**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 logica 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 contêm informações detalhadas sobre títulos, equipes de produção, colaboradores e pessoas envolvidas. Isso permite que a analise 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:

Texto

Descrição gerada automaticamente com confiança baixa

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:

Uma imagem contendo Texto

Descrição gerada automaticamente

**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 possivel a criação do cluster que armazenará os arquivos apontados através do esquema estrela, sendo eles:

[name.basics.tsv.gz](https://datasets.imdbws.com/name.basics.tsv.gz)

[title.akas.tsv.gz](https://datasets.imdbws.com/title.akas.tsv.gz)

[title.basics.tsv.gz](https://datasets.imdbws.com/title.basics.tsv.gz)

[title.crew.tsv.gz](https://datasets.imdbws.com/title.crew.tsv.gz)

[title.episode.tsv.gz](https://datasets.imdbws.com/title.episode.tsv.gz)

[title.principals.tsv.gz](https://datasets.imdbws.com/title.principals.tsv.gz)

[title.ratings.tsv.gz](https://datasets.imdbws.com/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.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

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](https://datasets.imdbws.com/title.akas.tsv.gz), [title.episode.tsv.gz](https://datasets.imdbws.com/title.episode.tsv.gz) e title.crew

Diagrama

Descrição gerada automaticamente

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

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

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:

Uma imagem contendo Tabela

Descrição gerada automaticamente

3.3 Qualidade dos dados:

3.3.1 Análise sobre qualidade dos dados

Tabela

Descrição gerada automaticamente

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 dificil 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 facil acesso:

Tabela

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

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.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

**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:

Tabela

Descrição gerada automaticamente

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;

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

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:



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.

Gráfico

Descrição gerada automaticamente

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;

Tabela

Descrição gerada automaticamente

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;

Tabela

Descrição gerada automaticamente

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;

Tabela

Descrição gerada automaticamente

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;

Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente

Oque 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.