**Food for Thought - Extract-Transform-Load-Project**

**Team**: Tolani, Joy, Andrea, Manuel

**Objective and Methodology**:

The internet is a great resource for aspiring chefs, casual cooks, and for people wanting to make healthier food and lifestyle choices in order to lose weight or feel better. One of the largest collections of recipes can be found on [allrecipes.com](https://www.allrecipes.com/). There are several categories of recipes, including Diet and Health, which includes a curated list of [healthy recipes](https://www.allrecipes.com/recipes/84/healthy-recipes/). In order to verify the nutritional value of each recipe, we used data from the United States Department of Agriculture’s [FoodData Central](https://fdc.nal.usda.gov/) download site. By merging the two datasets, we provide a vehicle to find not only tasty recipes, but nutritional ones as well.

**Data Sources:**

<http://www.allrecipes.com/recipes/84/Healthy-recipes/>

<https://fdc.nal.usda.gov/download-datasets.html>

**Data Cleanup & Analysis: Extract and Transform**

The notebook “Final\_recipes\_pulled\_into\_pandas” outlines the first step in this study: web scraping the Allrecipes.com website for healthy recipes. As part of the scraping process, we extract the recipe title, recipe link, and ingredients using Selenium, Beautiful Soup, PyMongo, and MongoDB. The results are then transformed to a DataFrame using Pandas. Please refer to the “Final\_recipes\_pulled\_into\_pandas” notebook for specific comments about the code.

The notebook “recipes\_csv” outlines the next steps in this study:

### Using Pandas to clean and join recipes with ingredients and add to MongoDB collection.

* Extract FoodData Central dataset and clean data using Pandas
* Convert datasets to json format and export to MongoDB using terminal

Recipe title and ingredient data were joined using Pandas and the combined DataFrame was organized by removing id columns added by MongoDB. Additionally, the scraped data contained unnecessary symbols and letters (“, b, ‘, / , etc.) Using Pandas, the data was exported as a csv, then reimported as a new DataFrame (“recipes\_df”). Unnecessary columns were again removed, leaving just the recipe title and ingredients.

The USDA FoodData Central download site provides nutritional information in the form of Excel files. Using Pandas’ read\_excel Method, a food DataFrame was created. 8790 ingredients were imported, and columns without complete nutritional information were removed. The data was further refined into a dictionary format by converting to a json format.

**Data Cleanup & Analysis: Load**

With the cleanup of data completed, the datasets were loaded into the non-relational database, MongoDB utilizing the terminal in VSCode. We chose to use this database due to the variance of ways to quantify recipe ingredients. Instead of trying to resolve this issue in this one week project, we took advantage of the NoSQL approach and imported ingredients as is.

MongoDB: allrecipes\_etl1

Collections:

collection = db.recipes

collection2= db.ingredients