UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Departamento de Computação

Banco de Dados para Ciência de Dados

Projeto Físico e Consultas - Cassandra - IMDB

Profa. Marilde Terezinha Prado Santos

Integrantes do Grupo:

Caio Ueda Sampaio, 802215, BCC
Gabrielly Castilho Guimarães, 805757, BCC
Lucas Maciel Balieiro, 800534, BCC
Vinícius de Oliveira Guimarães, 802431, BCC

08 de Fevereiro de 2023

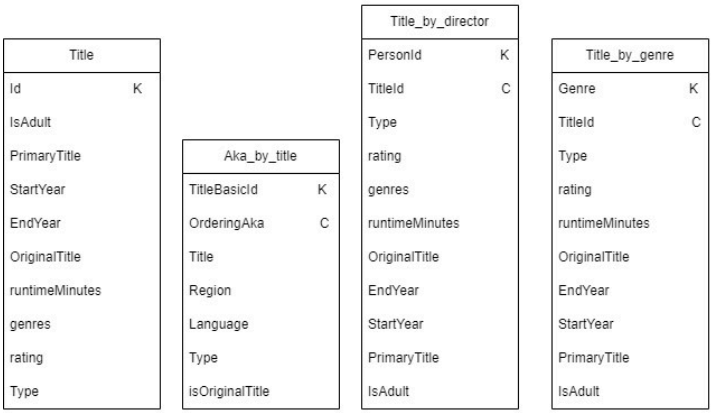
INTRODUÇÃO

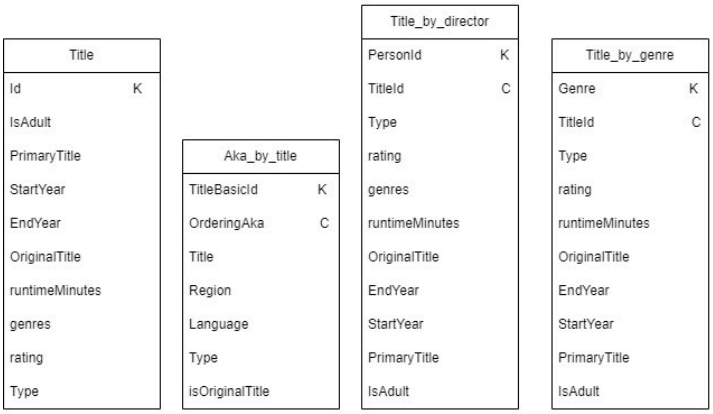
O projeto físico do nosso grupo e as consultas foram feitos com base no tema escolhido “IMDb Dataset - From 1888 to 2023” que contém as informações de filmes, séries, desenhos, participantes, diretores, escritores e entre outros, desde 1888 até 2023.

Link para o Repositório da parte do Cassandra, com todos os Scripts python criados para manipulação dos CSV’s necessários:
https://github.com/Viniciusog/bdcd-imdb/tree/main/EtapaCassandra

PROJETO FÍSICO

O nosso projeto físico foi criado para satisfazer as seguintes tabelas:





CÓDIGO PARA CRIAÇÃO DAS TABELAS QUE SERÃO UTILIZADAS PARA AS BUSCAS

```
CREATE KEYSPACE meu_keyspace
WITH replication = {'class': 'SimpleStrategy', 'replication_factor' :
1};

USE meu_keyspace;

// TEM QUE USAR O ARQUIVO DO TITLE BASICS COM RATING (Isso foi feito
usando o arquivo CreateCsvTitleWithRating)

// ------ CRIAÇÃO DAS TABELAS ------
//
tconst,titleType,primaryTitle,originalTitle,isAdult,startYear,endYear,r
untimeMinutes,genres,averageRating

CREATE TABLE title (
    tconst text PRIMARY KEY,
    titleType text,
    primaryTitle text,
    originalTitle text,
    isAdult boolean,
    startYear int,
    endYear text,
    runtimeMinutes text,
    genres text,
    averageRating float
);

// ------ IMPORTAÇÃO DOS DADOS ------

COPY title
(tconst,titleType,primaryTitle,originalTitle,isAdult,startYear,endYear,
runtimeMinutes,genres,averageRating) FROM
'/dados/ImdbTitleBasicsWithRating.csv' WITH DELIMITER=',' AND
HEADER=TRUE;
```

```
// ------ RELAÇÃO TITLE COM AKA ------

CREATE TABLE AkaByTitle (
    titleId text,
    ordering text,
    title text,
    region text,
    language text,
    types text,
    attributes text,
    isOriginalTitle boolean,
    PRIMARY KEY (titleId, ordering)
);

COPY AkaByTitle (titleId, ordering, title, region, language, types,
attributes, isOriginalTitle) FROM
'/dados/ImdbTitleBasicsAkaRelation.csv' WITH DELIMITER=',' AND
HEADER=TRUE;

// ------ RELAÇÃO TITLE COM GENRE ------
CREATE TABLE TitleByGenre (
   tconst text,
   titleType text,
   primaryTitle text,
   originalTitle text,
   isAdult boolean,
   startYear int,
   endYear text,
   runtimeMinutes int,
   genres text,
   averageRating float,
   genre text,
   PRIMARY KEY ((tconst, genre))
);

COPY TitleByGenre (tconst,titleType, primaryTitle, originalTitle,
isAdult, startYear, endYear, runtimeMinutes, genres, averageRating,
genre) FROM '/dados/ImdbTitleBasicsGenreRelation.csv' WITH
DELIMITER=',' AND HEADER=TRUE;

// ------ CRIAÇÃO EPISODES BY TITLE ------
```

```
CREATE TABLE episodes_by_title (
    tconst text,
    parentTconst text,
    seasonNumber text,
    episodeNumber text,
    PRIMARY KEY (tconst, parentTconst)
);

COPY Episodes_by_title (tconst, parentTconst, seasonNumber,
episodeNumber)
FROM '/tmp/ImdbTitleEpisode10000.csv' WITH DELIMITER=',' AND
HEADER=TRUE;

// ------ CRIAÇÃO TITLE BY DIRECTOR ------
CREATE TABLE title_by_director (
    tconst TEXT,
    titleType TEXT,
    primaryTitle TEXT,
    originalTitle TEXT,
    isAdult BOOLEAN,
    startYear TEXT,
    endYear TEXT,
    runtimeMinutes TEXT,
    genres TEXT,
    averageRating FLOAT,
    directors TEXT,
    director TEXT,
    PRIMARY KEY(director, tconst)
);


COPY title_by_director (tconst, titleType, primaryTitle, originalTitle,
isAdult, startYear, endYear, runtimeMinutes, genres, averageRating,
directors, director)
FROM '/dados/ImdbTitleBasicsDirectorRelation.csv'
WITH DELIMITER=',' AND HEADER=TRUE;

// ---- CRIAÇÃO PRINCIPALS PERSON ----
tconst,nconst,category,characters,primaryName,birthYear,deathYear,prima
ryProfession
```

```
CREATE TABLE person_by_title (
    tconst text,
    nconst text,
    category text,
    characters text,
    primaryName text,
    birthYear text,
    deathYear text,
    primaryProfession text,
    PRIMARY KEY (tconst, nconst)
);

COPY person_by_title (tconst, nconst, category, characters,
primaryName, birthYear, deathYear, primaryProfession) FROM
'/dados/ImdbTitlePersonByTitleRelation.csv' WITH DELIMITER=',' AND
HEADER=TRUE;

// ---- INSERINDO 10K DE PESSOAS -----
nconst,primaryName,birthYear,deathYear,primaryProfession,knownForTitles

CREATE TABLE person (
    nconst text,
    primaryName text,
    birthYear text,
    deathYear text,
    primaryProfession text,
    knownForTitles text,
    PRIMARY KEY (nconst)
);

COPY person (nconst, primaryName, birthYear, deathYear,
primaryProfession, knownForTitles) FROM '/dados/ImdbName10k.csv' WITH
DELIMITER=',' AND HEADER=TRUE;

// ---- TITLE BY START YEAR -----
tconst,titleType,primaryTitle,originalTitle,isAdult,startYear,endYear,r
untimeMinutes,genres,averageRating

CREATE TABLE title_by_start_year (
    tconst text,
```

```
    titleType text,
    primaryTitle text,
    originalTitle text,
    isAdult boolean,
    startYear text,
    endYear text,
    runtimeMinutes text,
    genres text,
    averageRating float,
    PRIMARY KEY (startYear, tconst)
);

COPY title_by_start_year
(tconst,titleType,primaryTitle,originalTitle,isAdult,startYear,endYear,
runtimeMinutes,genres,averageRating) FROM
'/dados/ImdbTitleBasicsWithRating.csv' WITH DELIMITER=',' AND
HEADER=TRUE;

// ---- TITLE EPISODES -----
tconst,parentTconst,seasonNumber,episodeNumber
CREATE TABLE episodes (
    tconst text,
    parentTconst text,
    seasonNumber text,
    episodeNumber text,
    PRIMARY KEY(parentTconst, tconst)
);

COPY episodes (tconst,parentTconst,seasonNumber,episodeNumber) FROM
'/dados/ImdbTitleEpisode10k.csv' WITH DELIMITER=',' AND HEADER=TRUE;
```

Verificando se as tabelas foram criadas com sucesso:

```
cqlsh:meu_keyspace> describe tables;

akabytitle  person           title              title_by_start_year
episodes    person_by_title  title_by_director  titlebygenre
```

CÓDIGO PARA AS CONSULTAS:

```
// PESSOAS
SELECT * FROM meu_keyspace.person;

SELECT * from meu_keyspace.person where nconst = 'nm0007191';

// TITLE BY GENRE
SELECT * FROM meu_keyspace.titleByGenre;

SELECT * FROM meu_keyspace.titleByGenre WHERE genre = 'Drama' ALLOW
FILTERING;

SELECT * FROM meu_keyspace.titleByGenre WHERE genre = 'Western' ALLOW
FILTERING;

SELECT * FROM meu_keyspace.titleByGenre WHERE titletype = 'short' ALLOW
FILTERING;

// TITLE BY START YEAR
SELECT * FROM title_by_start_year;

SELECT * FROM title_by_start_year WHERE startYear = '1918';

SELECT * FROM title_by_start_year WHERE startYear = '1918' AND
titleType = 'movie' ALLOW FILTERING;

// PERSON BY TITLE
SELECT * FROM person_by_title;

SELECT * FROM person_by_title WHERE tconst = 'tt0066986'

SELECT * FROM person_by_title WHERE category = 'writer' ALLOW
FILTERING;

SELECT * FROM person_by_title WHERE deathYear = '2020' ALLOW FILTERING;

// EPISODES BY TITLE
SELECT * FROM episodes;
```

```
SELECT * FROM episodes WHERE parentTconst = 'tt2162303';

SELECT * FROM episodes WHERE seasonNumber = '1' ALLOW FILTERING;

// TITLE BY DIRECTOR
SELECT * FROM title_by_director;

SELECT * FROM title_by_director WHERE director = 'nm0227020';

// AKAS BY TITLE
SELECT * FROM akabytitle;

SELECT * FROM akabytitle WHERE titleid = 'tt0004702';

// TITLE
SELECT * FROM title;

SELECT * FROM title WHERE tconst = 'tt0000382';

SELECT * FROM title where originalTitle = 'The Impostor' ALLOW
FILTERING;
```

Exemplo de consulta realizada para filtrar todos os filmes com diretor sendo ‘nm0227020’

```
cqlsh:meu_keyspace> SELECT * FROM title_by_director WHERE director = 'nm0227020';

 director  | tconst    | averagerating | directors | endyear | genres          | isadult | originaltitle          | primarytitle           | runtimeminutes | startyear | titletype
-----------+-----------+---------------+-----------+---------+-----------------+---------+------------------------+------------------------+----------------+-----------+-----------
 nm0227020 | tt0004106 |             0 | nm0227020 |       N |    Comedy,Short |   False |       His Taking Ways |       His Taking Ways |              N |      1914 |     short
 nm0227020 | tt0004188 |             0 | nm0227020 |       N |           Drama |   False |  The Key to Yesterday |  The Key to Yesterday |              N |      1914 |     movie
 nm0227020 | tt0004882 |             0 | nm0227020 |       N |    Comedy,Short |   False |        Almost a Widow |        Almost a Widow |              N |      1915 |     short
 nm0227020 | tt0004896 |             0 | nm0227020 |       N |    Comedy,Short |   False |     Anita's Butterfly |     Anita's Butterfly |              N |      1915 |     short
 nm0227020 | tt0006575 |             0 | nm0227020 |       N |    Comedy,Short |   False |       A Deep Sea Liar |       A Deep Sea Liar |              N |      1916 |     short
 nm0227020 | tt0006576 |             0 | nm0227020 |       N |    Comedy,Short |   False | Delinquent Bridegrooms | Delinquent Bridegrooms |              N |      1916 |     short
 nm0227020 | tt0006695 |             0 | nm0227020 |       N |    Comedy,Short |   False | For Ten Thousand Bucks | For Ten Thousand Bucks |              N |      1916 |     short
 nm0227020 | tt0006799 |             0 | nm0227020 |       N |    Comedy,Short |   False |       Hired and Fired |       Hired and Fired |              N |      1916 |     short
 nm0227020 | tt0006802 |             0 | nm0227020 |       N |    Comedy,Short |   False |          His Blowout |          His Blowout |              N |      1916 |     short
 nm0227020 | tt0006869 |             0 | nm0227020 |       N |    Comedy,Short |   False |         The Iron Mitt |         The Iron Mitt |              N |      1916 |     short
 nm0227020 | tt0007556 |             0 | nm0227020 |       N |    Comedy,Short |   False |        When Papa Died |        When Papa Died |              N |      1916 |     short
 nm0227020 | tt0008136 |             0 | nm0227020 |       N |    Comedy,Drama |   False |    Indiscreet Corinne |    Indiscreet Corinne |              N |      1917 |     movie
 nm0227020 | tt0008865 |           6.3 | nm0227020 |       N | Comedy,Thriller |   False |               Beans |               Beans |             50 |      1918 |     movie
 nm0227020 | tt0008882 |           7.1 | nm0227020 |       N |           Drama |   False |    Betty Takes a Hand |    Betty Takes a Hand |              N |      1918 |     movie
 nm0227020 | tt0009152 |             0 | nm0227020 |       N |    Comedy,Drama |   False |     Heiress for a Day |     Heiress for a Day |              N |      1918 |     movie
 nm0227020 | tt0009303 |             0 | nm0227020 |       N |          Comedy |   False |       Limousine Life |       Limousine Life |              N |      1918 |     movie
 nm0227020 | tt0009320 |             0 | nm0227020 |       N |    Comedy,Drama |   False |      The Love Swindle |      The Love Swindle |             50 |      1918 |     movie
 nm0227020 | tt0009418 |             0 | nm0227020 |       N |    Comedy,Drama |   False |      Nancy Comes Home |      Nancy Comes Home |             50 |      1918 |     movie
 nm0227020 | tt0009602 |             0 | nm0227020 |       N |    Comedy,Drama |   False |   She Hired a Husband |   She Hired a Husband |             50 |      1918 |     movie
 nm0227020 | tt0009975 |             6 | nm0227020 |       N |          Comedy |   False |      Burglar by Proxy |      Burglar by Proxy |              N |      1919 |     movie
 nm0227020 | tt0010133 |           4.2 | nm0227020 |       N |          Comedy |   False |      The Follies Girl |      The Follies Girl |             50 |      1919 |     movie

(21 rows)
```

Exemplo de consulta para pegar AKA com titleId = 'tt0004702';

```
cqlsh:meu_keyspace> SELECT * FROM akabytitle WHERE titleid = 'tt0004702';

 titleid   | ordering | attributes | isoriginaltitle | language | region | title                | types
-----------+----------+------------+-----------------+----------+--------+----------------------+-------
 tt0004702 |        1 |          N |           False |        N |     US | Through the Centuries |     N
```