**ETL Project**

Kenny Dao

Sean Pei

[Mickey Sandhu]

Our task was to source, clean up and load the data to a production database. We used the Pandas library for our data manipulation. We used Jupyter notebook to pull API from selected data source.

Also, we created DataFrames using CSV files about books data from Kaggle and Goodreads API.

Data source:

1. Kaggle: <https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m?select=book900k-1000k.csv> (dataset is being updated every 2 days)
2. API: <https://www.goodreads.com/api>
3. <https://www.kaggle.com/pelinsoylu/amazon-the-most-read-books-of-the-2019-dataset> (bestsellers with categories.csv)

We started with books data sourced from Kaggle, it’s a csv file with 40,890 rows and 20 columns. We chose to focus on the columns: BookId, Title, AuthorName, PublishYear, PublishMonth and PublishDay.

Created a new authors table using the AuthorName to pull from API and parse the data to include more detail information about an author, such as AuthorId, Gender, Hometown and books written by those authors.

Created author\_book table including AuthorId, BookId and BestSeller (to determine whether a book title is one of the best seller books.

Load

We loaded our CSV into Postgres as our database and make it available on-line using google cloud platform.