Data Catalogue

1. User Data

- **Description**: Synthetic user demographic data generated for 610 users with attributes nationality, gender, and age.
- **Source**: Generated using Raw_data_generator.ipynb.
- **Input Data**: Number of users in the Movielens dataset (610).
- **Transformation**: Randomized attribute generation structured into a CSV format.
- **Output**:
    - **File Name**: users.csv
    - **Location**: ../movie_data/
    - **Schema**:
        - user_id: Integer, Unique identifier for each user.
        - nationality: String, Nationality of the user.
        - gender: String, Gender of the user.
        - age: Integer, Age of the user.

2. Rating Data

- **Description**: Reformatted rating data from the Movielens dataset into JSON format, linked to external IMDB and TMDB datasets.
- **Source**:
    - Movielens dataset (ratings.csv, movies.csv, links.csv).
    - https://grouplens.org/datasets/movielens/latest/
- **Input Data**:
    - ratings.csv: User ratings of movies.
    - movies.csv: Movie metadata.
    - links.csv: Links between Movielens, IMDB, and TMDB IDs.

- **Transformation**:
    - Structured data converted to JSON format.
    - Additional complexity introduced via external dataset linking.
- **Output**:
    - **File Name**: raw_unstructured_data.json
    - **Location**: ../movie_data/
    - **Schema**:
        - user_id: Integer, Identifier for the user.
        - movie_id: Integer, Identifier for the movie.
        - rating: Float, User's rating of the movie.
        - imdb_id: String, IMDB identifier for the movie.
        - tmdb_id: String, TMDB identifier for the movie.

3. TMDB Data

- **Description**: Metadata about movies and credits extracted from TMDB datasets.
- **Source**:
    - TMDB datasets (tmdb_5000_movies.csv, tmdb_5000_credits.csv).
    - https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data?select=tmdb_5000_movies.csv
- **Input Data**:
    - tmdb_5000_movies.csv: Metadata about movies (genres, budget, release dates, etc.).
    - tmdb_5000_credits.csv: Cast and crew details.
- **Transformation**:
    - Data loaded into Pandas DataFrames for further preprocessing.
- **Output**:
    - **Files**:
        - tmdb.csv
        - cast.csv

        - crew.csv
    - **Location**: ../movie_data/
    - **Schemas**:
        - **TMDB Data (tmdb.csv)**:
            - tmdb_id: Integer, Unique movie identifier.
            - title: String, Title of the movie.
            - budget: Float, Budget of the movie.
            - genres: List, Genres associated with the movie.
            - release_date: Date, Release date of the movie.
            - Derived columns:
                - age_restriction: String, Indicates movie's target age group.
                - high_budget: Boolean, Flags movies with budgets exceeding $1,000,000.
        - **Cast Data (cast.csv)**:
            - actor_name: String, Name of the actor.
            - character: String, Character played by the actor.
            - gender: String, Gender of the actor.
        - **Crew Data (crew.csv)**:
            - director_name: String, Name of the director.
            - department: String, Department of the crew member.
            - role: String, Specific role within the movie production.

4. Processed Data

- **Description**: Data preprocessed and structured for database loading.
- **Source**: Processed from raw TMDB, Movielens, and synthetic user data.
- **Transformations**:
    - Merged ratings data with TMDB IDs.
    - Cleaned and transformed attributes (e.g., genres, keywords, cast, and crew).
    - Generated holiday and seasonal indicators based on release dates.

- **Output**:
    - **Files**:
        - ratings.csv
        - tmdb.csv
        - cast.csv
        - crew.csv
        - holidays.csv
    - **Location**: ../movie_data/
    - **Schemas**:
        - **Ratings Data (ratings.csv)**:
            - user_id: Integer, User identifier.
            - movie_id: Integer, Movie identifier.
            - rating: Float, Rating given by the user.
        - **Holiday Data (holidays.csv)**:
            - tmdb_id: Integer, TMDB movie identifier.
            - release_date: Date, Release date of the movie.
            - holiday_flag: Boolean, Indicates if the movie was released near a major holiday.
            - season: String, Season of release (e.g., summer, winter).

5. Database Schema