ETL Group Project

Our ETL project is based on Amazon ratings for cell phones. The dataset was pulled from the Kaggle website. The dataset comprises of two CSV files (items.csv, reviews.csv). The first file contains 10 columns of information on the cell phones e.g. phone brand, phone model, amazon id (asin) etc. The second CSV file contains 8 columns of ratings data by several reviewers e.g. reviews, name of each reviewer, rating, amazon id etc.

**EXTRACT**

The data in the two CSV files was extracted by reading the raw data into two Panda DataFrames. The DataFrames are called “phone\_data\_df” and “reviews\_data\_df”.

**TRANSFORM**

The following transformations were applied to the data from the csv files:

***Removed columns not needed:***

1. We chose to use three columns (asin, brand, title) out of the 10 contained in the “items” file.
2. We chose to use two columns (asin, ratings) out of the “reviews” file

***Cleaned Data***

1. Removed null values (NaN).
2. Removed items with no price.
3. Renamed columns from csv files to match what we have in the database. e.g. “asin” was renamed to “id”, “rating” was renamed to “avg\_rating” e.t.c

***Aggregate***

1. Grouped the “reviews” dataframe by “asin” to get the average rating.

**Joins**

1. Merged the two dataframes into one dataframe using the “id” column.

**LOAD**

The final dataframe was loaded into a postgres SQL table.