ETL Challenge - Project 1 Team 3

Team:

Dan Freda

Jenish Jariwala

Banu Nathan

Mona Peteet

We will be working with a dataset sourced from Kaggle, called The Movies Dataset. This dataset provides ratings from over 270,000 users of 45,000 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: *credits.csv*, *keywords.csv, links.csv, moives\_metadata.csv and ratings.csv.*

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

Movies database querying:

1. What are the most popular genres?
2. What is the revenue of a specific genre, by director/actor, etc..
3. What are the most popular genres, or the least popular genres?
4. What is the popularity by production studio?
5. How many movies were released in a specified timeframe (e.g.June 2019)?
6. Which user rated the most movies?
7. Did users rate the most popular movies?

<https://www.kaggle.com/rounakbanik/the-movies-dataset?select=ratings_small.csv>