**ETL Project Proposal**

**Group Members:** Veera Supraja Koppisetty, Felix Ogbodu, Sindhura Surapaneni, Chuck Youngman, Albert Burroughs.

**Project Description/Outline:**We will combine two datasets to create a summary that shows all the movies in theatres and the streaming platforms.

**Research Question to Answer:**

What is viewership with all movies in streaming platforms vs movie theaters?

**Datasets to be used:**

We will be using 2 csv Movie datasets.

1. One dataset is from Kaggle that have the movie streaming platforms data.

[Movies on Netflix, Prime Video, Hulu and Disney+ | Kaggle](https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney)

1. One dataset is from the IMDB movie dataset.

<https://drive.google.com/file/d/1IFDKr1Dv4iIhaAEQb29tQ8CmOWyVenQT/view>

**Extract:**

We will get following columns from dataset that contains streaming platforms:

1. Title
2. Year
3. Rotton Tomatoes score
4. Netflix
5. Hulu
6. Prime Video
7. Disney+
8. Directors
9. Genres
10. Country
11. Language
12. Runtime

We will get following columns from dataset that contains IMdb Movie Information:

1. Title
2. Release date
3. Genre
4. Language
5. Country
6. Lead Actor
7. IMdb Score
8. Duration
9. Gross Income
10. Budget

**Transform:**

In the Imdb dataset the year will be extracted from Release date column. We will create some aggregate columns.

**Load:**

We will be loading data into two separate tables in the Postgres named Movie Database.