**ETL Project – Group 5**

Jennifer Darby, Pankaj Tahiliani, Tobias Judd

Team’s github repo:

<https://github.com/Pankaj-gh/ETL-Project>

What data sources you chose, and why?

We selected two data sources. The first data source was available from [Wikipedia](https://d.docs.live.net/c0fbc2e3a74a4ba8/Documents/Wikipedia) and provided a list of “Top 100 Best American Movies” determined by the American Film Institute. This list of movies was based on films that displayed significant create and production elements and had included critical recognition, major award winning, popularity over time, and historical / cultural significance.

The second data source came from [Kaggle](https://d.docs.live.net/c0fbc2e3a74a4ba8/Documents/Kaggle) and provided a list of more than 16,000 films and identified which streaming services they were available on. Streaming Services included Netflix, Prime Video, Hulu, and Disney+. This set of information was available in a csv format and viewed as interesting because it included the following:

* Which streaming platform(s) can I find this movie on?
* Target age group movies vs the streaming application they can be found on
* The year during which a movie was produced and the streaming platform they can be found on

Detailing the process of the extraction, transformation, and loading steps

We first extracted the table from the Wikipedia page and stored it as a dataframe with the Film and Director columns. Then we went to Kaggle and looking for data containing films with what streaming apps they were showing on. This data was stored as a .csv file, we transformed using Python. We eliminated all the media types except the movie types. Then, we created a column that combined the streaming services together in a field called streaming\_channel which would identify all sources for which the movie could be watched.



We stored both datasets into PostgresDB. Using flask, we retrieved the datasets and assigned them routes. We originally converted them to jasonified lists. Because we had added time, we used HTML and CSS styling to transform them.

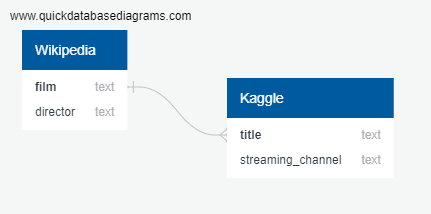
Explain why you have performed the types of transformation you did

We believed that people who wanted to view historical movies and or were in the market to subscribe to a streaming service that it would be beneficial to review prior to signing up. This combined dataset would allow individuals to select which type of service either offered the most number of Top 100 Movies from a historical context or to identify which streaming services may have movies that you have not seen before.

Why you chose the type of final database

We have chosen to use a relational database for the two datasets. This is so that we can tie together to datasets defined by the Title of the movies and their associated data. Both datasets were very structured with fit well with the PostgresDB SQL database.

Schema of the tables/collections in the final database



Plan is to extract and link the Top 100 movies mentioned on Wikipedia with Netflix’s Movie & TV Shows is 2020 to see what movies can be watched on the streaming service that is relevant to the Top 100. Likewise, one could understand how movie ratings / rankings have varied historically versus current rating.  The viewer could group and analyze by release year, director, genre, and other variables to understand movie trends.

Hypothetical use case(s) for your database

Plan is to extract and link the Top 100 movies mentioned on Wikipedia with Netflix’s Movie & TV Shows is 2020 to see what movies can be watched on the streaming service that is relevant to the Top 100. Likewise, one could understand how movie ratings / rankings have varied historically versus current rating.  The viewer could group and analyze by release year, director, genre, and other variables to understand movie trends.