos sobre o mesmo, como mostrado abaixo:

```
>
n_movies_without_actors = len(df_movies[df_movies.all_actors.isna()])
n_movies_without_directors = len(df_movies[df_movies.director.isna()])
n_movies_without_producers = len(df_movies[df_movies.producer.isna()])

print('FILMES SEM LISTA DE ATORES: ' + str(n_movies_without_actors))
print('FILMES SEM LISTA DE DIRETORES: ' + str(n_movies_without_directors))
print('FILMES SEM LISTA DE PRODUTORES: ' + str(n_movies_without_producers))

10]

++ FILMES SEM LISTA DE ATORES: 321
FILMES SEM LISTA DE DIRETORES: 267
FILMES SEM LISTA DE PRODUTORES: 930
```

Saída do código python - Dados faltantes

Acima, foi realizada uma busca por informações incompletas, onde constatou-se os seguintes números de filmes sem atores, diretores e produtores através do código python:

Filmes sem a presença de	
Atores	321
Diretores	267
Produtores	930

Nesta parte, foram verificados se os dados respeitavam a categoria as quais eram descritos no catálogo de dados através do código python, com o comando .info() da biblioteca do pandas.

Através dele, pode-se avaliar se os tipos das colunas, em Dtype, eram números, onde se classificariam como int64 ou float64 e objetos de textos, ao quais estariam na classificação de object. Através da coluna Non-Null Count, pode-se avaliar também se existem valores ausentes nas colunas.

Vejamos os resultados para os arquivos utilizados para a construção do Data Lake.

Movies


```
df_movies.info()
✓ 1.6s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15425 entries, 0 to 15424
Data columns (total 9 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   title                15425 non-null  object
1   year                 15425 non-null  int64
2   genre                15425 non-null  object
3   _id                  15425 non-null  object
4   director             15158 non-null  object
5   all_actors           15104 non-null  object
6   gender_percent       15024 non-null  float64
7   adjusted_budget      5501 non-null   float64
8   producer             14495 non-null  object
dtypes: float64(2), int64(1), object(6)
memory usage: 1.1+ MB
```

Saída do código python - .info()

Colunas	Valores nulos	Tipos
Title	0	Objeto
year	0	Números
Genre	0	Objeto
_id	0	Objeto
Diretor	0	Objeto
All_actors	0	Objeto
Gender_percent	0	Números
Adjusted_budget	0	Números
Producer	0	Objeto

Diante da análise acima, as colunas conferem com a descrição do catálogo de dados.

Ratings

```
df_ratings.info()
10] ✓ 0.1s
.. <class 'pandas.core.frame.DataFrame'>
Index: 1352361 entries, 0 to 1352360
Data columns (total 3 columns):
#   Column          Non-Null Count  Dtype
---  -
0   tconst          1352361 non-null object
1   averageRating   1352361 non-null float64
2   numVotes        1352361 non-null int64
dtypes: float64(1), int64(1), object(1)
memory usage: 41.3+ MB
```

Saída do código python - .info()

Colunas	Valores nulos	Tipos
Tconst	0	Objeto
averageRating	0	Números
numVotes	0	números

Diante da análise acima, as colunas conferem com a descrição do catálogo de dados.

Box Office

```
df_box_office.info()

[11] ✓ 0.0s

... <class 'pandas.core.frame.DataFrame'>
Index: 4803 entries, 0 to 4802
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   budget                 4803 non-null   int64
1   genres                 4803 non-null   object
2   homepage               1712 non-null   object
3   id                     4803 non-null   int64
4   keywords                4803 non-null   object
5   original_language      4803 non-null   object
6   original_title         4803 non-null   object
7   overview                4800 non-null   object
8   popularity              4803 non-null   float64
9   production_companies    4803 non-null   object
10  production_countries    4803 non-null   object
11  release_date            4802 non-null   object
12  revenue                 4803 non-null   int64
13  runtime                 4801 non-null   float64
14  spoken_languages        4803 non-null   object
15  status                  4803 non-null   object
16  tagline                 3959 non-null   object
17  title                   4803 non-null   object
18  vote_average            4803 non-null   float64
19  vote_count              4803 non-null   int64
dtypes: float64(3), int64(4), object(13)
memory usage: 788.0+ KB
```

Saída do código python - .info()

Colunas	Valores nulos	Tipos
Budget	0	Números
Genres	0	Objeto
Homepage	0	Objeto
Id	0	Números
Keywords	0	Objeto
Orginal_language	0	Objeto
Original_title	0	Objeto
Overview	0	Objeto
Popularity	0	Números
Production_companies	0	Objeto
Production_contries	0	Objeto
Release_date	0	Objeto
Revenue	0	Números
Runtime	0	Números
Spoken_languages	0	Objeto
Status	0	Objeto
Tagline	0	Objeto
Vote_average	0	Números
Vote_count	0	Números

Diante da análise acima, as colunas conferem com a descrição do catálogo de dados.

Nesta etapa foram avaliados a porcentagem dos valores faltantes para algumas colunas importantes para a criação do Data Lake.

Atores, diretores e produtores:

```
#porcentagem dos valores que estão faltando
percent_miss_actors = df_movies.all_actors.isna().mean().round(4)*100
percent_miss_director = df_movies.director.isna().mean().round(4)*100
percent_miss_producer = df_movies.producer.isna().mean().round(4)*100

print('Porcentagem de atores faltando: {}'.format(percent_miss_actors))
print('Porcentagem de diretores faltando: {}'.format(percent_miss_director))
print('Porcentagem de produtores faltando: {}'.format(percent_miss_producer))
]
```

Porcentagem de atores faltando: 2.08%
Porcentagem de diretores faltando: 1.73%
Porcentagem de produtores faltando: 6.03%

Código python da avaliação das porcentagens faltantes de atores, diretores e produtores

Colunas	Porcentagem Faltantes(%)
Atores	2.08
Diretores	1.73
Produtores	6.03

Budget e adjusted_budget

```
percent_miss_budget = df_box_office.budget.isna().mean().round(4)*100
percent_miss_adjusted_budget = df_movies.adjusted_budget.isna().mean().round(4)*100
print('Porcentagem de dados faltando em budget: {}'.format(percent_miss_budget))
print('Porcentagem de dados faltando em adjusted_budget: {}'.format(percent_miss_adjusted_budget))
5]
```

Porcentagem de dados faltando em budget: 0.0%
Porcentagem de dados faltando em adjusted_budget: 64.34%

Código python da avaliação das porcentagens faltantes das colunas budget e ajuste_budget

Colunas	Porcentagem de dados faltando(%)
Budget	0
Adjusted_budget	64.34

Nesta etapa foi verificado o cálculo da coluna `gender_percent`, para contestar sua validade, através de código python:

```
df_movies_consolidated['actress_percent'] =
df_movies_consolidated.actresses_count/(df_movies_consolidated.actors_count + df_movies_consolidated.actresses_count) * 100
df_movies_consolidated['actress_percent'] =
df_movies_consolidated['actress_percent'].apply(lambda x: round(x, 0))
#delta_percent = porcentagem calculada para verificação dos dados
df_movies_consolidated['delta_percent'] =
df_movies_consolidated.gender_percent -
df_movies_consolidated.actress_percent
```

Com isso, e através da coluna `delta_percent`, foi avaliado a diferença de valores, como mostrado abaixo:

#valores diferentes entre os dados carregados e da porcentagem calculada
df_movies_consolidated[df_movies_consolidated.delta_percent != 0][['gender_percent', 'actress_percent', 'delta_percent']]
se] ✓ 0.0s
**

	gender_percent	actress_percent	delta_percent
11	11.0	12.0	-1.0
12	29.0	30.0	-1.0
19	38.0	39.0	-1.0
24	20.0	21.0	-1.0
26	36.0	37.0	-1.0
...
2474	37.0	38.0	-1.0
2476	37.0	38.0	-1.0
2479	26.0	27.0	-1.0
2480	28.0	29.0	-1.0
2481	35.0	36.0	-1.0

964 rows x 3 columns

	gender_percent	actress_percent	delta_percent
11	11.0	12.0	-1.0
12	29.0	30.0	-1.0
19	38.0	39.0	-1.0
24	20.0	21.0	-1.0
26	36.0	37.0	-1.0
...
2474	37.0	38.0	-1.0
2476	37.0	38.0	-1.0
2479	26.0	27.0	-1.0
2480	28.0	29.0	-1.0
2481	35.0	36.0	-1.0

964 rows x 3 columns

Diferença entre os valores dos dados e os calculados

Com isso, pode-se perceber que os valores dos dados foram possivelmente truncados em diferença dos calculados que foram arredondados.

b. Solução do problema

Esta parte encontra-se em anexo ao trabalho.

6. Autoavaliação

a. Dificuldades encontradas

i. Data Factory e Pipeline

Realizei inicialmente o trabalho por esta plataforma, ingerindo meus dados pela plataforma da Data Factory para, posteriormente, seguir no processo de ETL na construção do pipeline utilizando as ferramentas que a Azure oferece.

Porém, na própria ingestão dos dados ocorreu um erro de inserção de dados no banco SQL, que não permitia inserir os dados no banco devido ao tipo de arquivo de origem, já que nele continha dados multivaloriados e isso não faz parte das regras de um banco dados SQL sem normalização, no caso. Por consequência, o pipeline e etl utilizando as ferramentas da Azure também não foram realizados. Não consegui sanar este problema já que para ele precisaria de justamente um banco de dados SQL para realizar a normalização e tratamento de dados antes de armazenar os dados em outro banco SQL em nuvem e a proposta do trabalho era de realizar os processos via nuvem.

ii. Databricks

Utilizei o Databricks como alternativa para criar um SQL warehouse, opção disponibilizada pela plataforma. Porém o seguinte erro abaixo na inicialização da base de dados foi informado:

The screenshot shows the Microsoft Azure Databricks web interface. On the left is a navigation sidebar with options like 'Novo', 'Espaço de trabalho', 'Recentes', 'Catálogo', 'Workflows', 'Computação', 'SQL', 'Editor de SQL', 'Consultas', 'Painéis', 'Alertas', 'Histórico de consultas', and 'SQL Warehouses'. The main area displays the 'teste' SQL Warehouse. A red error banner at the top of the main area reads 'Falha ao iniciar' (Failed to start). Below this, a detailed error message is shown in a yellow box: 'Clusters are failing to launch. Cluster launch will be retried. Details for the latest failure: Error: Error code: ResourceCountExceedsLimitDueToTemplate, error message: Subscription 7ef1d71a-e190-49a1-8cea-9768ad0901df has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 4 new resources of the this type which exceeds the quota. Please contact support to increase the quota for resource type PublicIpAddress Type: CLIENT_ERROR Code: AZURE_QUOTA_EXCEEDED_EXCEPTION Cluster-id (internal): 0923-010557-nlfcxyz Additional details: ("azure_error_code":"ResourceCountExceedsLimitDueToTemplate","azure_error_message":"Subscription 7ef1d71a-e190-49a1-8cea-9768ad0901df has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 4 new resources of the this type which exceeds the quota. Please contact support to increase the quota for resource type PublicIpAddress")'. Below the error message, the warehouse details are listed: Nome: teste (ID: fcd01cbd7eb5c356), Tipo: Pro, Tamanho do cluster: Pequeno, and Paragem automática: Após 45 minutos de inatividade.

Nome	teste (ID: fcd01cbd7eb5c356)
Tipo	Pro
Tamanho do cluster	Pequeno
Paragem automática	Após 45 minutos de inatividade

SQL Warehouses chamado teste

Falha ao iniciar

Clusters are failing to launch. Cluster launch will be retried.

Details for the latest failure: Error: Error code: ResourceCountExceedsLimitDueToTemplate, error message: Subscription 7ef1d71a-e190-49a1-8cea-9768ad0901df has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 4 new resources of the this type which exceeds the quota. Please contact support to increase the quota for resource type PublicIpAddress Type: CLIENT_ERROR Code: AZURE_QUOTA_EXCEEDED_EXCEPTION Cluster-id (internal): 0923-011010-vyryglkr Additional details: {"azure_error_code": "ResourceCountExceedsLimitDueToTemplate", "azure_error_message": "Subscription 7ef1d71a-e190-49a1-8cea-9768ad0901df has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 4 new resources of the this type which exceeds the quota. Please contact support to increase the quota for resource type PublicIpAddress"}

Falha ao iniciar o cluster

O erro refere-se aos recursos necessários para iniciar o SQL warehouse, informando que nesse tipo de conta não haveria recursos suficientes para sua criação.

Entretanto, foi possível a criação de um cluster, com a política “Personal Compute”:

The screenshot shows the Databricks interface with the 'Computation' tab selected. A table lists the following clusters:

Est...	Nome	Política	Runtime	Memó...	Núcleo...	DBU / ...	Origem	Criador	Noteb...
🟢	Ana Carolina Silvério's Pers...	Personal Compute	14.0 ML	14 GB	4 núcleos	0,75	UI	ana_krol72...	-
🔴	Ana Carolina Silvério's Clus...	Power User Compute	14.0 ML	-	-	-	UI	ana_krol72...	-
🔴	Ana Carolina Silvério's Clus...	-	13.3	-	-	-	UI	ana_krol72...	-

Clusters criados e suas políticas

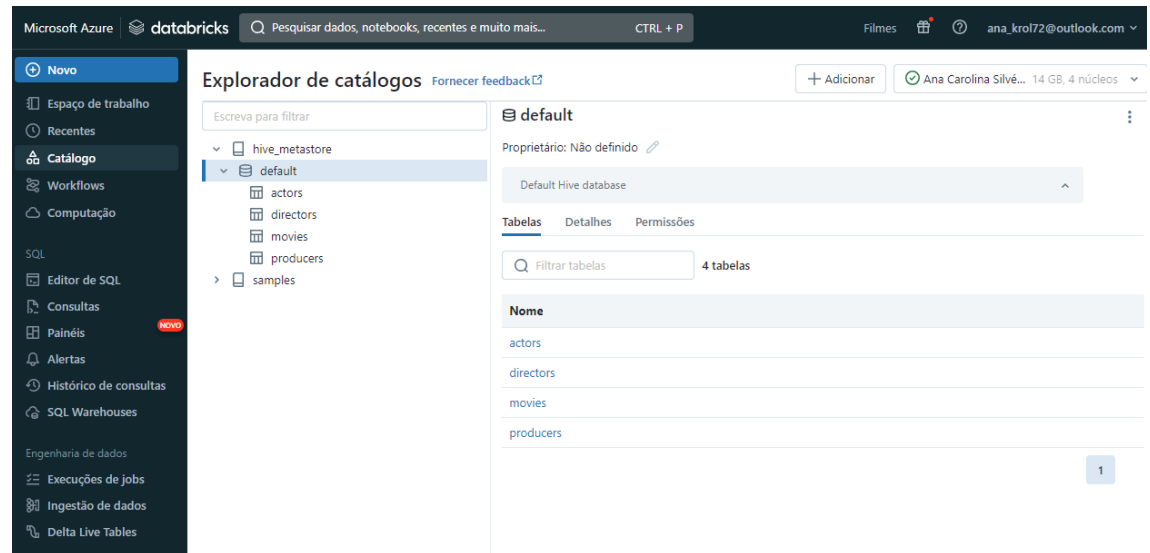
The screenshot shows the configuration page for the 'Ana Carolina Silvério's Personal Compute Cluster'. The 'Configuração' tab is active, displaying the following details:

- Política:** Personal Compute
- Modo de acesso:** Apenas um utilizador tem acesso
- Utilizador único:** Ana Carolina Silvério
- Desempenho:** Versão do Databricks Runtime: 14.0 ML (includes Apache Spark 3.5.0, Scala 2.12). Utilizar a aceleração do Photon: ☐
- Tipo de nó:** Standard_DS3_v2 (14 GB de memória, 4 núcleos)
- Terminar após:** 4320 minutos de inatividade
- Etiquetas:** Nenhuma etiqueta personalizada

A 'Resumo' sidebar on the right provides a quick overview: 1 controlador, 14 GB de memória, 4 núcleos; Runtime: 14.0.x-cpu-m1-scala2.12; Standard_DS3_v2, 0,75 DBU/h.

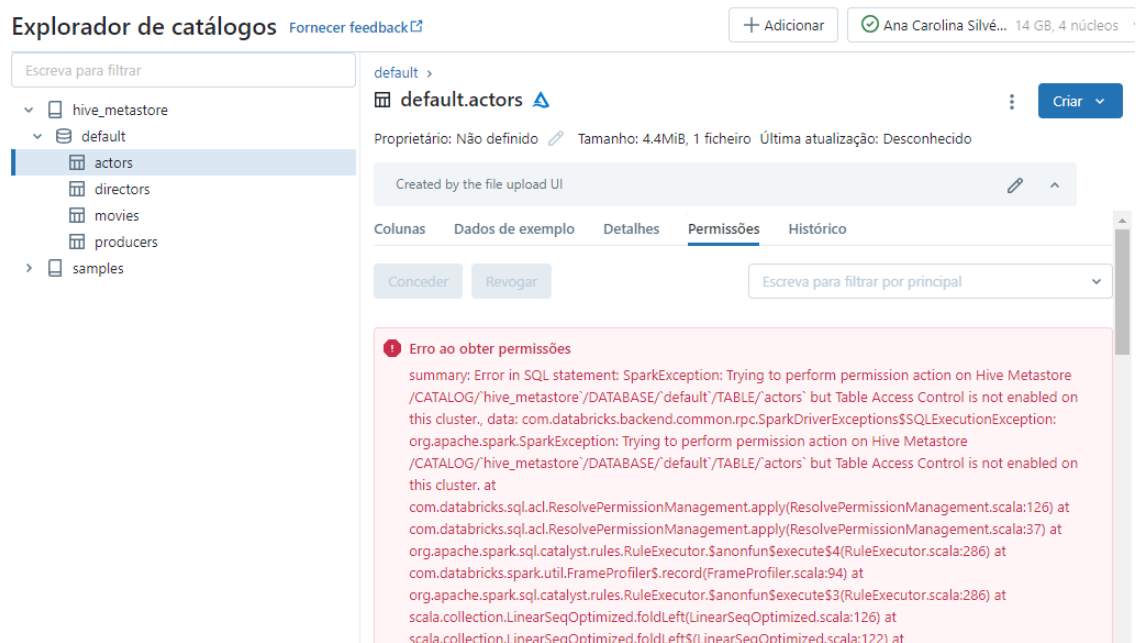
Cluster criado com a política “Personal Compute”

Através dele, conseguir adicionar os dados necessários para dar sequência a análise de dados:



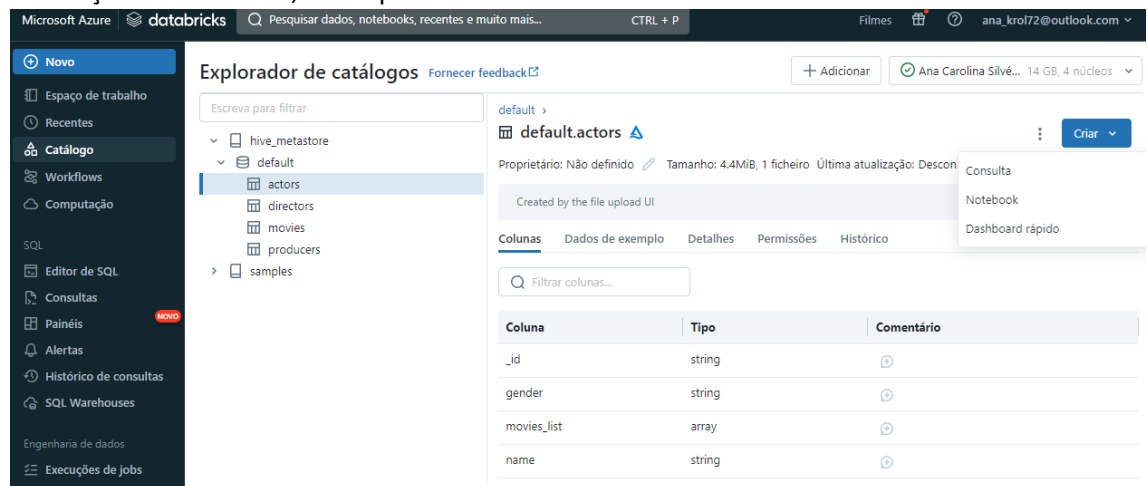
Dados adicionados ao cluster

Porém, os arquivos não tinham acesso externo, apenas na plataforma, como visto abaixo:



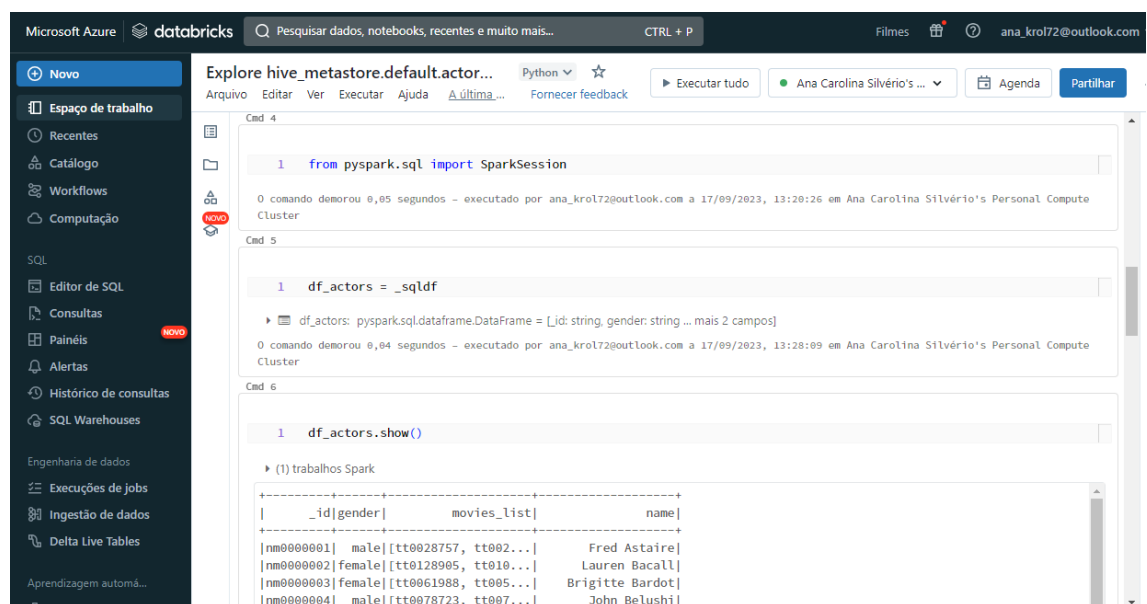
Permissão dos arquivos carregados

Todavia, era possível a análise de dados através da plataforma do Databricks através da criação do notebook, como pode ser visto abaixo:



Opção de criação do notebook para análise de dados

A partir dele foi possível a análise dos dados através das python, SQL, scala e R. A utilizada por mim para começo da análise foi python.



Notebook para análise de dados utilizando a linguagem python

Porém, devido a dificuldade de acesso aos dados da nuvem, limitando-se a análise dos dados na plataforma, e a restrição da criação do SQL warehouse pelo tipo de conta, esta opção de análise de dados não teve sequência.

b. Trabalhos futuros

Futuramente, há de se pensar em dados mais atuais para a análise, que reflitam uma situação mais moderna.

Há de se considerar também a qualidade dos dados adquiridos para a análise, os quais podem ser de origem mais rebuscada.

c. Objetivos delineados

Durante a análise dos dados foi possível determinar uma linha de raciocínio para estabelecer que somente depois de analisar os dados se é capaz de responder ao objetivo e aos indícios que o processo de análise acaba gerando.

7. Referências

U.S. movies with gender-disambiguated actors, directors, and producers -

[https://figshare.com/articles/dataset/U S movies with gender-disambiguated actors directors and producers/4967876](https://figshare.com/articles/dataset/U_S_movies_with_gender-disambiguated_actors_directors_and_producers/4967876)

AWS vs. Azure vs. Google: Comparação na nuvem – <https://blog.saninternet.com/aws-vs-azure-vs-google>

Introdução ao Azure Data Lake Storage Gen2 - <https://learn.microsoft.com/pt-br/azure/storage/blobs/data-lake-storage-introduction>

IMDb Non-Commercial Datasets - <https://developer.imdb.com/non-commercial-datasets/>

TMDB 5000 Movie Dataset - https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/?select=tmdb_5000_movies.csv

From The Screen To The Corner Office: What's Happening With The Gender Disparity In Hollywood? - <https://www.forbes.com/sites/joshwilson/2022/12/02/from-the-screen-to-the-corner-office-whats-happening-with-the-gender-disparity-in-hollywood/?sh=6e8e5d083af2>