Homework 2 (HBase) - Part 1

Dave Millman
Ben Polk
CSCI 540 - Fall 2021

October 21, 2021

Given the MyPyramid Raw Food Data Set, I wanted to answer the question, "What is the nutritional profile of any food that includes the word 'fried' in its display name?" More specifically, I set out to look up the caloric value of all fried foods in the dataset.

In order to answer this question, I decided to use a combination of food name and portion name for the row key. Since I would be searching for rows based on food name (containing the word 'fried'), it made sense to use this as part of the key. Portion name was included as part of the key in order to ensure unique keys from entries in the data file. Since the data file contained multiple entries for each food, but with different portions, relying on food name alone would have resulted in data being overwritten. In addition, including portion name in the key allows you to view how nutritional information changes for different portions of the same food.

My schema used a single column family of 'nutrition'; this seemed like an apt description for the type of information I cataloged from the file. Specifically, for each food I added columns of 'solid_fats', 'added_sugars', 'alcohol', 'calories', and 'saturated_fats' to the 'nutrition' column family. This would allow for easily querying to find certain nutritional aspects of foods.

Given this schema, answering my question was straightforward. I was able to use a RowFilter, matching on a regular expression to get any food names that contained the word 'fried'. I was also able to project the 'nutrition:calories' column, and thereby find the caloric content of all fried foods in the dataset.