**ETL Project**

Group 12:

Adrian Jackson

Katie Newman

**E**xtract: The U.S. Department of Agriculture has many datasets related to food, ingredients, and nutrition. Initially we attempted to use the API provided through FoodData Central. The API would only return data once an ID was specified. Our goal was to have a dataset with all food, so we decided to download the csvs provided on the USDA website. After viewing each of the files, we downloaded two particular csv files: “branded\_food” and “food.” Each csv file contained 260,370 rows of data, which we then converted into a pandas dataframes.

**T**ransform: First, the necessary columns were copied for each of the two dataframes, dropping the columns we did not need. Next we wanted to do some work on the branded\_food database. We started by separating out each of the ingredients in the column called ‘ingredients.’ With the help of ‘str.split’ (to separate by comma) and ‘melt’ (to transform the separated ingredients from columns into rows) we created a new dataframe. Since each product has a different number of ingredients, we were left with a lot of null values. The null values were dropped and we were left with 3,693,889 rows of data!

Meanwhile, we needed to loop through the ingredients to find which ones contained organic items. This was done using by reducing the branded\_food csv into two columns: fdc\_id and ingredients. Once the new dataframe was built, a regular expression was used to find the occurrences of the word “organic” (re.compile(“\W\*(ORGANIC)\W\*”). Using the regular expression inside a for loop all instances of organic ingredients were aggregated into two new columns. These new columns can be used later for further analytical insights.

**L**oad: We chose to load our data into SQLite. A connection to the database was established using Jupyter Notebook. Next, we checked to make sure that the connection was valid. We then did a join on our dataframes, combined\_df and ingred\_df.