**ETL Project – Group 5**

Jennifer Darby, Pankaj Tahiliani, Tobias Judd

Team’s github repo:

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

Proposed datasets:

<https://en.wikipedia.org/wiki/AFI%27s_100_Years...100_Movies>

<https://www.kaggle.com/intandea/netflix2020?select=USA_NF.csv>

What data sources you chose, and why? ● Detailing the process of the extraction, transformation, and loading steps ● Explain why you have performed the types of transformation you did ● Why you chose the type of final database ● Schema of the tables/collections in the final database ● Hypothetical use case(s) for your database

What useful investigation could be done with 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, understand how movie ratings / rankings have varied historically versus current rating. Could group and analyze by release year, director, genre, and other variables to understand movie trends.

Whether final database will be relational or non-relational. Why?

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.