Jadhav Amarjit
Joshi Payal

1. Domain Description and Data Source: Describe in English the domain you intend to build a database for, along with 20 questions (in English) that someone might want to ask about the domain. (You will be permitted to revise these questions later if needed.) Also describe what source you intend to use for data, and how you intend to ingest the data into your database. You should choose a domain where you can easily get several hundred rows of data.

The domain that we intend to use to build the database is based on Movies released in the year 1995. This dataset is an ensemble of data collected from TMDB and GroupLens. It contains metadata such as budget, revenue, release dates, languages, etc related to various movies. It has information related to ratings given to various movies by users and can be on scale of 1-5. It also contains the movie plot keywords for movies. Movies can be listed under more than one keyword and can belong to more than one genres. Genres range from comic to documentary.

Following questions can be asked about the domain:

Questions :

1. Which movies are not rated by any users?
2. Which movies have different titles originally?
3. Which movies have received ratings from maximum number of users?
4. Which movies are produced by more than one production companies?
5. For which movies no data is available about their production companies?
6. How many movies have original language as english?
7. Which movies have the highest average rating?
8. Which movies do not have imdb and tmdb IDs assigned to them?
9. How many movies appear under various distinct keywords?
10. Which movies belong to more than 2 genres and have been rated by users?
11. Which users have contributed the most to rating various movies?
12. Which company has produced maximum number of movies?
13. How much revenue did the movie having highest budget earned?
14. What is the budget and revenue of the least popular movie?
15. List the various distinct genres for which ‘Warner Bros’ produced movies?
16. What is the second most common language used for movies?
17. List the most popular movie per quarter.
18. What is the average runtime for movies of documentary genre?
19. List the details of the movie having lowest timestamp.
20. Give the user_id and his rating for the lowest budget movie.

Source of Data:

The source of movies data set is:
https://www.kaggle.com/rounakbanik/the-movies-dataset/data
The data we will be using for the project is downloaded in the form of CSV files.

List of files being used is as follows:
- keywords.csv
- links.csv
- movies_metadata.csv
- ratings.csv

Some of the fields in the movies_metadata such as genres, production companies, keywords, etc in movies_metadata.csv and keywords.csv files contain data in JSON object format.
We intend to convert such data into CSV format using below online tool:
https://json-csv.com

Further, we will be dividing keywords.csv data into 2 tables - keywords which will have metadata related to keywords and another one will represent many to many relation between movies and keywords associated with them.

In similar manner, we will be constructing a separate table to store production companies data and another table will have data regarding which movie is produced by which company/ companies.

Also, movies_metadata contain metadata of about 45,000 movies. However, in order to reduce the dataset size, we will be using metadata of only those movies which were released in the year 1995.

Ingesting the data into database:

We plan to import the data into database by uploading data from CSV file using ‘Copy’ command of postgresql.
Tentative steps would be:
1. Create tables in the Postgresql
2. Run the Copy command, syntax for which is as follows:
COPY table_name [(column_name [, ...])]
FROM { 'filename' | STDIN }
[[WITH] (option [, ...])]

Where,

table_name:
The name (optionally schema-qualified) of an existing table

column_name:
An optional list of columns to be copied. If no column list is specified, all columns of the table will be copied.

filename:
The absolute path name of the input or output file. Windows users might need to use an E'' string and double any backslashes used in the path name.

STDIN:
Specifies that input comes from the client application

option:
FORMAT format_name:
Selects the data format to be read or written: text, csv (Comma Separated Values), or binary.