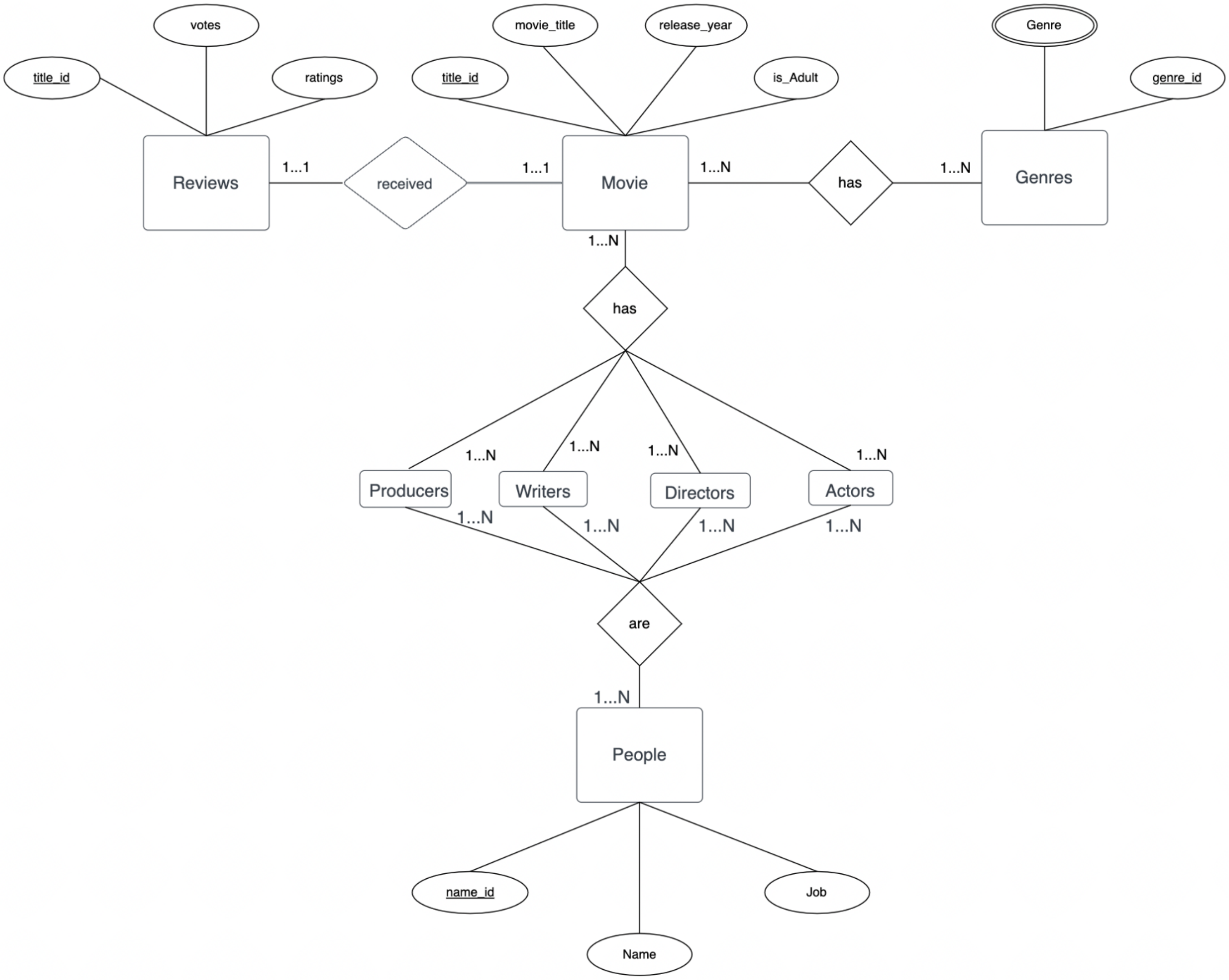
Assignment 1

Anikhet Mulky
am9559@g.rit.edu

Q1)



Q2

Movie = (title_id , movie_title, release_year, is_adult)
Person = (name_id , name)
People = (id, title_id,name_id, job)
Genre = (genre_id , genres)
Genre_table = (id , title_id, genre, genre_id)
Reviews = (title_id , ratings , votes)

SQL scripts are attached in the code file.

Q3
Dataset

1) Title.basics : This file contains basic attributes of a movie like id, name of the movie in English and its original name, its release year and end year along with its run time and genres.

Tconst : It is a unique id provided for each movie in the dataset.
primaryTitle : Movie name in english
originalTitle : The original movie name
isAdult : 1 if the film is for mature audience and 0 if the movie is fit for everyone
startYear: The year the movie was released .
runtimeMinutes : Amount of minutes the film runs for.
genres : This column gives information about the genre of the movies.

2) Title.principals: This file aims to connect the movie ids with the name ids and also gives description of occupation, job and the characters/ roles the person played.

Tconst : It is a unique id provided for each movie in the dataset.
Nconst : It is an id for the names of the people.
Category : Job description of every person related to nconst.
Job : Occupation of the person connected to nconst.
Characters : Characters played by the person (Generally actors and actresses)

3) Title.ratings: This file basically contains valid IMdb ratings and necessary information about votes for a particular movie title.

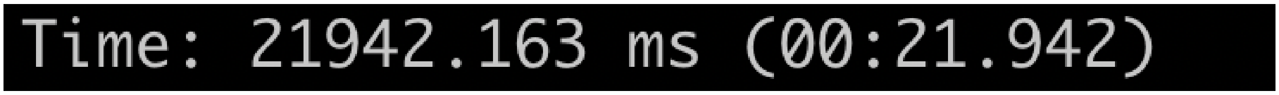
Tconst : It is an id which is provided for the title of each movie. It is in alphanumeric format and unique.

averageRating: It is an average of all the individual viewer/user ratings given for movies.
numVotes: It denotes the number of votes a movie title has earned.

4) Name.basics: This file basically represents collective information about names of people.
Nconst: It is an id for a name which is unique for each individual person. It is in alphanumeric format.
primaryName: It is a name used for identifying and addressing an individual.
birthYear: Date for an individual which represents the starting of his/her active years. In YYYY format.
deathYear: Date for an individual which represents the end or last of his/her active years. In YYYY format.
primaryProfession: Basic work/job of an individual. Top three of his/he jobs are shown.
knownforTitles: It represents the element for which a person or individual is known

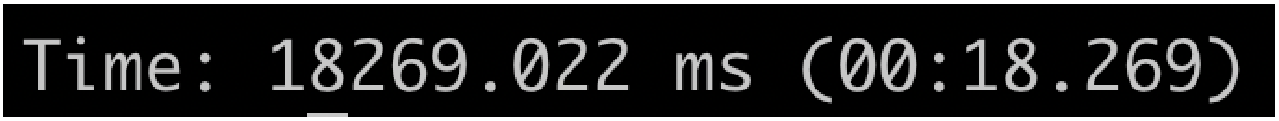
Q4

Movie table



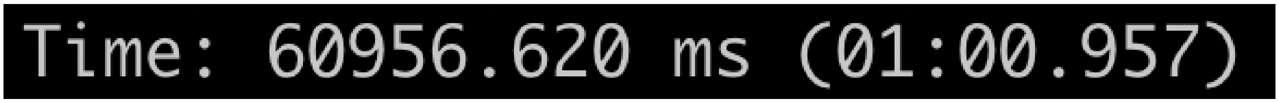
```
Time: 21942.163 ms (00:21.942)
```

Person table



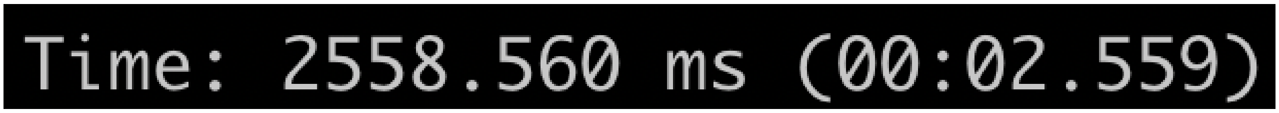
```
Time: 18269.022 ms (00:18.269)
```

People table



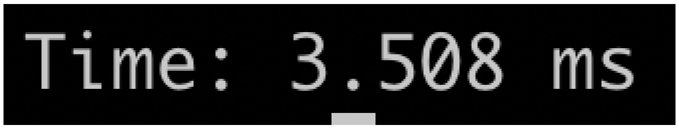
```
Time: 60956.620 ms (01:00.957)
```

Reviews table



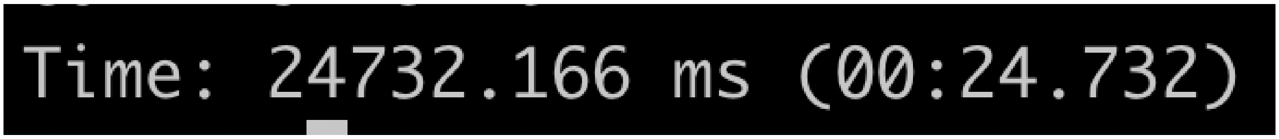
```
Time: 2558.560 ms (00:02.559)
```

Genre_table



```
Time: 3.508 ms
```

Genre



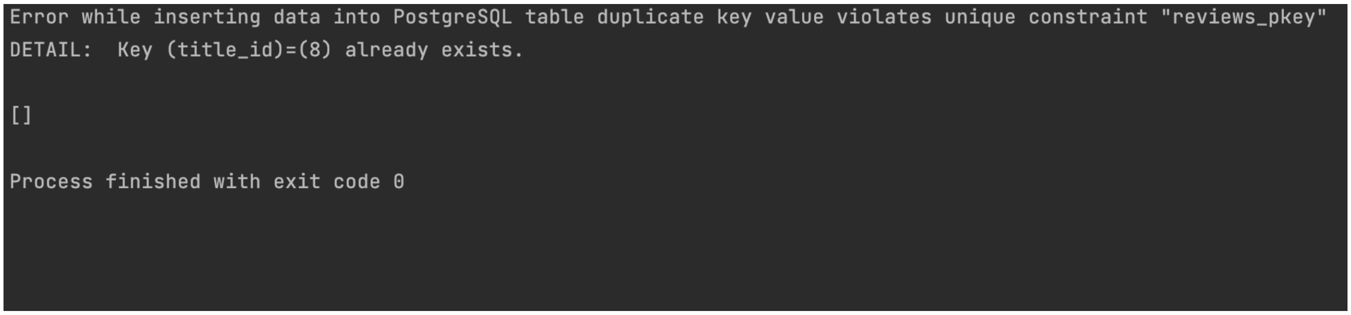
```
Time: 24732.166 ms (00:24.732)
```

The problem of invalid foreign keys was tackled by skipping rows which were not valid in the preprocessing of the data.
Above is the time taken by each Copy command to upload to the table

Q5

```
INSERT INTO reviews VALUES (8, 5.6, 1590) ,(8, 5.6, 1590), (3, 5.6, 1590);
```

This command tries to insert two same values for the primary key in the second row.
The transaction aborts and no rows are inserted.
Following image shows no rows were inserted.



```
Error while inserting data into PostgreSQL table duplicate key value violates unique constraint "reviews_pkey"
DETAIL:  Key (title_id)=(8) already exists.

[]

Process finished with exit code 0
```