**Group #6**

Cindy Brady

Patrice Butler

Ramario Napier

Brittany Chin

ETL Project

**December 1, 2020**

# Overview

Extract, Transform, Load. Find a data set to extract. Clean the data. Write a report on how the data could be used in production.

# Goals

1. **The sources of data that you will extract from.  
   Movies on Netflix, Prime Video , Hulu, & Disney+ - and Netflix Originals - Kaggle, both CSV**

**https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney**

**https://www.kaggle.com/swapnilbhange/netflix-original-movies**

1. **The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).  
   Joining, filtering, cleaning, aggregating**
2. **The type of final production database to load the data into (relational or non-relational).  
   RDBS - Postgres**
3. **The final tables or collections that will be used in the production database**

**TBD**

# Milestones

## Choose two datasets (potentially more)

Choose two similar datasets to join

## Clean and aggregate the data

Reduce columns, perform

## Load the data into a Postgres server

Create DB, create tables, and load in the cleaned data

## Complete a project report

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* **T**ransform: what data cleaning or transformation was required.
* **L**oad: the final database, tables/collections, and why this was chosen

## Project Presentation

ETL Report

At the end of the week, your team will submit a Final Report that describes the following:

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* Pulled CSV files from kaggle (listed above)
* **T**ransform: what data cleaning or transformation was required.
* Dropped unnecessary columns listing things like (age, ID, year etc.)
* Removed all null values
* Reformatted columns for data consistency, numerical data was reformatted to decimals
* Aggregated the data and found mean, and count of different streaming platforms (boolean to int)
* Output Director dataframe to CSV
* Merged the two datasets on movie title and identified netflix originals
* Uploaded new merged netflix originals dataframe to postgres using SQL queries and connection engine
* Selected data from the postgres database to confirm data and verify that data exists in the table
* **L**oad: the final database, tables/collections, and why this was chosen.
* We chose to load data to Postgres stored on a local database
* Database Title - Streaming\_DB
* Table Title - netflix\_originals

Why was this chosen? We had a group discussion about the last Netflix movie/series we watched. This brought up the topic of different streaming services and which services provided the same movies and TV shows. We decided to narrow our findings down to Netflix originals only.