John McBride and Vamsi Navuluri

1/15/2020

ETL Project Technical Writeup: Netflix and Golden Globes

**Extract:**

The data sources we used were sourced from <https://www.kaggle.com/datasets>; specifically we chose the Netflix <https://www.kaggle.com/shivamb/netflix-shows> and the Golden Globes (<https://www.kaggle.com/unanimad/golden-globe-awards>) datasets. The purpose of our project was to be able to cross-reference the two datasets to see what films/television shows featured on Netflix had won a Golden globe award. The datasets were formatted in CSV files that then we imported to our jupyter notebook and then into pgAdmin.

**Transform:**

Both exported CSV dataset files were then imported into a jupyter notebook to clean and append the data. In cleaning our data we had to remove all the Golden Globe award categories that did not contain the terms “Picture”, “TV”, “Film”, “Television”, so that our dataset was only left with either TV or Movies that we could cross-reference in the Netflix dataset. The Netflix dataset type contained the terms “TV Show” and “Movie” so that dataset was clean from the beginning. The next transformation in our Jupyter notebook that was required was to join the two cleaned datasets together so that we could filter and aggregate together the same shows on Netflix that had won a Golden Globe award. We added a primary key to our Golden globes dataset by creating a column called “Serial ID” which added 1 number to itself for the entire length of the data frame. In our Golden Globe dataset there were some “NaN” values in the film column so we cleaned the data by replacing the “NaN” values with the data in the “nominee” column.

**Load:**

After we had cleaned the data we connected with our jupyter notebook to the Postgresql pgAdmin database we had created and used pandas to load our csv converted Dataframe into the database. Then we confirmed our data had been added by querying both databases in our jupyter notebook with a pd.read\_sql\_query. We chose Postgressql/ pgAdmin due to its ease of creating, importing, and reviewing databases.