Data Engineering 1
Term Project 1 Documentation
Balint Thaler

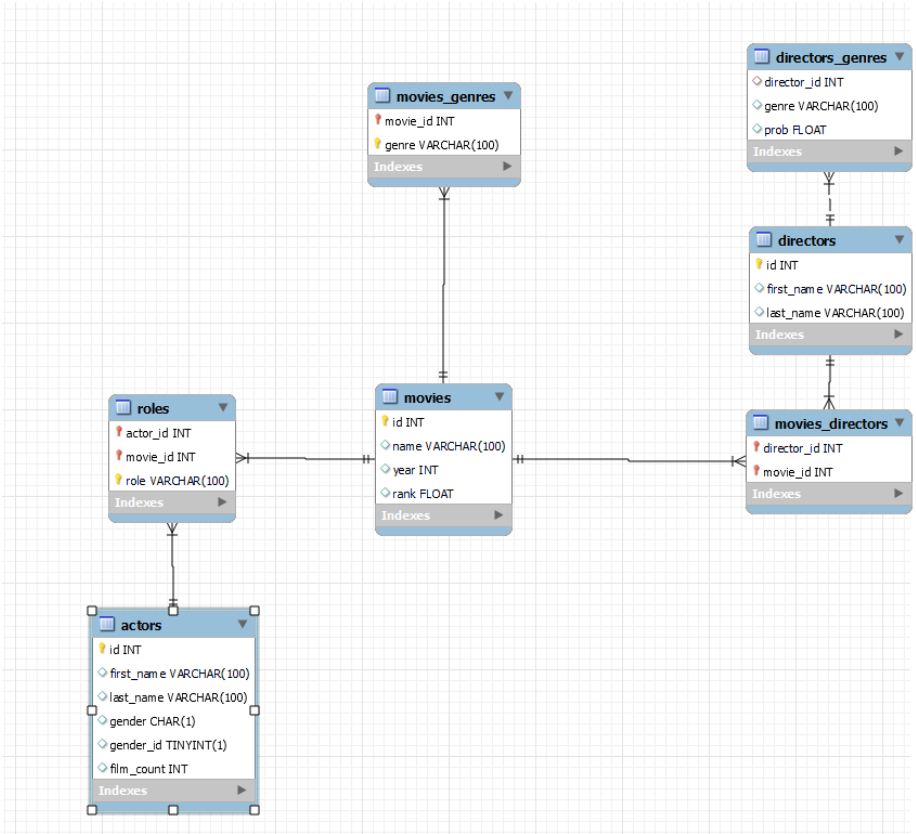
Documentation – imdb_small dataset

The data

The original dataset contains data about certain movies in the imdb database. It contains 7 data tables which are the following:

- *movies*: describing attributes of the movies themselves such as movie id, year, ranking, title
- *roles*: describing roles in movies
- *actors*: describing the actors who played certain roles
- *movies_genres*: it categorises each movie into genres
- *directors*: contains data on directors
- *movies_directors*: connects the director table with the movies table
- *directors_genres*: containing data on directors and the genres

The detailed EER diagram is shown below and is also part of the model that is uploaded on the GitHub repository.



The dataset contains data about 37 movies, 1907 actors, 1989 roles, genre data of all 37 movies, director data about 36 movies. The dataset was downloaded from the following website: https://relational.fit.cvut.cz/dataset/IMDb which is associated with the Czech Faculty of Information Technology at the Czech Technical University in Prague.

Analytics plan

The analysis of the dataset aims to identify potential differences among the different genres of movies. It aims to answer questions like: “Are certain genres ranking higher than others in the imdb ranking?”, “Do certain genres employ more or less male actors versus others?”, “Are genres with more or less male actors perceived as more successful by their ranking?”, “Are certain genres of movies that are older or younger employing more or less women or are they more or less successful?”.

Data wrangling

To answer the questions set out in the analytical plan, first the actors data table needs to be changed slightly. This table contains gender data on each actor as a string of M for males and F for females. Therefore, a new column is added in the table that takes the value of 1 for males and 0 for females.

The analytical layer

The analytical layer is created in a separate table called *genre_data.* The table is first created using 4 columns:

- genre: values for each of the 16 genres present in the dataset
- average_male_ratio: a grouped average of the ratio of male actors in the movies of each genre
- average_rank: a grouped average of the ranking of the movies of each genre
- average_movie_age: a grouped average of the age of the movies of each genre calculated by the year difference between the current date and the premier of the movie

Once the columns are created, the genre column is pre-filled with the unique values from the *movies_genre* table.
A stored procedure called *creategenredata()* stores the script for filling the *genre_data* table with the necessary data from the other tables using joins and some calculations around the average variables. Once the stored procedure is called, the data in the *genre_data* table is filed with the average values.

Data marts

Using internet search I have identified that the US movie scene is dominated by 5 major studios who are creating the vast majority of movies in Hollywood. These are Disney, WarnerBros, Universal, Paramount and Sony. Each studio is somewhat different from the other and create movies mostly in certain genres. Using ChatGPT[1], I have identified the genres that each of the 5 studios mostly operate in.

[1] Link to ChatGPT conversation: https://chat.openai.com/share/f0daffbc-e86a-463a-8aa5-e9dadb2fe692

The data marts in the model are reflecting these studio preferences as well: 5 views were created for each studio that filters for the values of the *genre_data* table according to the typical genres used by them.

Triggers

A trigger was created that calls the stored procedure called *creategenredata()* and therefore updates the values of the table accordingly once a new line is inserted into the *movies_genres* table. Given that the analytical layer uses movie-related data from the *movies* table (and that it is not possible to insert a new unique movie id into the *movies_genres* table before inserting the same id into the *movies* table), 2 new insert values are required to fire the trigger.

The script contains an example value – with obviously unrealistic numbers so that their effect is easily visible on the average values of the data marts or the *genre_data* table – for testing the trigger.

The trigger is suitable for a dataset of this size, however for a larger dataset the performance can be optimized by only refreshing the values of the *genre_data* table that have changed as a result of the new value insertion.

Summary

The analytical questions set out in the plan earlier can be answered by looking at the data in the analytical layer: we see that certain genres are more successful than others, some genres do indeed employ more men than others and some genres are significantly “younger” than others.