

Information Integration Project Proposal

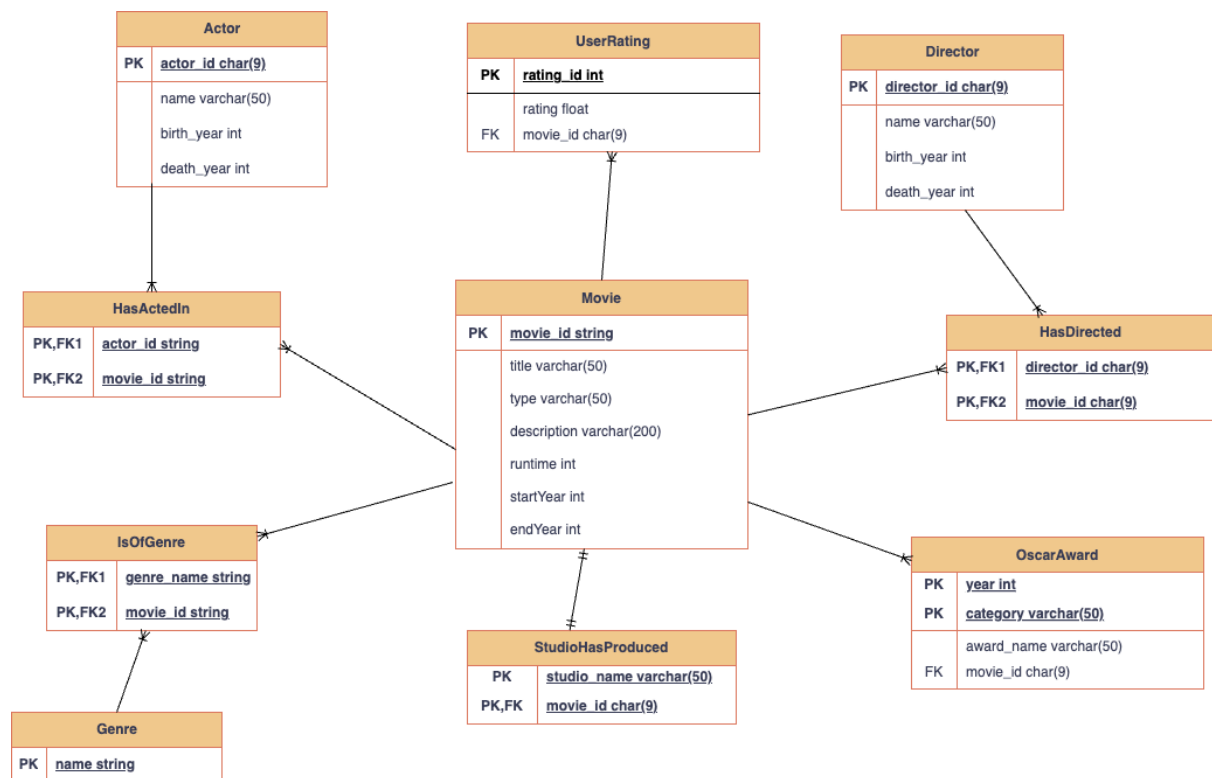
Giovanni Buracci 2097143

Project goal

The scope of the project is to integrate heterogeneous data regarding movies, actors, directors, ratings made by users and Oscar awards.

The retrieved global schema should be able to satisfy queries that span across the sources and that are able to provide information not directly collectible by querying the single sources.

Initial ER diagram



Please note that optionality in associations cardinality is not expressed

Data Sources



<https://developer.imdb.com/non-commercial-datasets/>

One of the largest sources of information regarding movies, series, cast and crew. It exposes data in the form of TSV files downloadable from the link above. In this project, it would provide information about movies, actors and directors.



<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>

For user ratings, data would be taken from the dataset that powers MoviLens, a recommendation engine for movies developed by Grouplens, a research lab of the University of Minnesota.

Data is downloaded in CSV format from Kaggle.



<https://www.kaggle.com/datasets/gsimonx37/letterboxd>

Letterboxd is a social network for film enthusiasts that also contains metadata about movies.

The creator of the dataset above, published on Kaggle as CSVs, scraped it obtaining information regarding the films listed.

It is useful because it contains organized data about production studios, and I would only use this portion of the dataset.



<https://www.kaggle.com/datasets/unanimad/the-oscar-award>

This dataset in the form of a CSV is the result of a scraping of The Academy Awards website, which exposes information about nominees and winners of the commonly known Oscars, from their foundation until these days.

Initial Queries

Below is an initial list of possible CQs that can be satisfied on the retrieved global schema:

- 1) Birth year of "Best Picture" Oscar-winning movies still alive
- 2) Name of genres associated with movies produced by "Paramount Pictures"
- 3) Surname of actors in movies directed by Christopher Nolan
- 4) Description of oscar-winning movies with a rating less than 2.5
- 5) Title of horror movies with actors born after the year 2000