

MVP de Engenharia de Dados

Roberto Harkovsky da Cunha

1 Definição do Problema

O IMDb (um acrônimo para Internet Movie Database) é um banco de dados on-line de informações relacionadas a filmes, séries de televisão, podcasts, vídeos caseiros, videogames e streaming de conteúdo on-line - incluindo elenco, equipe de produção e biografias pessoais, resumos de enredos, curiosidades, classificações e análises críticas e de fãs. Como complemento aos dados, o IMDb oferece uma escala de classificação que permite aos usuários votar e avaliar os filmes em uma escala de um a dez.

Neste escopo, objetivo deste projeto é o de realizar uma análise dos títulos publicados (filmes, seriados de TV) e responder as seguintes questões:

- Qual a popularidade dos filmes de James Bond
- Qual a pontuação dos filmes de James Bond
- Quais são os gêneros de filmes mais populares
- Quais os gêneros com as melhores pontuações
- Quais os 10 filmes com maior popularidade de Steven Spielberg
- Quais os 10 filmes com maior pontuação de Steven Spielberg
- Quais as 10 maiores médias de pontuação de diretores de filmes, com pelo menos 10000 votos mais de 5 filmes realizados?
- Qual é o tempo de execução típico para filmes de cada gênero?
- Quantos filmes foram feitos de cada gênero por ano entre 2020/22?
- Quem são os atores que interpretaram 'James Bond' em um filme?
- Quantas vezes eles fizeram o papel de 'James Bond'?
- Quantos filmes existem em cada gênero?

2 Visão Geral do Projeto

Utilizaremos neste projeto a nuvem Azure da Microsoft, e seus serviços. Foram utilizados serviços de repositórios de dados para armazenar os dados originais. Foram criados base de dados em servidores SQL serverless como

local para carga dos dados e para o processo de ETL foi utilizado o Azure Data Factory (ADF) que é o serviço ETL na nuvem do Azure para integração e transformação de dados sem servidor.

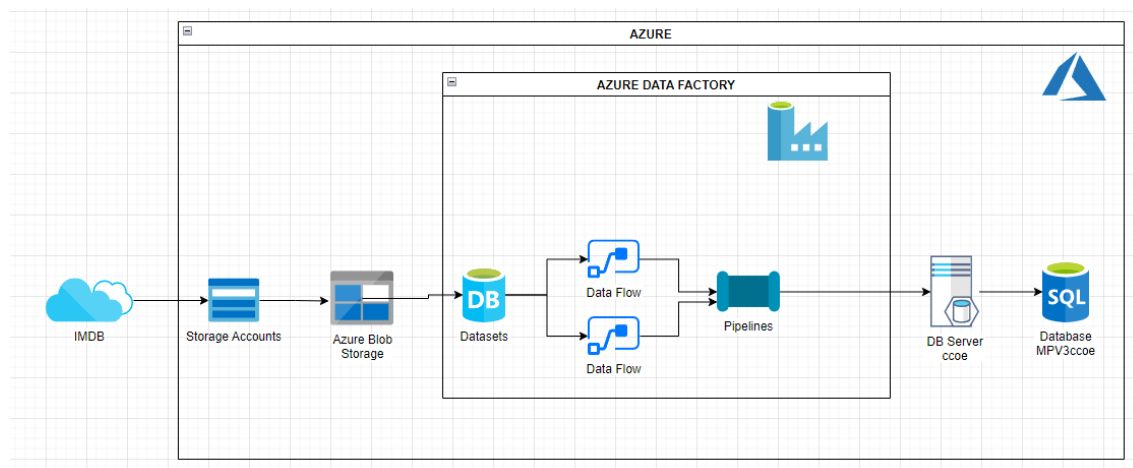
O desenvolvimento do projeto foi composto pelas seguintes etapas:

- Fonte de dados (Data Sourcing)
- Modelagem de dados
- Ingestão de dados
- Transformação de dados
- Carga de dados
- Resultados e Visualização de dados

Na etapa de fonte de dados foram utilizados arquivos públicos do portal do IMDB.

As etapas de ETL (ingestão, transformação e carga dos dados) foram realizadas na plataforma Azure e no Azure Data Factory. A figura abaixo apresenta uma visão geral destas etapas.

Figura 1 - Visão geral do Processo de ingestão, transformação e carga dos dados



A etapa de resultados e Visualização foi feita utilizando o SSMS da Microsoft.

Todas as etapas do projeto serão detalhadas nos itens a seguir.

3 Fonte de dados (Data Sourcing)

Para este projeto foi utilizado o Dataset IMDb, que consiste em 7 arquivos compactados, com valores separados por tabulação (*.tsv), que estão

disponíveis para download em <https://datasets.imdbws.com/>. Os arquivos selecionados são os seguintes:

- 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

Algumas informações adicionais sobre os dados deste dataset IMDB:

- Os dados são atualizados diariamente, embora os dados utilizados neste projeto tenham sido obtidos em 16/09/2023.
- Cada um desses arquivos compactados com valores separados por tabulação (TSV) formatados 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 faltando ou tem um valor NULL para esse título ou nome.

3.1 Detalhes dos Dados IMDB

O detalhamento do conteúdo dos arquivos está a seguir:

3.1.1 name.basics.tsv.gz

Contém as seguintes informações para nomes da equipe/atores:

Coluna	Descrição
nconst (string)	identificador alfanumérico exclusivo do nome/pessoa.
PrimaryName (string)	nome pelo qual a pessoa é creditada com mais frequência.
birthYear	no formato AAAA.
deathYear	no formato AAAA, se aplicável, caso contrário, “\N”.
primaryProfession (matriz de strings)	as 3 principais profissões da pessoa.
knownForTitles (matriz de tconsts)	títulos pelos quais a pessoa é conhecida.

3.1.2 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, etc).
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)	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.

3.1.3 title.akas.tsv.gz

Contém as seguintes informações extras para filmes:

Coluna	Descrição
titleId (string)	um tconst que é um identificador alfanumérico exclusivo do título.
ordenação (inteiro)	um número para identificar exclusivamente as linhas para um determinado titleId.
title (string)	o título localizado.
region (string)	a região para esta versão do título.
language (string)	o idioma do título.
types (array)	Conjunto enumerado de atributos para este título alternativo. Um ou mais dos seguintes: “alternativo”, “dvd”, “festival”, “tv”, “vídeo”, “trabalho”, “original”, “imdbDisplay”. Novos valores poderão ser adicionados no futuro sem aviso prévio.
attributes (array)	Termos adicionais para descrever este título alternativo, não enumerados.
isOriginalTitle (booleano)	0: título não original; 1: título original.

3.1.4 Title.crew.tsv.gz

Contém informações do diretor e escritor de todos os títulos da IMDb. Os campos incluem:

Coluna	Descrição
tconst (string)	identificador alfanumérico exclusivo do título.
directors (array de nconsts)	diretor(es) do título determinado.
writers (array de nconsts)	escritor(es) do(s) título(s) fornecido(s).

3.1.5 title.episode.tsv.gz

Contém as informações do episódio de TV. Os campos incluem:

Coluna	Descrição
tconst (string)	identificador alfanumérico do episódio.
parentTconst (string)	identificador alfanumérico da série de TV pai.
seasonNumber (inteiro)	número da temporada à qual o episódio pertence.
EpisodeNumber (inteiro)	número do episódio do tconst da série de TV.

3.1.6 title.principais.tsv.gz

Contém o elenco/equipe principal dos títulos:

Coluna	Descrição
tconst (string)	identificador alfanumérico exclusivo do título.
ordering (inteiro)	um número para identificar exclusivamente as linhas para um determinado titleId.
nconst (string)	identificador alfanumérico exclusivo do nome/pessoa.
category (string)	a categoria do trabalho em que a pessoa estava.
job (string)	o cargo específico, se aplicável, caso contrário, “\N”.
characters (string)	o nome do personagem interpretado, se aplicável, caso contrário “\N” (é realmente “[role1,role2,...]” ou “\N”).

3.1.7 title.ratings.tsv.gz

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

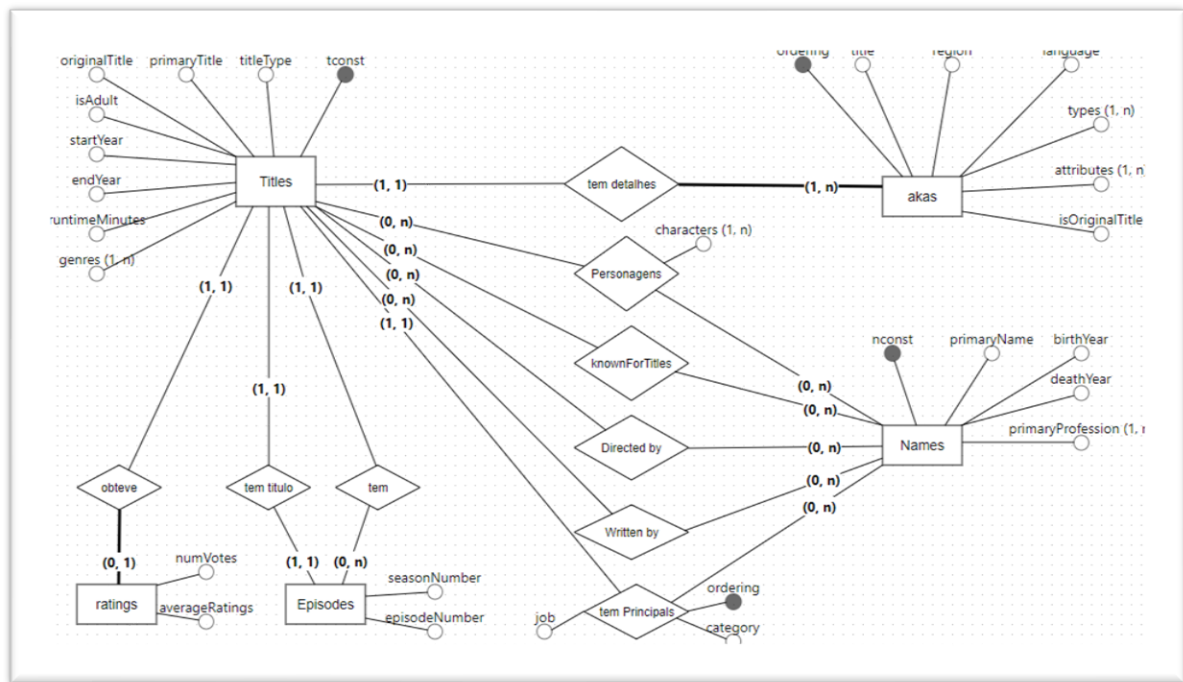
Coluna	Descrição
tconst (string)	identificador alfanumérico exclusivo do título.
AverageRating	média ponderada de todas as avaliações individuais dos usuários.
numVotes	número de votos que o título recebeu.

4 Modelagem de dados

O objetivo principal do projeto é responder perguntas ligados aos fatos “IMDBRating” e “NumVotes”, segundo as dimensões tempo (ano), diretor, escritor, gênero do filme, linguagem, personagens e episódios, bem como outras questões relacionadas.

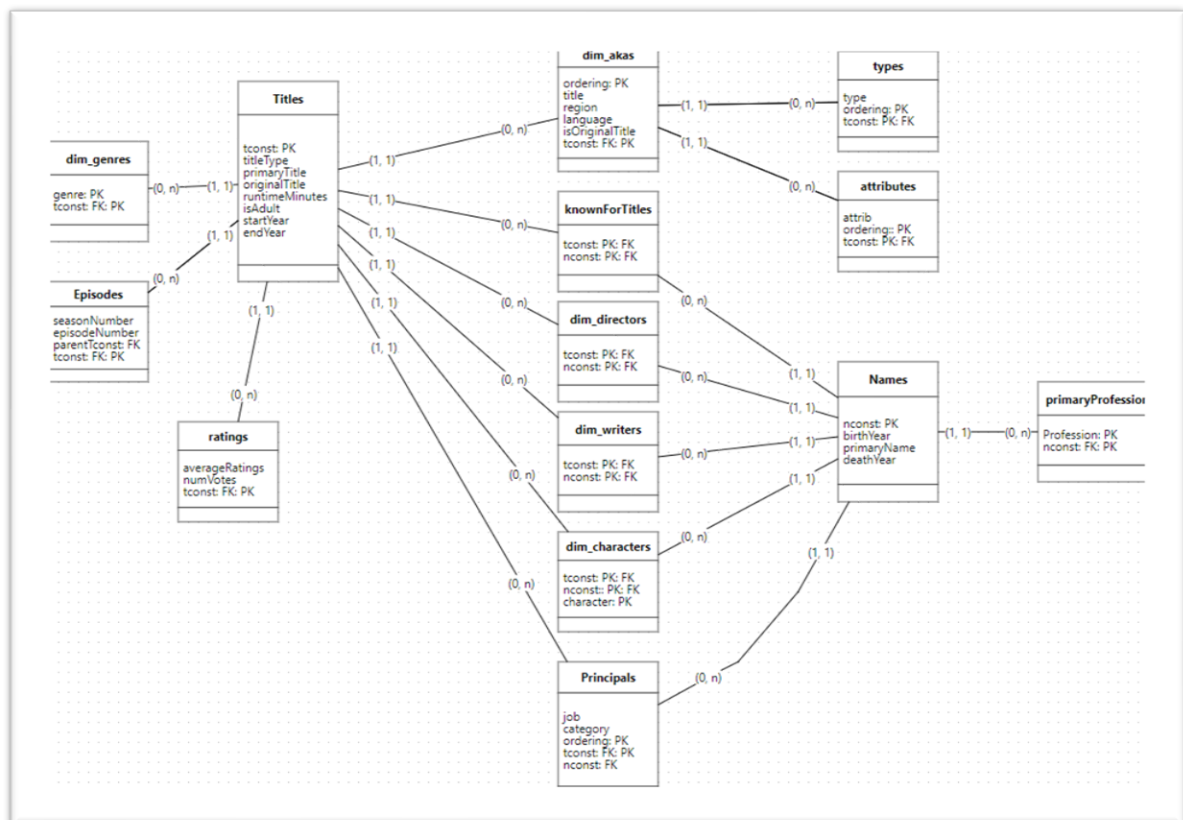
O modelo conceitual para alcançar estes objetivos é o seguinte:

Figura 2 - modelo conceitual do projeto



Já o modelo lógico derivado é o seguinte:

Figura 3 - modelo lógico do projeto



Este modelo lógico será utilizado como esquema de saída da transformação dos dados.

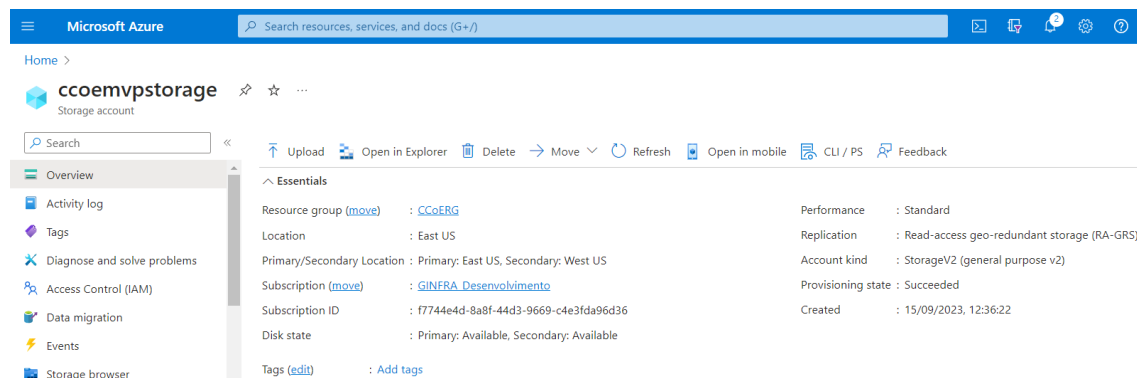
5 ETL

Como comentado no tópico “Visão geral”, para o processo de ETL foi utilizado o ADF. Para tal fim foi criada uma instância do ADF chamada “ccoemvpdfactory”, na qual será realizada a orquestração do ETL deste projeto.

5.1 Ingestão de dados

A primeira etapa foi a criação de um repositório no Azure de onde os arquivos serão ingeridos originalmente. O repositório foi criado por meio do serviço StorageAccount, nomeado “ccoemvpstorage” conforme a figura:

Figura 4 - storage account para repositório dos arquivos fontes



A seguir foram criados 2 containers: “imdbcontainer”, que contém os arquivos IMDB originais compactados, e o container “imdbcontainerunzip” que contém a versão descompactada dos arquivos.

A seguir foram criados 2 containers: “imdbcontainer” e “imdbcontainerunzip”.

Os arquivos originais compactados (extensão .gz) foram descompactados e carregados no storage Account “ccoemvpstorage” da seguinte forma: a versão

compactada foi carregada no container “imdbcontainer”, e a versão descompactada foi carregada no container “imdbcontainerunzip”.

Figura 5 - Container com os arquivos fonte

Microsoft Azure

Search resources, services, and docs (G+)

Home > ccoempvstorage

ccoempvstorage

Containers

Storage account

Search

Container

Change access level

Restore containers

Refresh

Delete

Give feedback

Search containers by prefix

Name	Last modified	Anonymc
<input type="checkbox"/> \$logs	15/09/2023, 12:36:47	Private
<input type="checkbox"/> bobcontainer	15/09/2023, 12:55:45	Private
<input type="checkbox"/> imdbcontainer	16/09/2023, 12:03:42	Private
<input type="checkbox"/> imdbcontainerunzip	16/09/2023, 13:01:34	Private

A figura a seguir evidencia a criação e o conteúdo do container “imdbcontainerunzip”.

Figura 6 - Container com arquivos IMDB descompactados

Home > ccoempvstorage | Containers >

imdbcontainerunzip

Container

Search

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Authentication method: Access key (Switch to Azure AD User Account)

Location: imdbcontainerunzip

Search blobs by prefix (case-sensitive)

Add filter

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> name.basics.tsv	16/09/2023, 13:08:40	Hot (Inferred)		Block blob
<input type="checkbox"/> title.akas.tsv	16/09/2023, 13:10:47	Hot (Inferred)		Block blob
<input type="checkbox"/> title.basics.tsv	16/09/2023, 13:07:48	Hot (Inferred)		Block blob
<input type="checkbox"/> title.crew.tsv	16/09/2023, 13:04:26	Hot (Inferred)		Block blob
<input type="checkbox"/> title.episode.tsv	16/09/2023, 13:03:29	Hot (Inferred)		Block blob
<input type="checkbox"/> title.principals.tsv	16/09/2023, 13:11:53	Hot (Inferred)		Block blob
<input type="checkbox"/> title.ratings.tsv	16/09/2023, 13:03:56	Hot (Inferred)		Block blob

Para este projeto utilizaremos apenas os arquivos descompactados oriundos deste container.

5.2 Transformação de dados

Foram criados 2 dataflows no processo de transformação dos dados: um para as tabelas de dimensões chamado de “dataflow_dim” e um segundo chamado de “dataflow_fact” para a tabela de fatos.

Para seleção dos atributos para as novas tabelas foi utilizado a técnica de projeção das colunas por meio do componente "Select".

Para construir a tabela de fatos, utilizamos o componente "Join" junção para anexar os valores numéricos de número de votos e classificação (rating) dos filmes.

Já o tratamento dados para estes campos, como mostrado na modelagem, foi de criar tabelas específicas par cada um deles. Para os campos multivalorados “genre” e “profession”, foram derivadas novas tabelas, por meio da utilização dos componentes “derived column” e “flatten”. E para garantir que não ocorrência de campos nulos, foi utilizado o componente de filtragem de conteúdo "Filter".

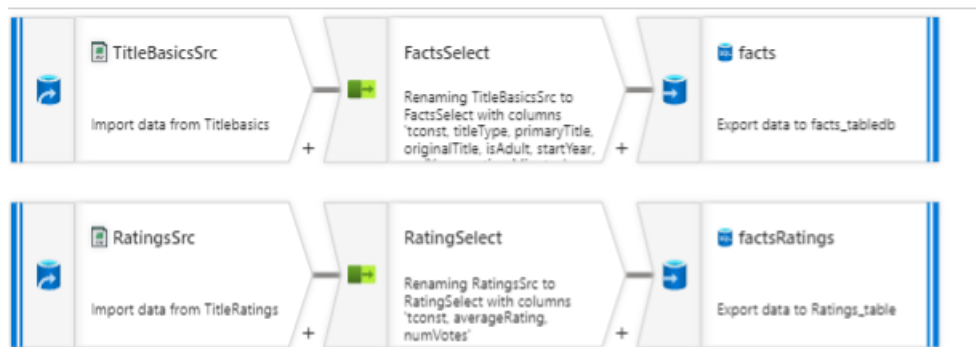
Pela modelagem, houve a necessidade de derivar algumas tabelas de campos com valores específicos de ocorrência, como o cargo de diretor e escritor. Assim para criação das tabelas "dim_diretor" e "dim_escrito" foi utilizado o componente filtro de conteúdo no campo job, procurando especificamente pelos valores "Director" e "writer" respectivamente.

O tratamento dos dados para cada fluxo está detalhado a seguir.

5.2.1 “dataflow_fact”

O Dataflow “dataflow_fact” para geração da tabela de fatos está apresentado abaixo:

Figura 7 - Fluxo Dataflow Facts



Para geração do dataflow de fatos, foram utilizadas como fonte as tabelas titleBasics e titleRatings (1)

Em cada uma delas foram projetadas as seguintes colunas (2)

- titleRatings (tconst, averageRating, numVotes)
- titleBasics (tconst, primarytite, originaltitle, isAdult, startYear, endYear, runtimeMinutes)

Figura 8 - Datasources dataflow facts - Fatos



5.2.2 “dataflow_dim”

O Dataflow “datalow_dim” para geração das tabelas de dimensão do projeto está apresentado na figura a seguir.

Figura 9 - Dataflow de dimensões



Para geração do dataflow de dimensões, foram utilizadas como fonte as tabelas titleBasics , titleEpisodes, names, titleprincipals (1)

Em algumas das tabelas foram projetadas as seguintes colunas (2)

- Tabela titleBasics: projetado o campo multivalorado “genre” para criação de uma tabela específica de gêneros de filmes;
- Tabela Names: projetadas as colunas “nconst”, “primaryName”, “birthDate”, “deathDate” para criação de uma tabela de apoio de nomes de pessoal;
- Tabela Names foi ainda projetado o campo multivalorado “profession” para criação de uma tabela específica de profissões na produção dos filmes.
- Da Tabela titlePrincipals foi projetado e transformado o campo multivalorado “characters” para criação de uma tabela específica de personagens de filmes;
- Da Tabela titlePrincipals foi projetado e transformado o campo “Director”, filtrando as linhas com a categoria “Director” para criação de uma tabela específica de diretores de filmes;
- Da Tabela titlePrincipals foram projetados os campos “tconst”, “ordering”, “nconst”, “category” e ‘job’ para uso nas consultas;

- A tabela Episodes não sofreu o processo de projeção.

Para os campos multivalorados “genre” e “profession”, foram feitas diversas transformações para derivar novas tabelas dos campos multivalorados.

(3)

Já para a tabela de personagens (character) foram feitas diversas transformações e limpas as possíveis ocorrências de nulos (null). (4)

Figura 10 - Dataflow_dim - dimensões

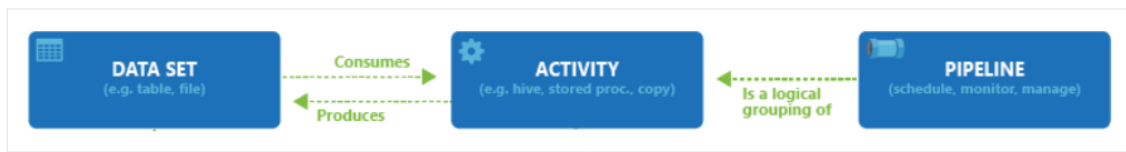


5.2.3 Pipelines

Um pipeline é um agrupamento lógico de atividades que juntas executam uma tarefa, que ingerem dados de um dataset e produzem novos dados. As atividades disponíveis são atividade de cópia, atividade de fluxo de dados

Desta forma as atividades de um pipeline definem as ações a serem executadas nos seus dados. No caso deste projeto, foram usadas atividades de fluxo de dados.

Figura 11 – visão geral do funcionamento de Pipelines (fonte:Microsoft)



Neste projeto foram criados 2 pipelines contendo cada um um dos fluxos principais dataflow_facts ou dataflow_dim. Um exemplo de execução de um deles está apresentado abaixo:

Figura 12 - exemplo de execução - Pipeline facts

Pipeline run ID: 05be433d-f4e7-4c86-9c03-3ef1fb44f970 **Pipeline status:** Succeeded

Data flow activity for this debug run will start as soon as the data flow debug session is ready.

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User property
dataflow_facts	Succeeded	Data flow	9/17/2023, 9:48:29 PM	4m 22s	AutoResolveIntegration	

5.3 Carga de dados

O processo de carga de dados envolveu 3 fases distintas:

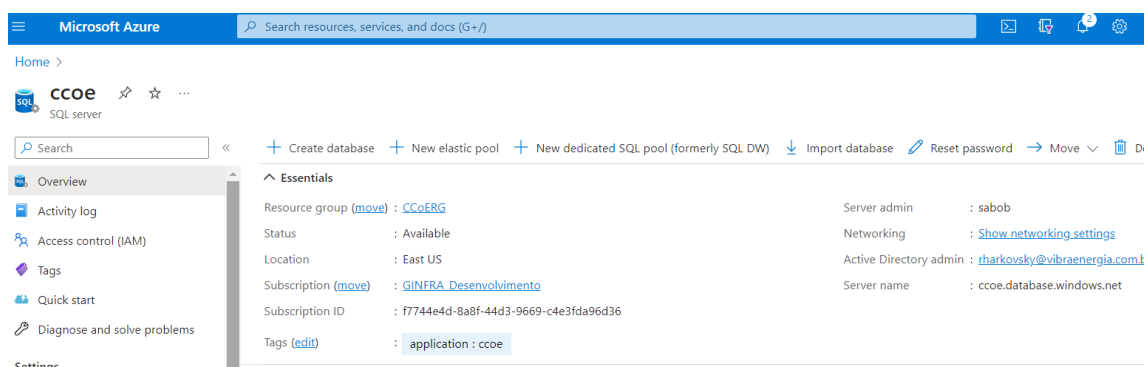
- Criação das tabelas no database mpv3db
- Geração das tabelas brutas sem restrições (pelo ADF)
- Carga final dos dados nas tabelas com restrições

Estas etapas estão descritas nos itens a seguir.

5.3.1 Criação das tabelas

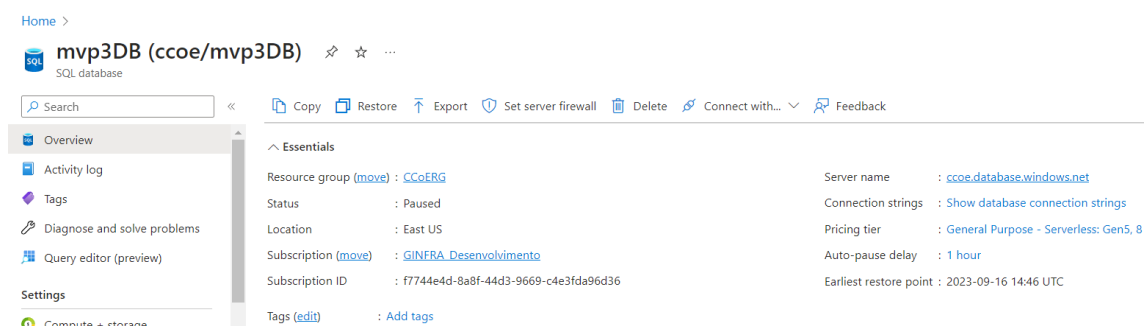
Para receber os dados oriundos do ETL foi criado no Azure um servidor de banco de dados SQL *serverless* chamado de “ccoe”, como evidenciado na figura a seguir.

Figura 13 - servidor de banco de dados “ccoe” no Azure



Em seguida, uma base de dados chamada de “mvp3DB” foi criada neste servidor (vide figura), para onde foram direcionados os dados de saída do modelo.

Figura 14 - Database no servidor “ccoe”



A próxima etapa consistiu em, a partir da modelagem realizada no item 4, proceder a criação propriamente dita das tabelas com as devidas restrições de

chave primária e estrangeira. O script da figura foi elaborado e aplicado ao database, resultando na criação das tabelas.

Figura 15 - Script de criação das tabelas

```
CREATE TABLE title_facts
(
  tconst varchar(10) PRIMARY KEY,
  titleType varchar(30),
  primaryTitle varchar(max),
  originalTitle varchar(max),
  isAdult INT,
  startYear INT,
  endYear INT,
  runtimeMinutes INT,
);
CREATE TABLE title_names
(
  nconst varchar(10),
  primaryName varchar(150),
  birthYear INT,
  deathYear INT,
  CONSTRAINT pk_names PRIMARY KEY (nconst)
);
CREATE TABLE dim_Profession
(
  nconst varchar(10),
  profession varchar(30) ,
  CONSTRAINT pk_profession PRIMARY KEY (nconst, profession)
);
CREATE TABLE dim_principals
(
  tconst varchar(10),
  ordering INT,
  nconst varchar(10),
  job varchar(max),
  category varchar(60),
  CONSTRAINT pk_principals PRIMARY KEY (tconst,ordering)
)
CREATE TABLE dim_episodes
(
  tconst varchar(10) PRIMARY KEY,
  parentTconst varchar(10),
  seasonNumber INT,
  episodeNumber INT,
);
CREATE TABLE [dbo].[dim_directors]
(
  tconst varchar(10),
  nconst varchar(10),
  CONSTRAINT pk_Director PRIMARY KEY (tconst, nconst)
);
CREATE TABLE dim_Characters
(
  tconst varchar(10),
  nconst varchar(10),
  characters varchar(30),
  CONSTRAINT pk_Character PRIMARY KEY (tconst, nconst, characters)
```

```

);

CREATE TABLE dim_genres
(
    tconst varchar(10),
    genre varchar(30),
    CONSTRAINT pk_genres PRIMARY KEY (tconst, genre)
);

CREATE TABLE title_Ratings
(
    tconst varchar(10) PRIMARY KEY,
    averageRating DECIMAL (5,1),
    numVotes INT
);

ALTER TABLE dim_episodes ADD FOREIGN KEY(parentTconst) REFERENCES title_facts (parentTconst)
ALTER TABLE dim_episodes ADD FOREIGN KEY(tconst) REFERENCES title_facts (tconst)

ALTER TABLE dim_principals ADD FOREIGN KEY(tconst) REFERENCES title_facts (tconst)
ALTER TABLE dim_principals ADD FOREIGN KEY(nconst) REFERENCES title_names (nconst)

ALTER TABLE dim_directors FOREIGN KEY(tconst) REFERENCES title_facts (tconst)
ALTER TABLE dim_directors FOREIGN KEY(nconst) REFERENCES title_names (nconst)

ALTER TABLE dim_Profession ADD FOREIGN KEY(nconst) REFERENCES title_facts (nconst)
ALTER TABLE dim_genres ADD FOREIGN KEY(genre) REFERENCES title_facts (genre)
ALTER TABLE dim_genres ADD FOREIGN KEY(tconst) REFERENCES title_facts (tconst)

ALTER TABLE dim_Characters ADD FOREIGN KEY(tconst) REFERENCES title_facts (tconst)
ALTER TABLE dim_Characters ADD FOREIGN KEY(nconst) REFERENCES title_names (nconst)
ALTER TABLE title_Ratings ADD FOREIGN KEY(tconst) REFERENCES title_facts (tconst)

```



5.3.2 Geração das Tabelas Brutas (sem restrições)

A saída do processo de ingestão e transformação dos dados foram tabelas SQL sem restrições, que foram armazenados num bando de dados SQL no Azure. Para tal, o componente “sink” do fluxo de transformação é responsável por apontar para o SQL server/database e carregar os dados na respectiva tabela. Nas figuras do item de “transformação” anteriormente apresentados, ele é o último componente como nome dim*.

Antes de criar um dataset, é preciso criar um serviço para vincular o repositório de armazenamento de dados ao ADF. Assim, o componente SINK implementa este serviço de conexão, ou *linked service*, com a base de dados do servidor ccoe. A configuração deste *linked service* segue abaixo:

Figura 16 - linked server coma base de dados no servidor ccoe

Edit linked service

 Azure SQL Database [Learn more](#) 

Name *

AzureSqlIDbMVP3

Description

Database MVP3

Connect via integration runtime * 

AutoResolveIntegrationRuntime

Connection string

Azure Key Vault

Account selection method 

☐ From Azure subscription ☒ Enter manually

Fully qualified domain name *

ccoe.database.windows.net

Database name *

mvp3DB

Authentication type *

SQL authentication

User name *

sabob


Password

Azure Key Vault

Password *

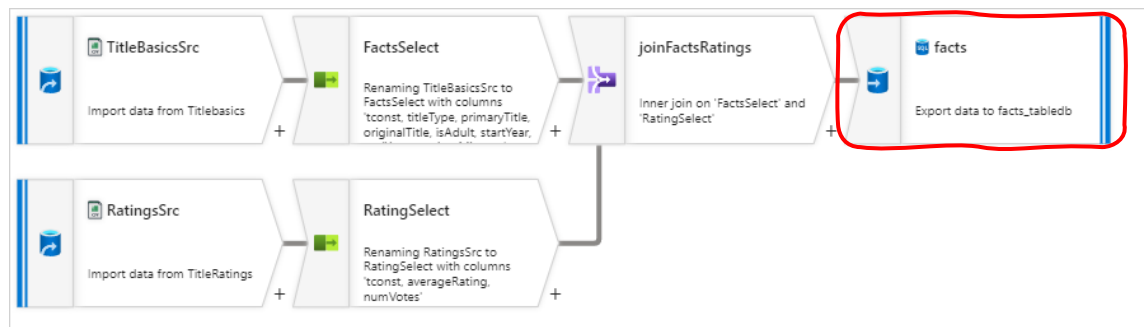
Apply

Cancel

 Test connection

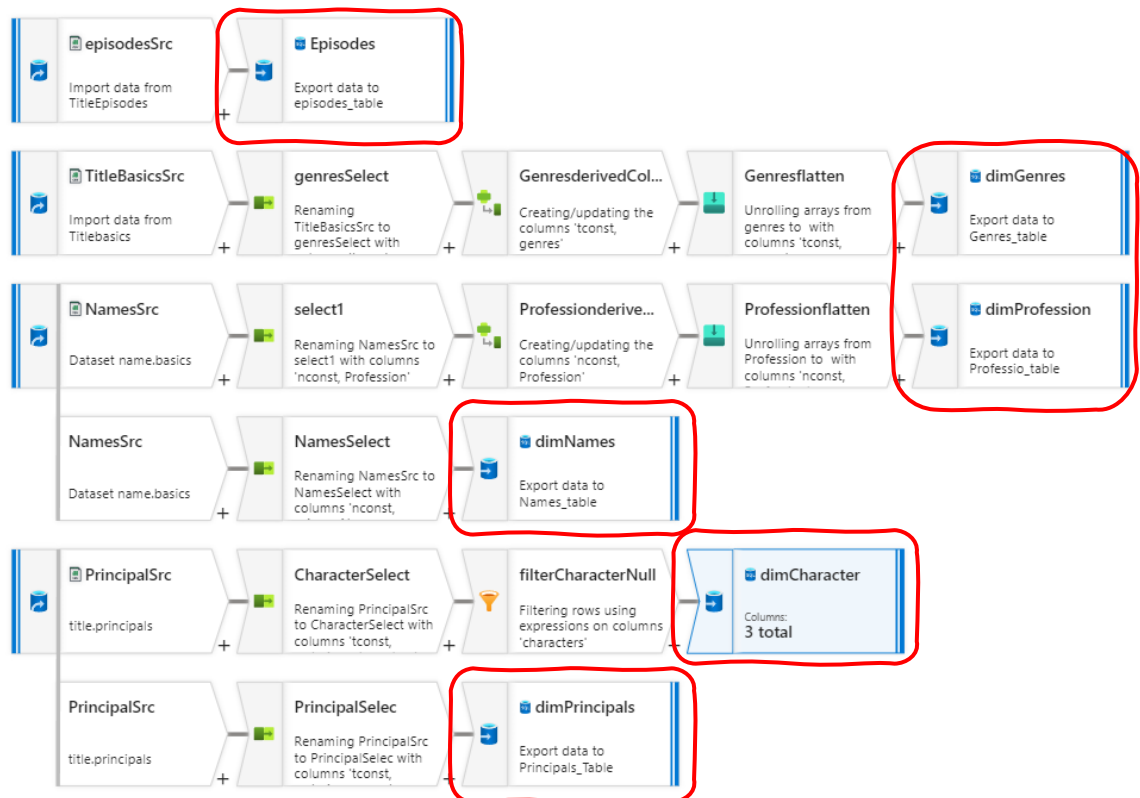
Para o fluxo de criação da tabela de fatos com o componente “Sink” está apresentado na figura, com nome “facts”.

Figura 17 - componente de saída da transformação SINK – tabela de fatos



Já para o fluxo de criação das tabelas de dimensão, temos vários sinks (um para cada tabela gerada) conforme mostrado na figura:

Figura 18 - componente de saída da transformação SINK – tabela de dimensões



5.3.3 Carga dos dados nas tabelas definitivas

A terceira etapa consistiu em carregar os dados das tabelas geradas pelo ADF (sem restrições) nas tabelas SQL criadas com as devidas restrições de chave Primária e estrangeira. Para isto foi utilizado o próprio SQL por meio de comandos “INSERT INTO” tendo como origem as tabelas oriundas do ADF, renomeadas para “*_old”. O script utilizado está mostrado na figura a seguir.

Figura 19 - Script INSERT INTO de carga final das tabelas

```
--INSERT INTO
INSERT INTO [dbo].[dim_title_facts]
SELECT * FROM [dbo].[dim_title_facts_old];

INSERT INTO [dbo].[dim_title_names]
SELECT * FROM [dbo].[dim_title_names_old];

INSERT INTO [dbo].[dim_Profession]
SELECT * FROM [dbo].[dim_Profession_old];

INSERT INTO [dbo].[dim_principals]
SELECT * FROM [dbo].[dim_principals_old];

INSERT INTO [dbo].[dim_episodes]
SELECT * FROM [dbo].[dim_episodes_old];

INSERT INTO [dbo].[dim_directors]
SELECT * FROM [dbo].[dim_directors_old];

INSERT INTO [dbo].[dim_episodes]
SELECT * FROM [dbo].[dim_episodes_old];

INSERT INTO [dbo].[dim_Characters]
SELECT * FROM [dbo].[dim_Characters_old];

INSERT INTO [dbo].[dim_genres]
SELECT * FROM [dbo].[dim_dim_genres_old];

INSERT INTO [dbo].[title_Ratings]
SELECT * FROM [dbo].[title_Ratings_old];
```

6 Resultados e Visualização de dados

Nas respostas abaixo consideramos popularidade como numero de voos do filme e pontuação como o valor do IMDB obtido.

6.1 Qual a popularidade dos filmes de James Bond

```
SELECT f.primaryTitle, SUM(r.numVotes) as Popularidade
FROM title_facts f inner join dim_Characters c on f.tconst=c.tconst
      join title_ratings r ON r.tconst= c.tconst
Where c.characters like 'James Bond' and f.titleType='movie'
GROUP by f.primaryTitle
order by Popularidade Desc
```

primaryTitle	Popularidade
Skyfall	716762
Casino Royale	678667
Quantum of Solace	462211
Spectre	456015
No Time to Die	428417
GoldenEye	265240
Die Another Day	225289
The World Is Not Enough	206468
Tomorrow Never Dies	201028
Goldfinger	197809
Dr. No	174824
From Russia with Love	141276
Thunderball	123931
You Only Live Twice	114603
The Spy Who Loved Me	113449
Live and Let Die	112628
Diamonds Are Forever	111339
Octopussy	110418
The Man with the Golden Gun	110394
Licence to Kill	109378
Moonraker	105995
For Your Eyes Only	105698
The Living Daylights	103325
A View to a Kill	102306
On Her Majesty's Secret Service	96554
Never Say Never Again	71231
10 Endrathukulla	1531
Mad Mission 3: Our Man from Bond Street	1105
Golden Chicken	1005
Deadly Hands of Kung Fu	388
One Fall	264
Goldenrock	79
Reflection of the Soul	57
The Shadow of Revenge	54
The Price of Loyalty	49
A Fool's Paradise	21
007: Shadows	21
Risque	12

6.2 Qual a pontuação dos filmes de James Bond

`SELECT f.primaryTitle, r.averageRating as Pontuacao`

`FROM title_facts f inner join dim_Characters c on f.tconst=c.tconst`

`join title_ratings r ON r.tconst= c.tconst`

Where c.characters like 'James Bond' and f.titleType='movie'

order by Pontuacao Desc

primaryTitle	Pontuacao
A Fool's Paradise	8.6
007: Shadows	8.4
Casino Royale	8.0
Skyfall	7.8
Goldfinger	7.7
Reflection of the Soul	7.6
The Shadow of Revenge	7.4
No Time to Die	7.3
From Russia with Love	7.3
Dr. No	7.2
GoldenEye	7.2
Golden Chicken	7.1
The Spy Who Loved Me	7.0
Thunderball	6.9
You Only Live Twice	6.8
Spectre	6.8
On Her Majesty's Secret Service	6.7
Live and Let Die	6.7
The Living Daylights	6.7
For Your Eyes Only	6.7
The Man with the Golden Gun	6.7
Licence to Kill	6.6
Quantum of Solace	6.6
Octopussy	6.5
Tomorrow Never Dies	6.5
Diamonds Are Forever	6.5
The Price of Loyalty	6.5
The World Is Not Enough	6.4
A View to a Kill	6.3
Moonraker	6.2
Never Say Never Again	6.1
Die Another Day	6.1
Mad Mission 3: Our Man from Bond Street	5.8
Deadly Hands of Kung Fu	5.3
Goldenrock	5.3
10 Endrathukulla	5.2
One Fall	5.1
Risque	4.2

6.3 Quais são os gêneros de filmes mais populares

SELECT g.genre, SUM(r.numVotes) As Popularidade

```
FROM dim_genres g INNER JOIN title_Ratings r on g.tconst=r.tconst
GROUP BY g.genre
ORDER BY NumFilmes desc
```

genre	Popularidade
Drama	727666892
Action	452560351
Comedy	425886436
Adventure	361059088
Crime	281332739
Thriller	204725179
Romance	158374971
Sci-Fi	155602172
Mystery	154278735
Horror	126988249
Fantasy	126255289
Animation	115494690
Biography	77677171
Family	63451833
History	39278113
Music	27457243
Documentary	26997430
War	26625783
Sport	21396309
Western	12207588
Musical	11015312
Short	9511853
Reality-TV	4474116
Film-Noir	3934127
Talk-Show	2241707
Game-Show	1863926
News	1299681
Adult	849689

6.4 Quais os gêneros com as melhores pontuações

```
SELECT g.genre, avg(r.averageRating) As MediaPontuação
from dim_genres g INNER JOIN title_Ratings r on g.tconst=r.tconst
group by g.genre
order by MediaPontuação desc
```

genre	MediaPontuação
History	7.345087
Documentary	7.268805
Biography	7.244111
Crime	7.132190

Mystery	7.125261
Adventure	7.119780
Animation	7.113325
Sport	7.104241
Family	7.085133
Fantasy	7.071544
Music	7.070868
Game-Show	7.055011
Drama	7.052183
Action	7.032853
War	7.022099
Comedy	6.998890
Reality-TV	6.982161
Western	6.979241
Talk-Show	6.906038
Romance	6.894680
Short	6.853250
News	6.822347
Sci-Fi	6.696568
Musical	6.658574
Film-Noir	6.464326
Thriller	6.418926
Adult	6.266529
Horror	6.127479

6.5 Quais os 10 filmes com maior popularidade de Steven Spielberg

`select top (10) f.primaryTitle, n.primaryName, r.numvotes as Popularidade`

`from title_ratings r inner join title_facts f on f.tconst=r.tconst`

`inner join dim_director d on d.tconst=f.tconst`

`INNER JOIN title_names n ON n.nconst=d.nconst`

`where n.primaryName='Steven Spielberg'`

`order by r.numVotes desc`

primaryTitle	primaryName	Popularidade
Saving Private Ryan	Steven Spielberg	1448909
Schindler's List	Steven Spielberg	1406454
Catch Me If You Can	Steven Spielberg	1047055
Jurassic Park	Steven Spielberg	1033739
Indiana Jones and the Raiders of the Lost Ark	Steven Spielberg	1012521
Indiana Jones and the Last Crusade	Steven Spielberg	791216
Jaws	Steven Spielberg	639216
Minority Report	Steven Spielberg	571395

Indiana Jones and the Temple of Doom	Steven Spielberg	522033
Indiana Jones and the Kingdom of the Crystal Skull	Steven Spielberg	480879

6.6 Quais os 10 filmes com maior pontuação de Steven Spielberg

```
select top (10) f.primaryTitle, n.primaryName, r.averagerating as Pontuacao
from title_ratings r inner join title_facts f on f.tconst=r.tconst
      inner join dim_director d on d.tconst=f.tconst
      INNER JOIN title_names n ON n.nconst=d.nconst
where n.primaryName='Steven Spielberg'
order by r.averagerating desc
```

primaryTitle	primaryName	Pontuacao
Schindler's List	Steven Spielberg	9.0
Saving Private Ryan	Steven Spielberg	8.6
Indiana Jones and the Raiders of the Lost Ark	Steven Spielberg	8.4
Indiana Jones and the Last Crusade	Steven Spielberg	8.2
Jurassic Park	Steven Spielberg	8.2
Catch Me If You Can	Steven Spielberg	8.1
Jaws	Steven Spielberg	8.1
E.T. the Extra-Terrestrial	Steven Spielberg	7.9
Murder by the Book	Steven Spielberg	7.7
The Color Purple	Steven Spielberg	7.7

6.7 Quais as 10 maiores médias de pontuação de diretores de filmes, com pelo menos 10000 votos mais de 5 filmes realizados?

```
SELECT top (10) n.primaryName, CAST(AVG(r.averagerating) AS DECIMAL(10,2)) as
MediaPontuaçãoDiretor, count(*) as NumFilmes
from title_ratings r inner join title_facts f on f.tconst=r.tconst
      inner join dim_director d on d.tconst=f.tconst
      INNER JOIN title_names n ON n.nconst=d.nconst
Where NumVotes >10000 and f.titleType='movie'
Group by n.primaryName
HAVING (COUNT(*) > 5)
Order by MediaPontuaçãoDiretor desc
```

primaryName	MediaPontuaçãoDiretor	NumFilmes
-------------	-----------------------	-----------

Ertem Egilmez	8.74	7
Sergio Leone	8.22	6
Christopher Nolan	8.20	12
Akira Kurosawa	8.06	17
Andrei Tarkovsky	8.00	7
Hayao Miyazaki	7.96	11
Quentin Tarantino	7.94	14
Frank Capra	7.94	8
Fritz Lang	7.93	7
Mani Ratnam	7.93	8

6.8 Qual é o tempo de execução típico para filmes de cada gênero?

```
SELECT g.genre, AVG(f.runtimeMinutes) as MediaMinutos
from dim_genres g inner join title_facts f on g.tconst=f.tconst
Where f.titleType='movie'
GROUP BY g.genre
Order by g.genre
```

genre	MediaMinutos
Action	100
Adult	80
Adventure	92
Animation	81
Biography	88
Comedy	92
Crime	94
Documentary	78
Drama	96
Family	91
Fantasy	93
Film-Noir	82
Game-Show	45
History	93
Horror	88
Music	87
Musical	98
Mystery	93
News	75
Reality-TV	84
Romance	98
Sci-Fi	90
Sport	88
Talk-Show	80
Thriller	95

War	95
Western	76

6.9 Quantos filmes foram feitos de cada gênero por ano entre 2020/22?

```
SELECT f.startYear, g.genre, COUNT(*) as FilmesporGenero
from dim_genres g inner join title_facts f on g.tconst=f.tconst
Where f.titleType='movie' and f.startYear is not null
GROUP BY f.startYear, g.genre
Order by f.startYear DESC
```

startYear	genre	FilmesporGenero
2020	Biography	427
2020	Thriller	1204
2020	Game-Show	3
2020	Western	53
2020	Action	1013
2020	Documentary	5167
2020	Fantasy	398
2020	Animation	353
2020	Romance	801
2020	Music	441
2020	Sport	313
2020	Drama	4972
2020	News	14
2020	Musical	150
2020	War	150
2020	Adventure	555
2020	Mystery	481
2020	Adult	203
2020	Talk-Show	24
2020	Reality-TV	74
2020	Horror	1245
2020	Family	428
2020	Sci-Fi	384
2020	History	322
2020	Crime	650
2020	Comedy	2260
2021	News	10
2021	Action	1295
2021	Drama	5955
2021	Musical	187
2021	Mystery	542
2021	Game-Show	2
2021	Western	61
2021	Horror	1401

2021	Sci-Fi	441
2021	Comedy	2461
2021	Romance	935
2021	Fantasy	438
2021	Sport	336
2021	Music	434
2021	Crime	796
2021	Talk-Show	21
2021	Animation	431
2021	Family	428
2021	Reality-TV	35
2021	History	325
2021	War	109
2021	Thriller	1461
2021	Biography	504
2021	Adult	111
2021	Adventure	598
2021	Documentary	5448
2022	Talk-Show	10
2022	Mystery	650
2022	Musical	177
2022	Biography	502
2022	Fantasy	486
2022	Family	513
2022	Thriller	1859
2022	Action	1396
2022	Adventure	660
2022	Western	70
2022	Animation	391
2022	Sci-Fi	456
2022	News	11
2022	Romance	1109
2022	Horror	1640
2022	Documentary	5270
2022	Music	383
2022	War	117
2022	History	357
2022	Reality-TV	30
2022	Crime	958
2022	Sport	283
2022	Adult	86
2022	Game-Show	3
2022	Drama	6608
2022	Comedy	3114

6.10 Quem são os atores que interpretaram 'James Bond' em um filme?

```
SELECT f.primaryTitle, f.originalTitle, n.primaryName, characters, f.startYear
FROM title_names n inner join dim_Characters c on c.nconst=n.nconst
      join title_facts f on f.tconst=c.tconst
WHERE c.characters like 'James Bond' and f.titleType='movie'
ORDER BY f.startYear
```

primaryTitle	originalTitle	primaryName	characters	startYear
Death Is Forever	Death Is Forever	Tom Smith	James Bond	NULL
Goldpingas	Goldpingas	Jack Bilsborough	James Bond	NULL
Death Collector	Death Collector	Liam Fountain	James Bond	NULL
Dr. No	Dr. No	Sean Connery	James Bond	1962
From Russia with Love	From Russia with Love	Sean Connery	James Bond	1963
Goldfinger	Goldfinger	Sean Connery	James Bond	1964
Thunderball	Thunderball	Sean Connery	James Bond	1965
You Only Live Twice	You Only Live Twice	Sean Connery	James Bond	1967
On Her Majesty's Secret Service	On Her Majesty's Secret Service	George Lazenby	James Bond	1969
Diamonds Are Forever	Diamonds Are Forever	Sean Connery	James Bond	1971
Live and Let Die	Live and Let Die	Roger Moore	James Bond	1973
The Man with the Golden Gun	The Man with the Golden Gun	Roger Moore	James Bond	1974
The Spy Who Loved Me	The Spy Who Loved Me	Roger Moore	James Bond	1977
Deadly Hands of Kung Fu	Li san jiao wei zhen di yu men	Alexander Grand	James Bond	1977
Moonraker	Moonraker	Roger Moore	James Bond	1979
For Your Eyes Only	For Your Eyes Only	Roger Moore	James Bond	1981
Never Say Never Again	Never Say Never Again	Sean Connery	James Bond	1983
Octopussy	Octopussy	Roger Moore	James Bond	1983
Mad Mission 3: Our Man from Bond Street	Zui jia pai dang 3: Nu huang mi ling	Jean Mersant	James Bond	1984
A View to a Kill	A View to a Kill	Roger Moore	James Bond	1985
The Living Daylights	The Living Daylights	Timothy Dalton	James Bond	1987
Licence to Kill	Licence to Kill	Timothy Dalton	James Bond	1989
GoldenEye	GoldenEye	Pierce Brosnan	James Bond	1995
Tomorrow Never Dies	Tomorrow Never Dies	Pierce Brosnan	James Bond	1997
The World Is Not Enough	The World Is Not Enough	Pierce Brosnan	James Bond	1999
Goldenrock	Goldenrock	Kristoffer Hatlestad	James Bond	1999
Bharat India Hindustan	Bharat India Hindustan	Biswajeet Chatterjee	James Bond	2000
Die Another Day	Die Another Day	Pierce Brosnan	James Bond	2002
Golden Chicken	Gam gai	Eric Tsang	James Bond	2002
Casino Royale	Casino Royale	Daniel Craig	James Bond	2006
Quantum of Solace	Quantum of Solace	Daniel Craig	James Bond	2008
The Price of Loyalty	The Price of Loyalty	Tom Smith	James Bond	2008

The Shadow of Revenge	The Shadow of Revenge	Tom Smith	James Bond	2010
Skyfall	Skyfall	Daniel Craig	James Bond	2012
A Fool's Paradise	A Fool's Paradise	Mac Jolly	James Bond	2013
Reflection of the Soul	Reflection of the Soul	Tom Smith	James Bond	2013
Risque	Risque	Paul Cusack	James Bond	2014
10 Endrathukulla	10 Endrathukulla	Vikram	James Bond	2015
Spectre	Spectre	Daniel Craig	James Bond	2015
One Fall	One Fall	Marcus Dean Fuller	James Bond	2016
007: Shadows	Shadows	Mac Jolly	James Bond	2020
No Time to Die	No Time to Die	Daniel Craig	James Bond	2021

6.11 Quantas vezes eles fizeram o papel de 'James Bond'?

`SELECT n.primaryName, count(*) as filmesRealizados`

`FROM title_names n inner join dim_Characters c on c.nconst=n.nconst`

`join title_facts f on f.tconst=c.tconst`

`WHERE c.characters like 'James Bond' and f.titleType='movie'`

primaryName	filmesRealizados
Eric Tsang	1
George Lazenby	1
Jack Bilsborough	1
Jean Mersant	1
Kristoffer Hatlestad	1
Liam Fountain	1
Marcus Dean Fuller	1
Paul Cusack	1
Alexander Grand	1
Biswajeet	
Chatterjee	1
Vikram	1
Timothy Dalton	2
Mac Jolly	2
Pierce Brosnan	4
Tom Smith	4
Daniel Craig	5
Roger Moore	7
Sean Connery	7

6.12 Quantos filmes existem em cada gênero?

`SELECT genre, COUNT(*) As NumFilmes`

`FROM dim_genres`

GROUP BY by genre

genre	NumFilmes
Animation	513029
Sci-Fi	109979
Western	30148
Family	750829
Thriller	170300
Musical	88704
Film-Noir	886
Talk-Show	1238672
History	148923
Documentary	956911
Horror	181758
Game-Show	356825
Adult	315756
Fantasy	206108
Biography	108840
Comedy	2033882
Crime	417089
Action	419221
Mystery	203965
Reality-TV	564381
Drama	2888489
Romance	969600
News	892443
Sport	240835
Short	1113099
Adventure	398238
War	34360
Music	390803

7 Conclusão

O objetivo deste projeto foi o de realizar uma análise dos títulos (filmes, seriados de TV) publicados e responder a diversas perguntas mais comuns sobre o mercado de mídia, utilizando as informações disponibilizadas pela plataforma IMDB.

A execução deste projeto envolveu a realização das seguintes atividades:

- Entender os dados no database disponibilizado pelo IMDb.
- Modelar o banco de dados usando a técnica o modelo Entidade-Relacionamento (ER) e diagramas de esquema lógico relacional.
- Projetar um banco de dados relacional

- Criar um servidor relacional AZURE SQL Server, e um banco de dados relacional com as devidas restrições de chave primária e estrangeira (para onde os dados transformados foram carregados).
- Criar um repositório de dados no Azure e armazenar os dados oriundos do IMDB nele;
- Realizar um ETL, utilizando o Azure Data Factory, onde extraímos os dados dos arquivos tsv (separados por tabulações) e transformando-os, e carregando-os em tabelas normalizadas e reestruturadas, segundo a modelagem prevista;
- Carregar os dados gerados nas tabelas com restrições de chave primária e estrangeira, e;
- Responder as questões colocadas

Sobre os dados fonte do database IMDB

O conjunto de dados IMDB é um conjunto de dados onde existem alguns problemas com dados faltantes que afetaram a forma como adicionamos restrições ao banco de dados.

Há dados faltantes no conjunto de dados da IMDB, principalmente nos arquivos name.basics.tsv.gz e title.basics.tsv.gz. Esses dados ausentes causam problemas quando tentamos impor certas restrições de chave estrangeira.

Este projeto mostrou apenas algumas possibilidades do que pode ser feito com esses dados do IMDB. Com estes dados poderíamos realizar diversas outras análises. Uma adição interessante a estes dados seria a inclusão dos dados das bilheterias obtidas pelos títulos.

Por fim, uma possível extensão do uso destes dados seria investigar mais detalhadamente as tendências, realizando análises estatísticas e possivelmente até usando alguns algoritmos de aprendizado de máquina.

