GROUP -23

DBMS
ASSIGNMENT-2
REPORT

We have converted our Entity-Relationship diagram made in the previous assignment to a relational model using the IMDb dataset given to us (in the assignment) and crawling the unavailable data using APIs.

STEP-1: CRAWLING THE MISSING DATA:

1. We first downloaded the available data as TSV files from the IMDb dataset link (given in the assignment pdf).
2. Using the OMDb API link (given in the assignment pdf) we crawled and obtained the necessary data corresponding to movie/tv series plots, Awards and nominations for the movie/tv-series, and the production companies.
3. We have used python scripts (attached in the submission) to automate the crawling process and get the necessary data.
4. We have written two python scripts -- one for missing data for movies/short and the other one for the missing data for tv-series
5. We were not able to obtain a large amount of data corresponding to all of the movies and TV-series in the data set due to the restriction on the number of API calls per day and the premium account being paid.
6. The crawled data was finally stored as '.tsv' files.

We now have the missing data as well and the next step is preprocessing this available data.

STEP-2: PRE-PROCESSING THE DATA SOURCE:

1. We created temporary tables in the data to store the data from the tsv files.
2. Using

```
COPY “table_name” FROM “path to file” DELIMITER “delimiter_character”;
```

command we copied the data from the TSV files into the corresponding tables created using SQL scripts.

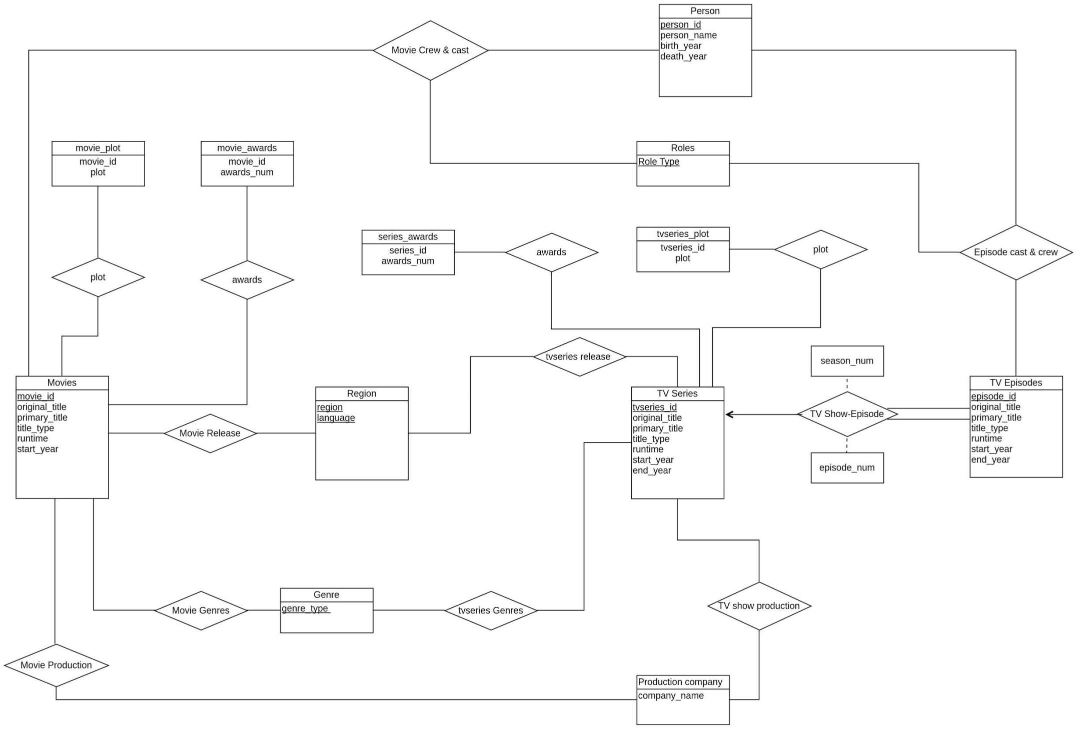
3. We then started processing this source data. The processing included the following:

- Removing the headers from the TSV data as they were of a different data type and unnecessary.
- Casting the data to its corresponding data type
- Separating the columns with multivalued attributes to multiple columns.
- Separating the entries with multivalued attributes into multiple rows.

STEP-3: MAKING OUR DATABASE

1. We first modified our ER diagram based on the final available data after crawling. We deleted the relations for which the data was still unavailable online.

2. The final ER diagram after the changes is as follows (.JPG file is attached along with the submission):-



3. We then used the updated ER diagram and created the tables with the attributes in the ER diagram using the ‘create table’ command in SQL (DDL).
4. All the other tables of the database were created using the sql commands mentioned in the “group23_completeSQL.sql” file attached in the submitted zip folder

By Group-23:

CH18BTECH11008	Ch Vinay kumar
CH18BTECH11009	D Lakshmi Manohar
CH18BTECH11012	Hritik Sarkar
CH18BTECH11015	M Dinesh
CH18BTECH11033	Vrushank K