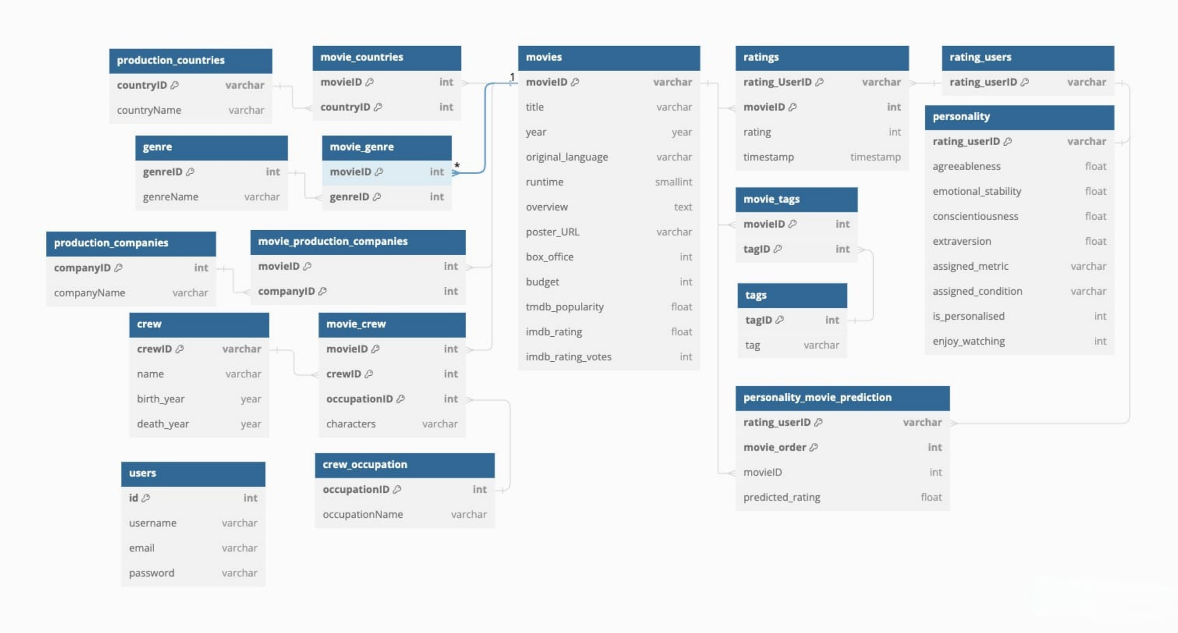
Database Architecture

The movie dataset referenced in this database contains a large number and variety of song metadata, which are mainly arranged by movie name and user name, and contain large number of duplicate data. This poses a great challenge for the performance of data screening as well as data storage. To address this issue, we designed a database schema for the original movie data, aimed to minimise redundancy and dependency by organise movie data into well-structured tables. The figure below expresses our database schema in the form of Entity-Relationship Diagram(ERD):



Entity Relationship Diagram of the Movie Database

Here, different states of normalisation for different data types are explained:

First Normal Form (1NF)

First normal form is applied for all tables contain atomic values, with no repeating groups.

- `movies`, `production_countries`, `genre`, `production_companies`, `crew`, `crew_occupation`, `rating_users`, and `tags` tables all have atomic attributes.
- Composite primary keys are utilised where necessary, such as in junction tables like `movie_crew`, `movie_genre`, `movie_countries`, `movie_production_companies`, `ratings`, `movie_tags`, and `personality_movie_prediction`.

Second Normal Form (2NF)

In second normal form, a table is in 1NF and does not have partial dependencies on the primary key. In this schema:

- No non-prime attributes are dependent on only a portion of a composite primary key.

- For instance, attributes like `title`, `release_year`, `original_language`, `runtime`, `overview`, `poster_URL`, `box_office`, `budget`, `tmdb_popularity`, `imdb_rating`, `imdb_rating_votes` in the `movies` table depend solely on the primary key `movieID`, not on any subset of it.

Third Normal Form (3NF)

Our database schema also implies third normal form, a table is in 2NF and does not have transitive dependencies. In this schema:

- Attributes are not transitively dependent on the primary key.
- For example, `genreName` in the `genre` table depends directly on the primary key `genreID`, not on any non-prime attributes.

Discussion

The provided database schema is well-normalized up to 3NF. It effectively eliminates redundancy and ensures data integrity by structuring data into distinct tables and establishing appropriate relationships between them. Key points contributing to normalization include the use of composite primary keys in junction tables, proper identification of primary keys, and avoidance of partial and transitive dependencies.

Conclusion

Normalisation is essential for efficient database management, pivotal for maintaining data consistency, minimising redundancy, and streamlining maintenance procedures. This database schema showcases a commendable adherence to normalisation principles, upholding a robust structure compliant with Third Normal Form (3NF). This underscores its capability to efficiently manage and manipulate data. Consequently, the schema lays a solid foundation for conducting further analyses on diverse types of movie data, including evaluation in correlation across them.