ETL Project - Grace, Jim, Lori and Meghan

We chose to create a database that combines data about movies with which streaming services (Netflix, Amazon Prime, Hulu, and Disney+) offers the movie. By combining the datasets, our database will be more useful to users because they can see all pertinent information about a movie, including reviews, and can also quickly see which streaming source it’s available to watch. The two datasets are then used to create a relational database that is normalized in PostgreSQL.

Steps taken:

1. Database schema defined on quickdatabasediagrams.com:

A screenshot of a cell phone

Description automatically generated

1. Exported from [quickdatabasediagrams.com](http://quickdatabasediagrams.com/) to predefine the database in PostgreSQL.
2. Run ETL process once the database is configured.

EXTRACT:

First, the following datasets were loaded. CSV files are in the Resources folder

1. MoviesOnStreamingPlatforms\_updated.csv - source kaggle.com - <https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney>
2. IMDB movies.csv - source kaggle.com - <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv>

TRANSFORM:

1. Streaming Data:

* remove all unnecessary columns (determined during the schema creation)
* drop any rows that have data missing (NaN values)

1. IMDB:

* remove all unnecessary columns (determined during the schema creation)
* drop any rows that have data missing (NaN values)

LOAD:

Since the tables have been precreated in PostgreSQL, loading of the data can happen only once (or primary keys will be violated). A test query combining all tables shows it worked correctly.