DJ Bonifacic, Ian Pearlstein and Diego Robayo

ETL Project

10/19/19

Executive Summary

The end of the year movie season is upon us, so we decided to create a database involving different stats for films. Our first step for this ETL process was to find our data. Using Kaggle, we were able to find datasets related to movie sales, ratings and genre.

Next, we needed to transform our data, so we could eventually merge it together. We cleaned the three files using pandas by dropping columns, removing NaNs, setting titles to lower case and splitting data into two columns using delimiters. After that, we merged our datasets based on the title of each movie. We also created a unique title\_ID as well. Once the data was ready, we split our merged data back into a Movie\_Genre, Movie\_ID and Movie\_Rating. This way, each table had a unique ID that would help merge them back together.

Lastly, we needed to load our data. We created a database in pgAdmin with the five different data tables we listed above and uploaded each one. Once the data was ready, we were able to create different queries based off the year, title, gross sales, etc.

