1.3 Technologies and resource contributions

Our team for “Beautiful Bacon” consists of Amit Patel ,Austin Wen and Preston Hinkel. The 3 of us worked on different websites to collect our data.

Mr. Wen looked at the “Bureau of Labor Statistics Data”, <https://data.bls.gov/timeseries/APU0000704111>, to obtain the historical data over the slice of bacon over the years.

Mr. Hinkel looked at the “Lean Hogs Futures Historical Prices”, <https://www.investing.com/commodities/lean-hogs-historical-data>, to obtain the stock price of hogs.

Mr. Patel looked at the Kroger website, <https://www.kroger.com/search?query=Smithfield%20Thick%20Cut&searchType=natural&fulfillment=all>, to obtain the price of “Smithfield Bacon” as well as the date.

1.4 Definitions, Acronyms and Abbreviations

Python - Is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

BeautifulSoup - Is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. It is available for Python 2.7 and Python 3.

Sqlalchemy - Is an open-source SQL toolkit and object-relational mapper for the Python programming language.

Postgres - PostgreSQL, also known as Postgres, is a free and open-source relational database management system emphasizing extensibility and SQL compliance. It was originally named POSTGRES, referring to its origins as a successor to the Ingres database developed at the University of California, Berkeley

Splinter - is an open source tool for testing web applications using Python. It lets you automate browser actions, such as visiting URLs and interacting with their items.

ETL – Stands for Extract Transform and Load. We extract our data from the 3 websites and converted into a dataframe. We transform our data by cleaning it up. Then we load them into a database.

2. ETL DETAILS

2.1 Data Import/Extract Sources and Method

The website that we used to collect the data on the price of Smithfield Bacon was [www.Kroger.com](http://www.Kroger.com). We used web scraping to collect the price of bacon. What will be loaded is the price of Smithfield bacon to our PG Admin.

2.2 Data Acquisition

The range for the Smithfield Bacon is dynamic, the price always changes, so the price is never the same, it is always updating. To obtain the price we had to scrape Kroger website to look for the price and then store it as a list to convert it into a dataframe and then extract it to PG Admin. Some special attributes are, if there is a special price on the bacon when it changes, then the code has to be updated.

2.3 Data Transform

What we had to clean up was to collect only the price, so we had to strip the data to reveal just the price and the date.

2.4 Data Integrity

The reliability of the source data is always updating week to week. If there are any notifications it can be received thru an email. If the price for bacon has been changed then a notification will be sent out thru email.

2.5 Data Refresh Frequency

The frequency which this ETL process will be weekly

2.6 Data Security

No data security