ETL Challenge - Project 1 Team 3

Team:

Dan Freda

Jenish Jariwala

Banu Nathan

Mona Peteet

We will be working with a dataset sourced from grouplens.org, called MovieLens Latest Datasets. This dataset provides ratings from over 600 users of 9000 movies. Analysis of the specific data files to use will be finalized based on exploration of the questions we would seek to answer. The files under consideration include: *movies.csv*, *ratings.csv, tags.csv, links.csv.*

Data extraction, transform and load activities will include relational database design, diagramming and creation. Data transformation activities will be performed in Pandas and will include cleaning and filtering the data. The ability to execute queries that answer the following questions will be designed.

Movies database querying:

1. How many movies were released in a specified timeframe (e.g.2019)? [PostgreSQL]
2. How are movies classified among genres? [Pandas]
3. Being able to get results of specified criteria (e.g. movie titles) [PostgreSQL]
4. What are the most frequently occurring genres (barchart)? [Pandas]
5. What is the distribution of movies by rating (pie)? [Pandas]

<https://grouplens.org/datasets/movielens/latest/>