

Parte 1. Objetivo e apresentação do MVP

Ao escolher o dataset do IMBD, fiquei com curiosidade de entender como funciona a estrutura hierárquica desses projetos e qual o padrão de cargos envolvidos em filmes extremamente bem ranqueados, além disso quais são os atores que mais aparecem dentre esses filmes?

Para isso será utilizado um esquema em que traga os atores e cargos principais como fatos e toda a estrutura em volta como dimensões.

Parte 2. Estrutura conceitual: Esquema estrela

Como ideia inicial de planejamento de estruturação lógica do projeto, é notado que existem algumas “tabelas” predefinidas pelo próprio sistema IMDB, sendo elas identificadas como esquema relacional induzido.

Como tradução necessária ao esquema inicial, fica-se imaginado um esquema estrela em que os “protagonista” ou “fatos” são os dados da pergunta feita na parte 1 do projeto encontrados no link disponibilizado pelo professor (<https://datasets.imdbws.com/>), sendo assim:

1. Tabela de Fatos:

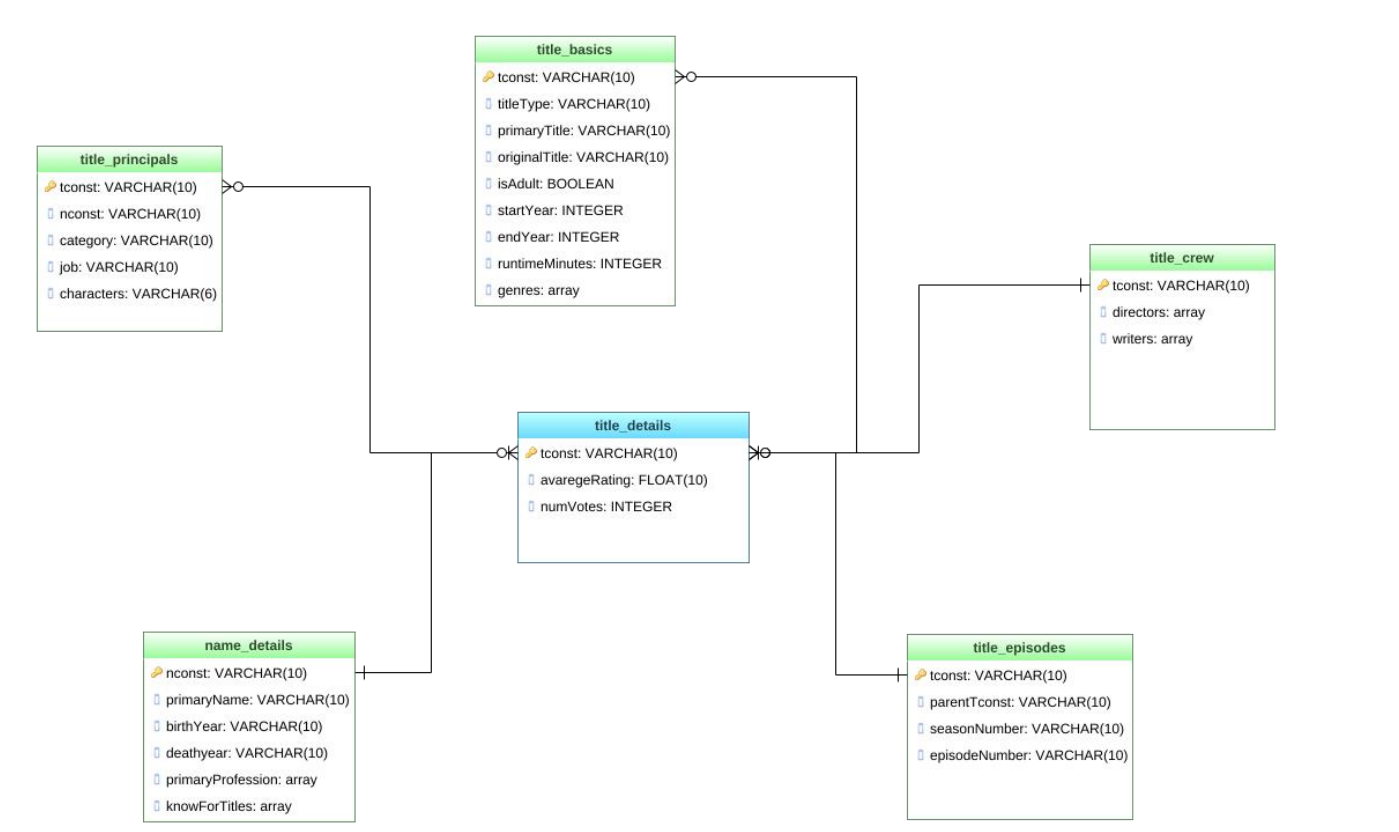
- **Título Detalhes** (title_ratings):
 - tconst (string) - chave estrangeira para a tabela title.basics.tsv.gz
 - averageRating (float) - média ponderada das avaliações individuais dos usuários
 - numVotes (integer) - número de votos recebidos pelo título

2. Tabelas Dimensionais:

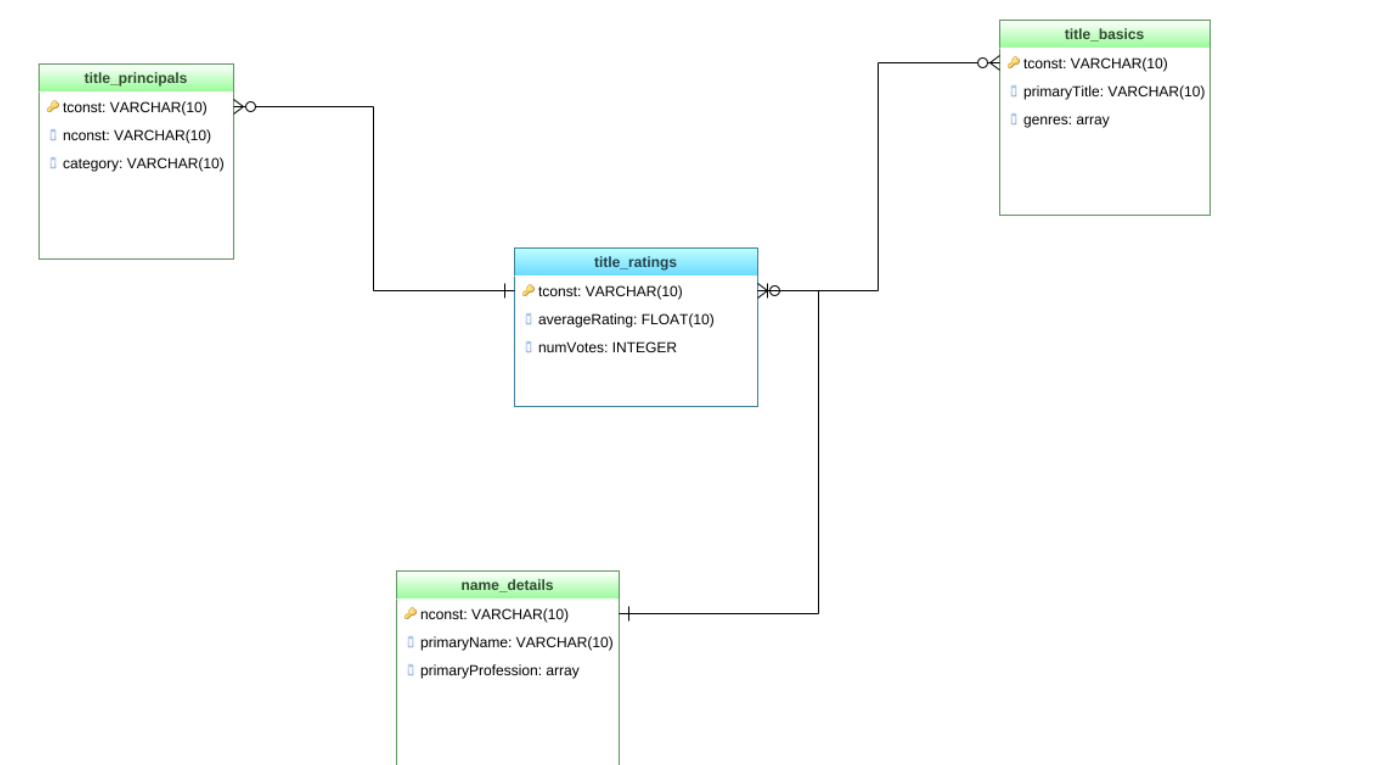
- **Título Básico** (title_basics):
 - tconst (string) - chave primária
 - primaryTitle (string) - título mais popular / título usado pelos cineastas
 - genres (array) - até três gêneros associados ao título
- **Principais Colaboradores** (title_principals):
 - tconst (string) - chave primária
 - nconst (string) - chave estrangeira para a tabela name.basics.tsv.gz
 - category (string) - categoria do trabalho da pessoa
- **Detalhes dos nomes** (name.basics):
 - nconst (string) - chave primária
 - primaryName (string) - nome pelo qual a pessoa é mais frequentemente creditada
 - primaryProfession (array) - top-3 profissões da pessoa

A tabela de fatos contém as métricas de avaliação (averageRating e numVotes) para os títulos, enquanto as tabelas dimensionais contém informações detalhadas sobre títulos, equipes de produção, colaboradores e pessoas envolvidas. Isso permite que a análise de como as características dos filmes se relacionam com as avaliações, bem como os padrões de equipe de produção e a relevância dos colaboradores.

De uma maneira da qual fossem englobadas e enquadradas todas as informações cruas do dataset inicial, a abordagem em esquema estrela resultada seria a seguinte:



Como gostaria de aumentar o desempenho e mostrar os dados de maneira mais limpa, decidi seguir com um esquema reduzido, do qual serão excluídas as colunas e tabelas que não respondam exatamente a pergunta do objetivo declarado na primeira parte:



Parte 3. ETL/Dados para banco em nuvem

3.1 Extração e transformação dos dados

Utilizando o Google cloud e seu serviço de cluster Google Proc, será possível a criação do cluster que armazenará os arquivos apontados através do esquema estrela, sendo eles:

- [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](#)

Sendo assim, é possível armazena-los através das ferramentas de upload oferecidas e de manipulação por exemplo, a que será utilizada Hive.

Como os arquivos disponibilizados são incorporados em extensão tsv e compactados em GZ, foi necessária a descompactação dos mesmos e para a importação no HDFS primeiramente serem colocados no Google Storage e a transferência para o cluster com o comando:

```
hadoop distcp gs://dataproc-staging-us-central1-345813319426-s9wcsy4e/google-cloud-dataproc-metainfo/5cb1e64a-daaa-482b-9f1c-46f13a234f81/IMDB/*.tsv hdfs:///user/dataproc/imdb/
```

3.1 Extração e transformação dos dados – Alteração na ferramenta e método

Após o plantão de duvidas, a forma de inserção e ETL dos dados no GC se tornou mais fácil e interativa, sem a necessidade da inserção no HDFS e manipulação através do HIVE.

Partindo dos passos antes feitos de inserção dos dados no Google Storage, a intenção agora é utilizar o Data Fusion com o BigQuery funcionando como banco de dados como passo final para a entrada dos dados. Será necessário para a alocação dos dados nos buckets somente o envio dos arquivos para os novos buckets normal e temporário.

← Detalhes do bucket

ATUALIZAR SAIBA MAIS

dataproc-staging-us-central1-345813319426-s9wcsy4e

Local

Classe de armazenamento

Acesso público

Proteção

us-central1 (Iowa)

Standard

Sujeito a ACLs de objeto

Nenhum

OBJETOS

CONFIGURAÇÃO

PERMISSÕES

PROTEÇÃO

CICLO DE VIDA

OBSERVABILIDADE

RELATÓRIOS DE INVENTÁRIO

Intervalos

dataproc-staging-us-central1-345813319426-s9wcsy4e

google-cloud-dataproc-metainfo

5cb1e64a-daaa-482b-9f1c-46f13a234f81

IMDB

FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR PASTA

TRANSFERIR DADOS

GERENCIAR RETENÇÕES

FAZER O DOWNLOAD

EXCLUIR

Filtrar apenas pelo prefixo do nome

Filtro

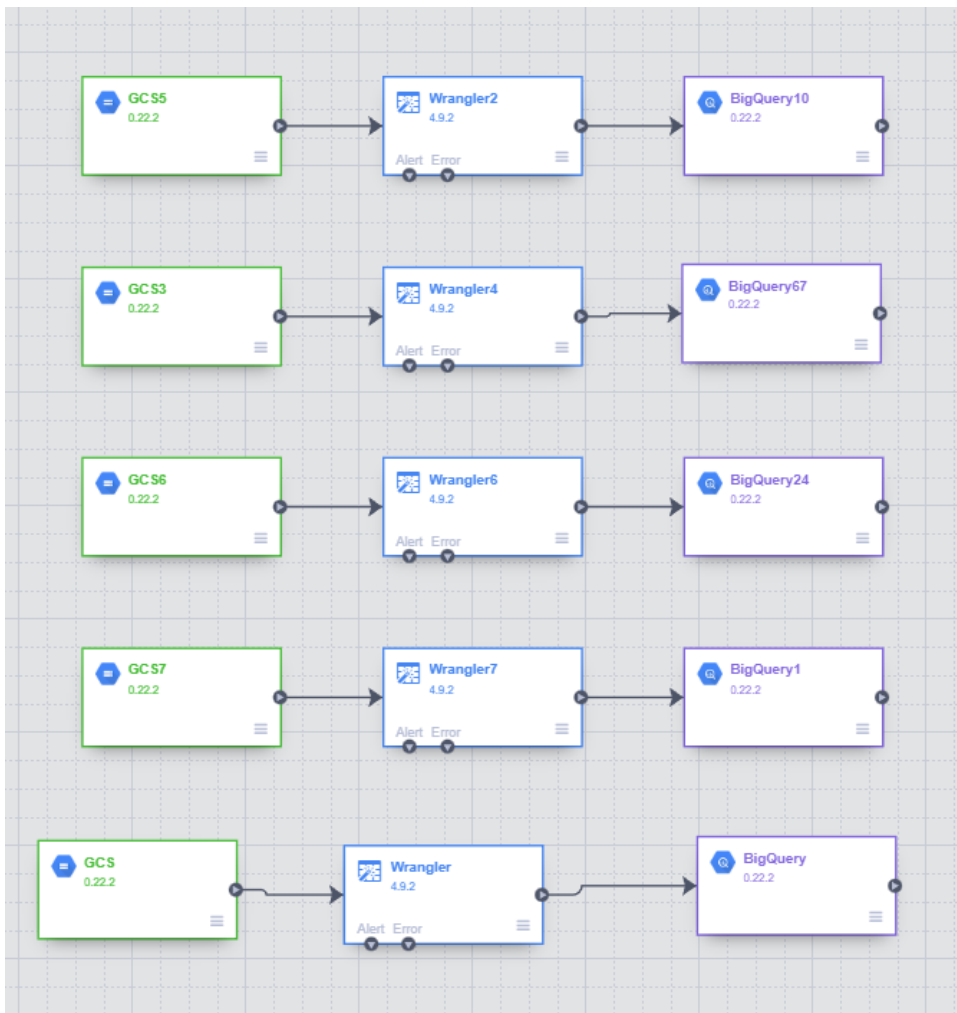
Filtrar objetos e pastas

Mostrar dados excluídos

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público	Histórico de versões	Criptografia
<input type="checkbox"/>	name.basics.tsv	742,3 MB	application/octet-stream	29 de ago. de 2023 23:50:41	Standard	29 de ago. de 2023 23:50:41	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.akas.tsv	1,7 GB	application/octet-stream	29 de ago. de 2023 23:52:00	Standard	29 de ago. de 2023 23:52:00	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.basics.tsv	825,1 MB	application/octet-stream	29 de ago. de 2023 23:51:49	Standard	29 de ago. de 2023 23:51:49	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.crew.tsv	318,3 MB	application/octet-stream	29 de ago. de 2023 23:52:17	Standard	29 de ago. de 2023 23:52:17	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.episode.tsv	191 MB	application/octet-stream	29 de ago. de 2023 23:52:16	Standard	29 de ago. de 2023 23:52:16	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.principals.tsv	2,4 GB	application/octet-stream	29 de ago. de 2023 23:52:43	Standard	29 de ago. de 2023 23:52:43	Não público	—	Gerenciada pelo Gc
<input type="checkbox"/>	title.ratings.tsv	22,2 MB	application/octet-stream	29 de ago. de 2023 23:51:52	Standard	29 de ago. de 2023 23:51:52	Não público	—	Gerenciada pelo Gc

Assim, como segundo passo após o upload dos dados TSV nos buckets do google storage, o passo seguinte é a transformação dos dados dentro do Data fusion utilizando o Wrangle.

Logo, como primeiros passos, a exclusão de algumas tabelas que não fazem sentido:
[title.akas.tsv.gz](#), [title.episode.tsv.gz](#) e [title.crew](#)



No processo de transformação foram excluídas todas as colunas e tabelas que não fazem sentido para o objetivo.

Recipe

```
1 drop :birthYear
2 drop :deathYear
3 drop :knownForTitles
```

WRANGLE

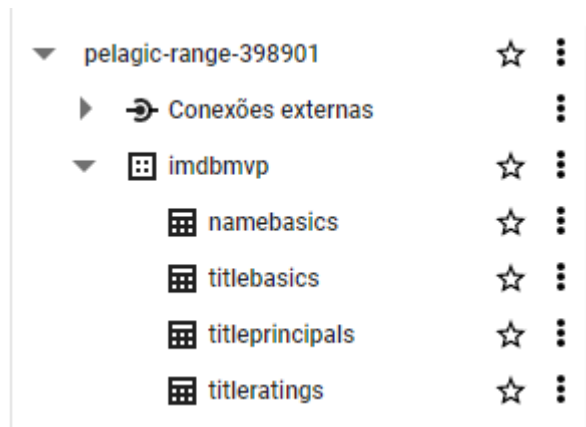
Recipe

```
1 drop :originalTitle,:isAdult,:startYear,:endYear,:runtimeMinutes,:titleType
```

WRANGLE

Como toda informação remanescente das tabelas pode ser importante para as consultas, não foram necessárias adequações dos dados inclusos nas colunas.

Dataset inicial no BigQuery:



3.3 Qualidade dos dados:

3.3.1 Análise sobre qualidade dos dados

titlebasics					
CONSULTA					
COMPARTILHAR COPIAR SNAPSHOT EXCLUIR EXPORTAR					
ESQUEMA		DETALHES	VISUALIZAR	LINHAGEM	PERFIL DE DADOS
QUALIDADE DOS DADOS					
Linha	tconst	titleType	primaryTitle	originalTitle	
1	tt6825766	tvEpisode	Episode #1.1417	Episode #1.1417	
2	tt7131918	tvEpisode	DAY6, SEVENTEEN, Baek a Yeo...	DAY6, SEVENTEEN, Baek a Yeo...	
3	tt23034056	tvEpisode	Episode #1.4	Episode #1.4	
4	tt2312592	tvEpisode	Episode #1.87	Episode #1.87	
5	tt23768910	videoGame	The Baptized	The Baptized	
6	tt13739088	tvEpisode	Episode #1.136	Episode #1.136	
7	tt14021266	tvEpisode	Episode #1.513	Episode #1.513	
8	tt16431512	movie	Carriage Hill	Carriage Hill	
9	tt0195953	movie	Norman and God	Norman and God	
10	tt3985070	tvEpisode	Episode #1.51	Episode #1.51	
11	tt4013110	tvEpisode	Episode #2.26	Episode #2.26	
12	tt4031514	tvEpisode	Episode #1.135	Episode #1.135	
13	tt24852638	movie	Popular	Popular	
14	tt1154652	tvEpisode	Eggs	Eggs	
15	tt7704026	tvEpisode	Breaking All the Rules	Breaking All the Rules	
16	tt18283258	tvEpisode	Episode #1.2774	Episode #1.2774	
17	tt18332164	tvEpisode	Episode #1.3360	Episode #1.3360	

Para a tabela titlebasics, as coluna que mais apresentam inconsistências são as relacionadas a títulos.

Existem muitas formas diferentes de representar os números dos episódios, podendo ser eles com muitas casas decimais após o primeiro número, tornando-se muito difícil entender a qual episodio o subconjunto está se referindo.

Uma maneira de resolver é buscando seus verdadeiros significados através de algum padrão encontrado nas linhas, e com isso, criar uma categorização mais simples através da substituição de todos os números após “Episode #”

Quanto as outras tabelas, todas as informações utilizadas estão bem definidas e de fácil acesso:

titeratings		CONSULTA	COMPARTILHAR
ESQUEMA	DETALHES	VISUALIZAR	LINHAGEM
Linha	tconst	averageRating	numVotes
1	tt0000024	4.2	117
2	tt0000025	3.9	45
3	tt0000036	4.4	611
4	tt0000037	4.4	68
5	tt0000038	4.2	204
6	tt0000040	4.0	68
7	tt0000044	3.9	48
8	tt0000052	4.2	105
9	tt0000076	4.4	541
10	tt0000078	3.7	88
11	tt0000108	4.4	550
12	tt0000109	4.5	531
13	tt0000110	4.4	537
14	tt0000111	4.4	553
15	tt0000112	4.5	530

titleprincipals

CONSULTA

COMPARTILHAR

COPIAR

SNAPSHOT

ESQUEMA

DETALHES

VISUALIZAR

LINHAGEM

PERFIL DE DADOS

QUALI

Linha	tconst	nconst	category
1	tt11528406	nm9506111	actor
2	tt13608964	nm0001769	actor
3	tt12721794	nm7255714	actor
4	tt5711016	nm8124389	actor
5	tt0003599	nm0168621	actor
6	tt14748922	nm6021864	actor
7	tt3218912	nm5977384	actor
8	tt14600862	nm8135362	actor
9	tt12477560	nm3102870	actor
10	tt1249105	nm0585171	actor
11	tt26426096	nm1340051	actor
12	tt6932758	nm9018281	actor

Linha	nconst	primaryName
1	nm12841848	Ali Shakeri Zand
2	nm11806888	Lily Beer
3	nm7946759	Trish Robertson
4	nm4991335	Marshall Martinez
5	nm2178491	Doug McGovern
6	nm14675637	Samet Yüce
7	nm8450815	Leon Blanda
8	nm5443765	Louise Rodgers
9	nm9485566	David Graf
10	nm11637590	Shane Newsham
11	nm12433638	Carvajal Carlos
12	nm6749959	DeAngelo Alexander
13	nm6882452	Isabelle Brandauer

As definições dos subconjuntos estão claros e totalmente entendíveis, tornando muito mais fácil os processos de consulta em porterior.

3.3.2 Catalogação do dataset e de suas tabelas.

Foram adicionadas descrições detalhadas a todas as tabelas e suas colunas.

titlebasics

☆ STAR

+ ANEXAR TAGS

🔍 ABRIR NO BIGQUERY

📊 EXPLORAR COM O LOOKER STUDIO

⋮

SAIBA MAIS

📄

us

imdbmvp

Administrador: ✎

DETALHES

ESQUEMA

LINHAGEM

PERFIL DE DADOS

QUALIDADE DOS DADOS

Detalhes da Tabela do BigQuery

Tipo

TABLE

Sistema

BIGQUERY

Tipo de tabela

Tabela do BigQuery

Data/hora de criação

13 de set. de 2023 15:47:07

Horário da última modificação

13 de set. de 2023 15:47:07

Tempo de expiração

Nunca

Local

us

Consultas (últimos 30 dias)

4

URL do recurso

[pelagic-range-398901.imdbmvp.titlebasics](#)

Marcadores

[Editar no BigQuery](#)

Nome totalmente qualificado

bigquery:pelagic-range-398901.imdbmvp.titlebasics

Descrição


Visão geral

Tabela com detalhamento sobre titulo do projeto com informações como:


Titulo original, titulo popular, minutagem, data de inicio das gravações e termino e chave primaria.

✎


EDITAR VISÃO GERAL


 Filtro Insira o nome ou o valor da propriedade


<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas 	Descrição
<input type="checkbox"/>	tconst	STRING	NULLABLE					Chave primaria que representa nome do titulo
<input type="checkbox"/>	titleType	STRING	NULLABLE					Tipo de titulo (Filme/Serie/Doc)
<input type="checkbox"/>	primaryTitle	STRING	NULLABLE					Titulo popular
<input type="checkbox"/>	originalTitle	STRING	NULLABLE					Titulo original
<input type="checkbox"/>	isAdult	INTEGER	NULLABLE					É um projeto para adultos?
<input type="checkbox"/>	startYear	STRING	NULLABLE					Inicio das gravações
<input type="checkbox"/>	endYear	STRING	NULLABLE					Termino das gravações
<input type="checkbox"/>	runtimeMinutes	STRING	NULLABLE					Tempo do projeto
<input type="checkbox"/>	genres	STRING	NULLABLE					Genero


 Filtro Insira o nome ou o valor da propriedade

<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas 	Descrição
<input type="checkbox"/>	nconst	STRING	NULLABLE					Chave primaria para nome do colaborador
<input type="checkbox"/>	primaryName	STRING	NULLABLE					Nome do colaborador
<input type="checkbox"/>	primaryProfession	STRING	NULLABLE					Profissões que é capaz de exercer

 Filtro Insira o nome ou o valor da propriedade

<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas 	Descrição
<input type="checkbox"/>	tconst	STRING	NULLABLE					Chave para titulo do projeto
<input type="checkbox"/>	nconst	STRING	NULLABLE					Chave para nome do colaborador envolvido
<input type="checkbox"/>	category	STRING	NULLABLE					Categoria do cargo exercido pelo colaborador
<input type="checkbox"/>	job	STRING	NULLABLE					Cargo especifico exercido pelo colaborador
<input type="checkbox"/>	characters	STRING	NULLABLE					Personagem interpretado. Podendo ser N

 Filtro Insira o nome ou o valor da propriedade

<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas 	Descrição
<input type="checkbox"/>	tconst	STRING	NULLABLE					Chave para nome do projeto
<input type="checkbox"/>	averageRating	FLOAT	NULLABLE					Avaliação média
<input type="checkbox"/>	numVotes	INTEGER	NULLABLE					Número de votos

Parte 4. Resposta ao objetivo

A primeira consulta necessária é descobrir quais são os filmes mais bem ranqueados do dataset disponibilizado pelo IMDB e logo em seguida criar uma tabela com os resultados:

```
1 SELECT t1.primarytitle, t2.averagerating, t2.numvotes
2 FROM `pelagic-range-398901.imdbmvp.titlebasics` t1
3 INNER JOIN `pelagic-range-398901.imdbmvp.titleratings` t2
4 ON t1.tconst = t2.tconst
5 WHERE t2.numvotes > 1000000
6 ORDER BY t2.averagerating DESC;
```

Resultados da consulta				
INFORMAÇÕES DO JOB		RESULTADOS	JSON	DETALHES DA EXECUÇÃO
Linha	primarytitle	averagerating	numvotes	
1	Breaking Bad	9.5	2023816	
2	The Shawshank Redemption	9.3	2787176	
3	Game of Thrones	9.2	2194822	
4	The Godfather	9.2	1940872	
5	The Dark Knight	9.0	2766563	
6	The Lord of the Rings: The Retu...	9.0	1909611	
7	Schindler's List	9.0	1402085	
8	The Godfather Part II	9.0	1318901	
9	Pulp Fiction	8.9	2138214	
10	Friends	8.9	1041248	
11	Inception	8.8	2456328	

```
CREATE OR REPLACE TABLE `pelagic-range-398901.imdbmvp.top20filmes`
AS
SELECT t1.primarytitle, t2.averagerating, t2.numvotes, t1.tconst
FROM `pelagic-range-398901.imdbmvp.titlebasics` t1
INNER JOIN `pelagic-range-398901.imdbmvp.titleratings` t2
ON t1.tconst = t2.tconst
WHERE t2.numvotes > 1000000
ORDER BY t2.averagerating DESC
LIMIT 20;
```

top20filmes					
CONSULTA					
COMPARTILHAR					
COPIAR					
SNAPSHOT					
ESQUEMA	DETALHES	VISUALIZAR	LINHAGEM	PERFIL DE DADOS	QUALIC
Linha	primarytitle	averagerating	numvotes	tconst	
1	Breaking Bad	9.5	2023816	tt0903747	
2	The Shawshank Redemption	9.3	2787176	tt0111161	
3	The Godfather	9.2	1940872	tt0068646	
4	Game of Thrones	9.2	2194822	tt0944947	
5	The Godfather Part II	9.0	1318901	tt0071562	
6	Schindler's List	9.0	1402085	tt0108052	
7	The Lord of the Rings: The Retu...	9.0	1909611	tt0167260	
8	The Dark Knight	9.0	2766563	tt0468569	
9	Friends	8.9	1041248	tt0108778	
10	Pulp Fiction	8.9	2138214	tt0110912	
11	Inception	8.8	2456328	tt1375666	
12	Forrest Gump	8.8	2167837	tt0109830	
13	The Lord of the Rings: The Fell...	8.8	1937797	tt0120737	
14	Fight Club	8.8	2221698	tt0137523	
15	The Lord of the Rings: The Two...	8.8	1723185	tt0167261	
16	Goodfellas	8.7	1209339	tt0099685	
17	One Flew Over the Cuckoo's Nest	8.7	1040058	tt0073486	
18	The Matrix	8.7	1982650	tt0133093	
19	Interstellar	8.7	1971755	tt0816692	
20	Stranger Things	8.7	1268334	tt4574334	

Agora, entender a estrutura hierárquica:

```
SELECT t1.primarytitle, t2.category, COUNT(t2.category) AS category_count
FROM `pelagic-range-398901.imdbmvp.top20filmes` t1
LEFT JOIN `pelagic-range-398901.imdbmvp.titleprincipals` t2
ON t1.tconst = t2.tconst
GROUP BY t1.primarytitle, t2.category
ORDER BY t1.primarytitle, category_count DESC;
```

Do qual tenho o objetivo de (em ordem decrescente) entender quantos atores, atrizes, escritores, diretores e produtores existem em cada uma das produções de cinema.

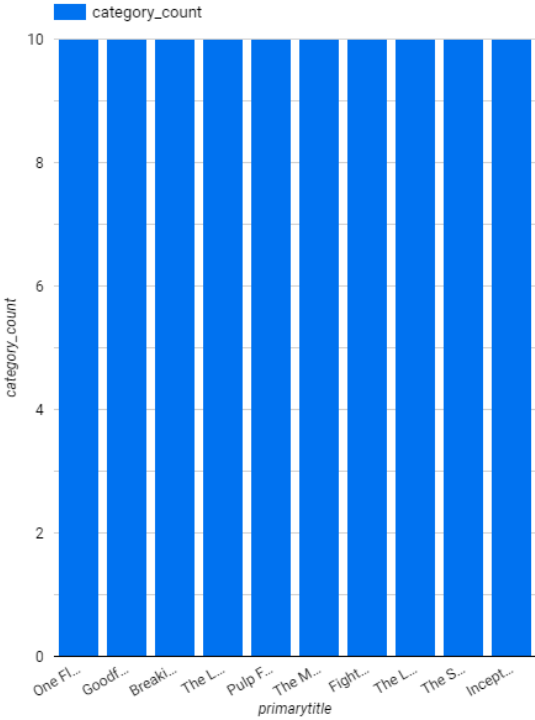
Resultando em uma consulta de 105 linhas, segue uma lista de correspondências encontradas:

primarytitle	category	category_count
Breaking Bad	actor	7
Breaking Bad	actress	2
Breaking Bad	writer	1
Fight Club	actor	4
Fight Club	producer	3
Fight Club	writer	2
Fight Club	director	1
Forrest Gump	producer	3
Forrest Gump	actor	2
Forrest Gump	actress	2
Forrest Gump	writer	2
Forrest Gump	director	1
Friends	actor	5
Friends	actress	3
Friends	writer	2
Game of Thrones	actress	4
Game of Thrones	actor	4
Game of Thrones	writer	2
Goodfellas	actor	3
Goodfellas	editor	2
Goodfellas	writer	1
Goodfellas	actress	1
Goodfellas	director	1
Goodfellas	cinematographer	1
Goodfellas	producer	1

Com estes números, é possível encontrar um padrão que possa apresentar alguma regra ou padrão dentre os filmes apresentados no dataset:

O total de pessoas em todas as equipes de alguma maneira sempre é de 10 pessoas, sendo elas divididas entre atores, atrizes, escritores, produtores, diretores ou cinematografista.

	primarytitle	category_count
1.	One Flew Over the Cuckoo's Nest	10
2.	Goodfellas	10
3.	Breaking Bad	10
4.	The Lord of the Rings: The Return of the King	10
5.	Pulp Fiction	10
6.	The Matrix	10
7.	Friends	10
8.	Forrest Gump	10
9.	Schindler's List	10
10.	The Dark Knight	10
11.	The Godfather	10
12.	Stranger Things	10
13.	Game of Thrones	10
14.	The Lord of the Rings: The Fellowship Ring	10
15.	Interstellar	10
16.	The Godfather Part II	10
17.	Fight Club	10
18.	The Lord of the Rings: The Two Towers	10
19.	The Shawshank Redemption	10
20.	Inception	10



Além disso, é possível notar que os filmes mais bem ranqueados tem como distribuição de cargos através da seguinte query:

```
SELECT
    t2.category,
    COUNT(*) AS role_count
FROM
    `pelagic-range-398901.imdbmvp.TOPCARGOS` t2
GROUP BY
    t2.category
ORDER BY
    role_count DESC;
```

Resultados da consulta			
INFORMAÇÕES DO JOB		RESULTADOS	JSON
Linha	category	role_count	
1	actor	20	
2	writer	17	
3	director	16	
4	producer	15	
5	actress	13	
6	cinematographer	8	
7	editor	7	
8	composer	7	
9	production_designer	2	

E pode ser calculada a média das classificações (averagerating) para cada categoria de cargo. Isso pode ajudar a determinar se há uma correlação entre o cargo desempenhado no filme e a classificação média no IMDB:

```
WITH CategoryAverageRatings AS (
    SELECT
        t2.category,
        AVG(t1.averagerating) AS average_rating
    FROM
        `pelagic-range-398901.imdbmvp.titleprincipals` t2
    LEFT JOIN
        `pelagic-range-398901.imdbmvp.top20filmes` t1
    ON
        t1.tconst = t2.tconst
    GROUP BY
        t2.category
)
SELECT
    category,
    ROUND(average_rating, 2) AS rounded_average_rating
FROM
    CategoryAverageRatings
ORDER BY
    average_rating DESC;
```

Resultados da consulta

INFORMAÇÕES DO JOB		RESULTADOS	JSON
Linha	category ▼	rounded_average_rat	
1	composer	8.96	
2	actor	8.95	
3	actress	8.94	
4	editor	8.92	
5	writer	8.91	
6	cinematographer	8.91	
7	director	8.88	
8	producer	8.87	
9	production_designer	8.85	
10	self	null	
11	archive_footage	null	

Quanto a pergunta final indagada na parte 1. Objetivo: quais são os atores que mais aparecem dentre esses filmes:

Para isso, se tornou necessário criar uma nova tabela TOPATORES adicionando a chave de nome com filtro somente para atores e cruza-la com a base namebasics para descobrir seus nomes.

QUERY 1

```
CREATE OR REPLACE TABLE `pelagic-range-398901.imdbmvp.topatores`
AS SELECT t1.primarytitle, t1.averagerating, t2.category, t2.tconst, t2.nconst
FROM `pelagic-range-398901.imdbmvp.top20filmes` t1
LEFT JOIN `pelagic-range-398901.imdbmvp.titleprincipals` t2
ON t1.tconst = t2.tconst
where category = 'actor';
```

QUERY 2

```
CREATE OR REPLACE TABLE `pelagic-range-398901.imdbmvp.topatores`
AS SELECT t1.primarytitle, t1.averagerating, t1.category, t1.tconst, t1.nconst, t2.primaryName
FROM `pelagic-range-398901.imdbmvp.topatores` t1
LEFT JOIN `pelagic-range-398901.imdbmvp.namebasics` t2
ON t1.nconst = t2.nconst;
```

ESQUEMA	DETALHES	VISUALIZAR	LINHAGEM	PERFIL DE DADOS	QUALIDADE DOS DADOS	
Linha	primarytitle	averagerating	category	tconst	nconst	primaryName
1	The Dark Knight	9.0	actor	tt0468569	nm0000288	Christian Bale
2	The Dark Knight	9.0	actor	tt0468569	nm0000323	Michael Caine
3	The Dark Knight	9.0	actor	tt0468569	nm0001173	Aaron Eckhart
4	The Dark Knight	9.0	actor	tt0468569	nm0005132	Heath Ledger
5	Schindler's List	9.0	actor	tt0108052	nm0000146	Ralph Fiennes
6	Schindler's List	9.0	actor	tt0108052	nm0000553	Liam Neeson
7	Schindler's List	9.0	actor	tt0108052	nm0001426	Ben Kingsley
8	The Godfather Part II	9.0	actor	tt0071562	nm0000199	Al Pacino
9	The Godfather Part II	9.0	actor	tt0071562	nm0000134	Robert De Niro
10	The Godfather Part II	9.0	actor	tt0071562	nm0000380	Robert Duvall
11	The Lord of the Rings: The Retu...	9.0	actor	tt0167260	nm0001557	Viggo Mortensen
12	The Lord of the Rings: The Retu...	9.0	actor	tt0167260	nm0089217	Orlando Bloom
13	The Lord of the Rings: The Retu...	9.0	actor	tt0167260	nm0000704	Elijah Wood
14	The Lord of the Rings: The Retu...	9.0	actor	tt0167260	nm0005212	Ian McKellen
15	Breaking Bad	9.5	actor	tt0903747	nm2666409	RJ Mitte
16	Breaking Bad	9.5	actor	tt0903747	nm0606487	Dean Norris
17	Breaking Bad	9.5	actor	tt0903747	nm0644022	Bob Odenkirk
18	Breaking Bad	9.5	actor	tt0903747	nm0666739	Aaron Paul
19	Breaking Bad	9.5	actor	tt0903747	nm0052186	Jonathan Banks
20	Breaking Bad	9.5	actor	tt0903747	nm0186505	Bryan Cranston
21	Breaking Bad	9.5	actor	tt0903747	nm2366374	Steven Michael Quezada
22	Goodfellas	8.7	actor	tt0099685	nm0000582	Joe Pesci
23	Goodfellas	8.7	actor	tt0099685	nm0000501	Ray Liotta
24	Goodfellas	8.7	actor	tt0099685	nm0000134	Robert De Niro
25	The Matrix	8.7	actor	tt0133093	nm0915989	Hugo Weaving
26	The Matrix	8.7	actor	tt0133093	nm0000401	Laurence Fishburne
27	The Matrix	8.7	actor	tt0133093	nm0000206	Keanu Reeves
28	Stranger Things	8.7	actor	tt4574334	nm1082086	David Harbour

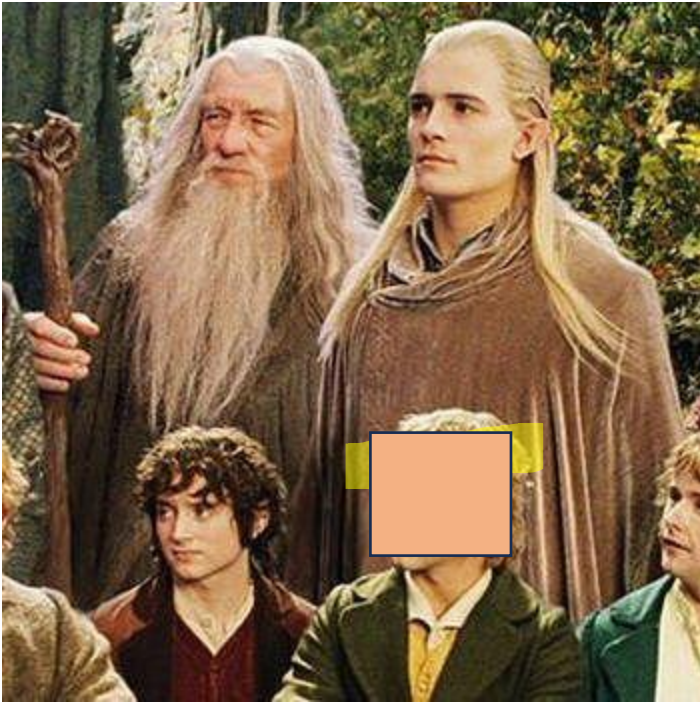
Agora, para finalizar, uma query para encontrar os 6 atores que estejam em mais de uma obra dentre as maiores 20 ranqueadas.

```
SELECT
    primaryname,
    COUNT(primaryname) AS name_count,
    AVG(averagerating) AS average_rating
FROM
    `pelagic-range-398901.imdbmvp.topatores`
GROUP BY
    primaryname
ORDER BY
    name_count DESC,
    average_rating DESC
limit 6;
```

Resultados da consulta

INFORMAÇÕES DO JOB	RESULTADOS	JSON	DETALHES DA EXI
Linha	primaryname	name_count	average_rating
1	Ian McKellen	3	8.866666666666...
2	Orlando Bloom	3	8.866666666666...
3	Elijah Wood	3	8.866666666666...
4	Al Pacino	2	9.1
5	Viggo Mortensen	2	8.9
6	Robert De Niro	2	8.85

O que nos dá a resposta de trindade dos atores e personagens mais bem ranqueados: Orlando Bloom(Legolas), Ian McKellen(Gandalf) e Elijah Wood(Frodo):



Como os três participam de uma das maiores trilogias dos cinemas, grande crédito também pode ser dado as lendas Al Pacino e Robert Deniro.



Parte 5. Autoavaliação

Ao me autoavaliar, noto resiliência e esforço para completar e entender todos os pontos dos materiais oferecidos, tanto quanto entregar de maneira completa seus trabalhos e o MVP.

Ao passar do tempo e entender a melhor maneira para fazer a entrega, entendendo e melhorando o trabalho de acordo com os pedidos e necessidades declaradas pelos professores, como a necessidade de todo o trabalho ser feito através de uma plataforma na nuvem, me vi também me esforçando para acertar todo o MVP de acordo com a demanda e as dificuldades encontradas no caminho.

Espero ter alcançado o objetivo proposto. Obrigado.