Project 2 – **e**xtract **T**ransform & **L**oad

# Jim holland – fnu sandhiya – julie novic – sarah steimle

## Project Proposal

We will be performing the ETL on Google Play Store Apps datasets. The purpose of this ETL project is to produce a dataset suitable for market research purposes, looking at the review content to better prepare for updates to the apps.

## Extract

The source data was gathered from Kaggle.com via two (2) CSV files containing Google Play Store app information. The first file contained app meta-data, including app name, average rating, number of downloads, etc. The second file contained the top 100 most relevant reviews for the same apps.

## Transform

The two files were loaded in Pandas and merged on the app name with a right outer-join based on the app meta-data data-frame. Once the data was merged, some columns were dropped that were deemed unnecessary for our purposes. Finally, any reviews with null values were dropped from the data set. The result was 72,566 rows of clean data.

## Load

SQLAlchemy was used to create the database engine. Then, Pandas converted the data frame into the SQL database using pymySQL. Using the engine, the database was queried to print out all the records contained in the data frame to ensure the database was built successfully.

## Database meta-data

|  |  |
| --- | --- |
| Field Name | Description |
| App | App name |
| Category | App genre |
| Rating | Average user rating (sum of ratings / # of ratings) |
| Reviews | Number of reviews made on the app |
| Installs | Number of unique installations on devices |
| Type | Free or Paid |
| Price | Listed price in app store |
| Current Ver | Current app version available in the app store |
| Android Ver | Current Android version compatible with app |
| Translated\_Review | User free-text input, translated into English |
| Sentiment | Categorizes overall user experience, from negative to positive |