ETL-Project Movie and TV Show Data Base

Arthur Ibanez

Peyton Harsh

Osvaldo Meza

Description

To help people with their sudden newfound boredom, we have assembled a movie and TV-Show data base that will help sort through the many options available. The following is a small breakdown of the process undertaken to establish the database and make it accessible for a user.

Extracting Data

Movie and TV-Show data was gathered through Kaggle.com and data.world. CSV files were downloaded for Netflix data, TMDB (The Movie Data Base), and IMDB (Internet Movie Database) from Kaggle, as well as a separate movie database was used from data.world. All data sets and files were saved in a “Data” folder and read into a jupyter notebook to clean the data and sort through unnecessary pieces.

Transforming Data

After all csv files were read into a jupyter notebook, the data was condensed to only include relevant columns that would have a correlation to the other data sets that were extracted. There were 2 movie data sets where certain elements were imbedded as a list of dictionaries in a single column; Python was used to run a for loop and nested loop to extract the data. Once extracted, the cast, character, and id were saved into lists that could then be used to create separate tables. These new data frames were then exported as SQL files to create a data base where primary keys were assigned to link together the different tables that were created. SQL schemas for additional tables were produced through pgAdmin4, these tables were populated by importing CSV files that had been cleaned with Pandas.

Loading Data

Data was lastly stored on SQL databases where the transformed datasets were used to create relational tables that could then be used to query movie, cast information, character information, release date, along with additional movie and TV show information. SQL Alchemy was also used to load the data sets into postgres.