

Extraction, Transformation, and Load Technical Report

bacon pricing database construction and maintenance



Beautiful Bacon Group

Amit Patel, Austin Wen, Preston Hinkel

Contents

[Introduction 2](#_Toc41737710)

[Scope 2](#_Toc41737711)

[Project Members 2](#_Toc41737712)

[Definitions, Acronyms and Abbreviations 3](#_Toc41737713)

[ETL Methodology 3](#_Toc41737714)

# Introduction

The purpose of this Extraction, Transformation, and Load (ETL) Technical Report is to help Beta Alpha Kappa National Fraternity office (the Client) to track daily bacon pricing. Bacon represents over 80 percent of the Client’s annual food expenditure; therefore, the Client engaged Beautiful Bacon Group to create an in-house process to monitor and to store bacon related data. First, the Client would like to obtain a daily bacon pricing feed from a specified grocer to measure the immediate cost of consumption. Second, the Client would also like to track bacon’s consumer price index (CPI) to better understand the effect of inflation. Finally, the Client would want to receive lean hogs commodity pricing data for future hedging strategy.

# Scope

The scope of this report covers the following:

1. Smithfield Thick Cut Naturally Hickory Smoked Bacon 16oz pricing data from Kroger.
2. Bacon, sliced, per lb. in U.S. city average price from US Bureau of Labor Statistics.
3. Lean Hogs futures from Investing.com.

Beautiful Bacon Group should create an automated process to obtain all above data from their respective sources. The obtained data should be stored in a database for the Client to retrieve easily.

# Project Members

Our team Beautiful Bacon Group consists of three data scientists: Amit Patel, Austin Wen, and Preston Hinkel. Mr. Amit Patel focused on Kroger pricing data acquisition and Github repository maintenance. Mr. Austin Wen focused on obtaining the US Bureau of Labor Statistics bacon CPI data. Mr. Preston Hinkel focused on retrieving Lean Hogs futures pricing from Investing.com.

# Definitions, Acronyms and Abbreviations

This report is technical in nature. Below is a list of definitions, acronyms and abbreviations for the Client to better understand the report.

1. Python – An interpreted general-purpose programming language that Beautiful Bacon Group used as the main engine for this ETL processes.
2. Python Module – A Python file contains additional useful statements and definitions in additional to the default. The modules used in this project are: Pandas, JSON, Requests, Sqlalchemy, Splinter, and BeautifulSoup.
3. PostgreSQL – An open-source relational database management system for data storage.
4. CPI – Consumer Price Index. A measure for changes in the weighted price level of a given product or basket.

# ETL Methodology

Kroger Smithfield Thick Cut Naturally Hickory Smoked Bacon 16oz pricing data

**Extraction**

Kroger is the Client’s preferred retail grocery store to procure Smithfield Thick Cut Naturally Hickory Smoked Bacon 16oz. Given the Client’s interest in tracking the daily price of the product, Beautiful Bacon Group created a Python function that access Kroger.com website directly, and then reads in all available HTML data. The initial HTML data contains more than the pricing data we are interested in. The Python function can pinpoint and extract the pricing data from the HTML raw data.

**Transformation**

Once the price is located, we used Python DateTime Module to give the pricing data a date timestamp. And then convert the data into a Python Pandas DataFrame for storage.

**Load**

Once the data is in a Python Pandas DataFrame format, we deployed PostgreSQL, an open-source relational database management system, to store the data into the database bacon\_db under Table kroger\_price.

US Bureau of Labor Statistics Bacon CPI Data

**Extraction**

The United States Bureau of Labor Statistic (BLS) publishes sliced bacon per pound consumer price index data on a monthly basis. The data history goes back to year 1980, and it is freely available via their website <https://data.bls.gov/timeseries/APU0000704111>. The website also provides an Application Programming Interface (API) method for users to download data in JSON format. We created a Python function (download\_bls\_api) that connects to BLS’s API and downloads bacon CPI data. The function defaults to pull every data point since the beginning of the previous year. However, it can be adjusted to accommodate wider time period if needed.

**Transformation**

The initial bacon CPI data download is in JSON format. The raw data comes with excess information such as download status and empty footnotes. We created another Python function (transform\_bls\_data) that processes and distills the raw data to two columns, a year/month column and a CPI value column. The clean result is then converted into a Python Pandas DataFrame for easy display and storage.

**Load**

Like the previous Kroger pricing data ETL, we store the clean CPI data into a PostgreSQL database bacon\_db. The function we created is called store\_cpi\_data\_to\_db. The function establishes connection to the PostgreSQL server, and then stores the DataFrame into the database. It has a build-in mechanism to add only the newly available CPI data to avoid duplicates in the database.