Large Scale Database Project Report

University of Texas San Antonio

Sean Adames

Omar Aedh

General Purpose Data Scraper and webpage to look at your food ingredients

Abstract:

In today's world, there are millions of products available in the market to select from. Many products have different ingredient and it's important to know what you select and eat. In our LSDM project, we developed a webpage of products from three countries. The web service is linked to a snowflake database schema using SSL connection for security connection between the two ends. The webpage has an https certificate for secure browsing.

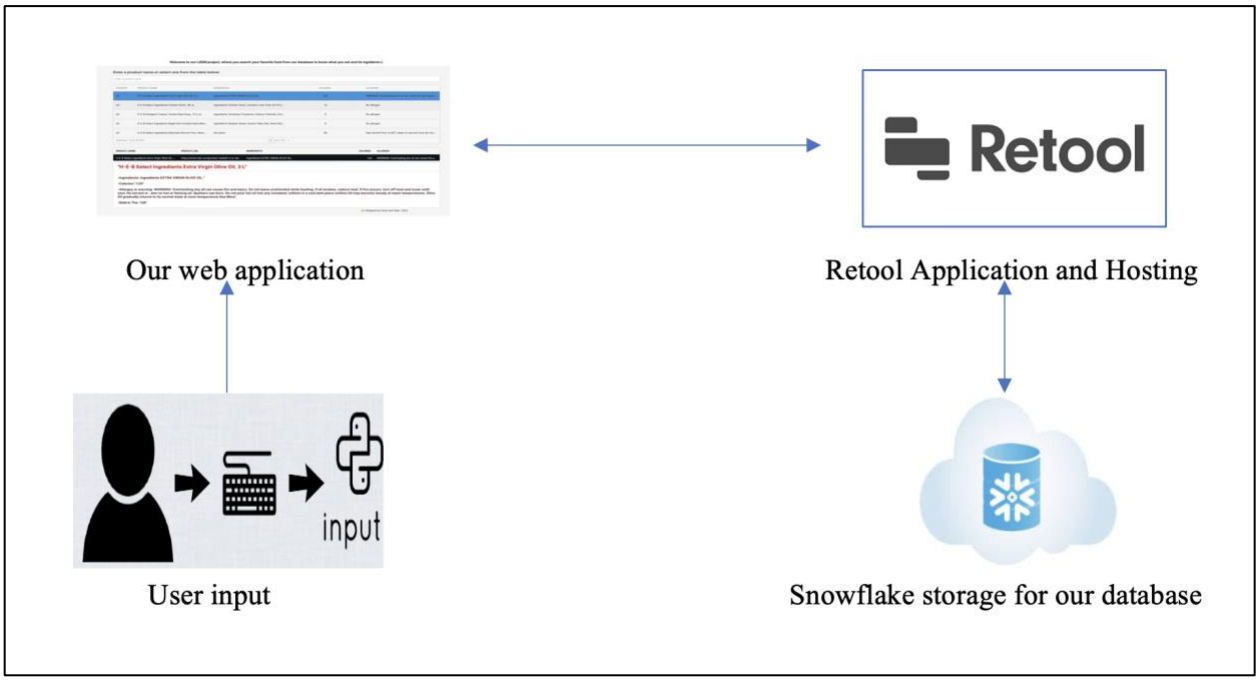
Introduction:

In this project, we provided a useful food products resource from the USA, UK, and AU so the customer can search find a product easily through our webpage. The user interface is easy and direct to use where it requires an input form the customer to find desired product and look its ingredients, calories, if allergen or not.

For our Large-Scale Database Project, we have selected to utilize the data scraping technique "Web Scraping" of grocery stores located in the US, Australia and UK. Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. The web scraping software may directly access the world wide web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page which a browser does when a user views a page. Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and telephone numbers, or companies and their URLs, or e-mail addresses to a list (contact scraping). We felt as though data scraping is a powerful utility in the space of large-scale data and wanted to explore it more and have a better understanding of it.

Our purpose behind selecting popular grocers and their respective stock as our data set was to have a data set that is not only user relatable but also important to everyday use in the world. The

data scraper project consists of three phases, in the first phase we scrape the web and available sources of these grocery stores to gather data regarding their inventory, more specifically their food products. Data should consist of name of product, price, allergens, most information one should see if they were to search for the item on the respective website. In phase two, we then process our data from a CSV file to a dedicated database. Databases are better for long-term storage of records that will be subject to changes. Databases have a far greater storage capacity than spreadsheets, which is why the transition from an Excel sheet to a database is necessary. In phase three, we compile everything we scraped from the web and processed into a database and allow users to organize the set using built queries and logic in a simple user-friendly application. This is accomplished using Retool, a free application building web tool.



Project architecture

Data collection :

One of the main tools we used to scrape for data was Octoparse. Octoparse is a modern visual web data extraction software. Both experienced and inexperienced users would find it easy to use Octoparse to bulk extract information from websites, for most of the scraping tasks no coding is really needed. It makes it easier and faster to get data from the web without having to hard code functions. We opted to use Octoparse over similar data scraping tools because it seemed the most user friendly especially to users new in the world of data science. The basic steps of scrapping the web for data is simple through these scraper tools. They utilize API's without having to hardcode them on your platform, everything is implemented via webpages and diagrams. Octoparse operates in a straightforward manner, you select your desired website and then setup your workflow. The workflow consists of several components that connect in order to

create a spreadsheet. One most manage which components the data scraper is looking for, items such as web image, price, allergens, once the parameters to scrape for are determined, you need to implement how your scraper traverses the pages via loops and pathing. Once the data scraper has what it is looking for and it is determined how to traverse the webpage/site, our workflow is complete, and we can begin to run the script to gather our data. Octoparse utilizes our customized workflow to fill a spreadsheet which can be saved and utilized in the next phase of the project which is storing the created CSV into a dedicated database.

Data processing and storage:

To process and store our data, a database is required, for this MongoDB was initially the selected tool for the project. MongoDB is a document-oriented NoSQL database used for high volume data storage. Instead of using tables and rows as in the traditional relational databases, MongoDB makes use of collections and documents. Documents consist of key-value pairs which are the basic unit of data in MongoDB. Collections contain sets of documents and function which is the equivalent of relational database tables. This was also selected due to its ability to integrate with several web tools, more specifically Retool, which will be used in the creation of the application for the web scraping project. Implementation of MongoDB is straightforward, after a CSV is created from data collection, it is then simply imported into our database.

Since extended JSON did not work as intended for our retool application, our team had to integrate over to Snowflake. Snowflake is self-described as a “best-in-class”, unified analytics platform that is independent from any underlying infrastructure. Connecting Retool to Snowflake simply took just a few minutes and lets you to build user interfaces quickly on top of the collected data. For example, you could build a tool to modify metadata tables. You can read data from Snowflake, and then write data back to it either directly, or via a separate API. Our main reason for using Snowflake is to use SQL queries instead of JSON.

To keep our data organized and allow users to query search based on their needs we opted to use a web tool to assist us, “Retool”. Retool is a fast and easy way to build and maintain internal tools. Integrating a data source like PostgreSQL, Salesforce, Firebase, MongoDB, and several others is user friendly and from there building queries and logic in SQL or JavaScript is made simple. Connecting built queries and logic to prebuilt components like tables, text inputs, and buttons is additionally made new user friendly with Retools help and tutorial features from there organizing and connecting these components to an application is made simple. The main appeal of Retool was its ability to integrate with Snowflake, not only did it save time but additionally saved the manpower required to create an html-based browser application.

Challenges and self-improvement:

The general-purpose data scraper project was not without its challenges and roadblocks. One challenge we faced as a team was the general concept of our project itself, web scraping. Web scraping was a new concept introduced to us as students, not covered in our course, so a great

deal of research was needed to be implemented to simply understand the main concept of our project. A roadblock with the data scraping was understanding how to traverse each page and analyze each grocery item in a unique way as to avoid repeating data, apprehending this error and determining the fix was slightly time consuming, since there was a large mass of data to be processed in each test. Another issue in the data scraping phase was the inability to access some websites and scrape through them. For instance, some grocer's websites in EU countries block the ability to scrape through them using generic scraping methods, this issue did not take too long to figure out however, we did not opt to look for a work around for this issue, rather we sample from a different source. In the second phase of our project where we needed to convert our CSV file to a database, we did not encounter many issues however, similar to web scraping, mongoDB was a relatively new interface and required a bit of learning to understand how to use it. When it came to extended JSON, retool ran into several issues with the querying. Several days of testing were costed only to result in needing to integrate our data through Snowflake In phase 3, where we then had to integrate our database into Retool, we encountered a few setbacks. As with learning any new interface or tool, Retool required a bit of learning to understand how to properly use it so we could develop our application. Understanding how to use these tools (MongoDB, Retool, Octoparse) was not exceptionally hard, only time consuming. Another issue with retool that we faced was the query portion of the application. We wanted users to be able to query specific items from the database but were having trouble implementing it correctly.

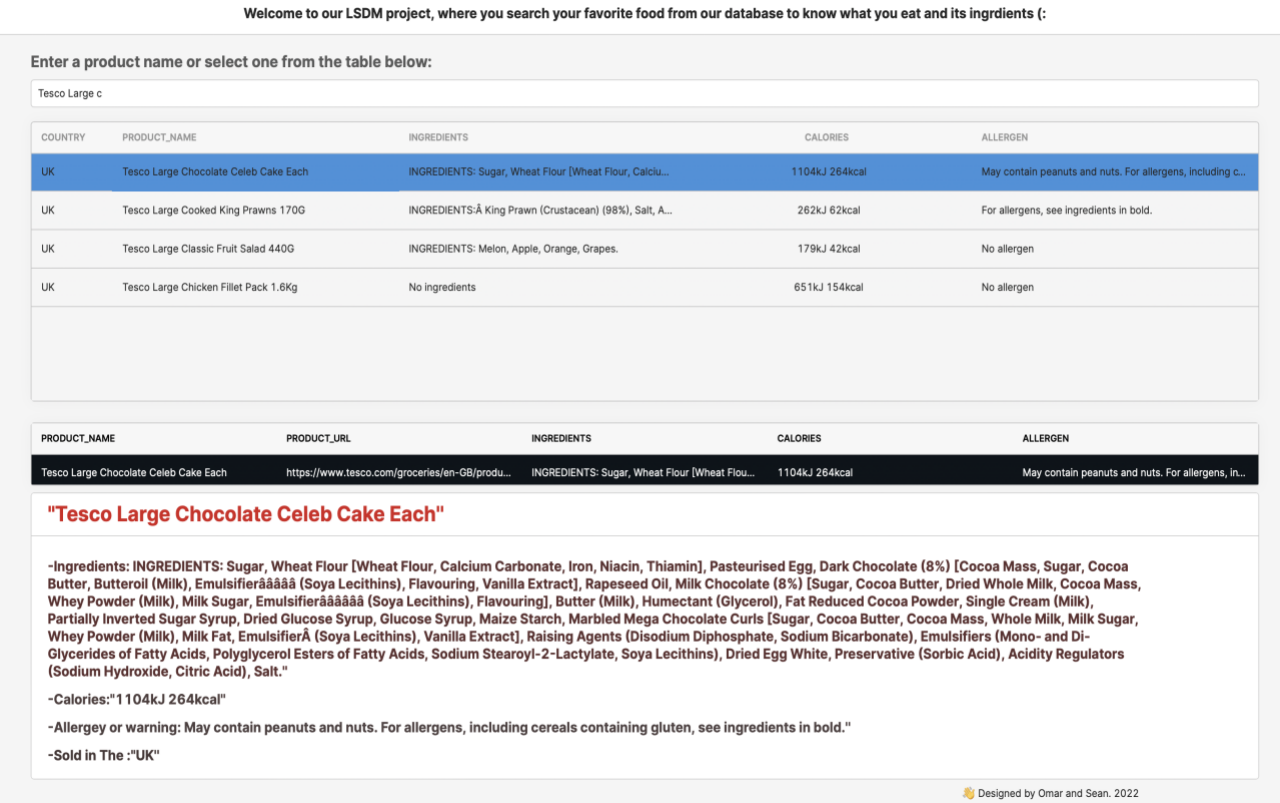
After successfully completing the data scraping project, it was insightful to learn and develop new skills in the realm of data science. Data scraping ended up being an extremely powerful tool for general data gathering, in gathering data from a company, in an Ecommerce setting, this would be an invaluable tool to conjure an understanding of one's competition in the market space. From the project, understanding how to data scrape and use this method to collect data proved to be an effective asset for our team, not simply for the course work, but for our individual skill set as computer scientists. Database management and creation was a skill we developed over the course of our class however, implementing it in our group project was a true test to our skills and what we have learned. MongoDB while being user friendly, can be cryptic and confusing to new users fortunately, with the skills we have developed over the semester creation of our own database proved to be uncomplicated. Completion of the project yielded many new skills, comparing the final production to the initial plan, our team is content that we did not deviate very far from our initial framework and kept our core ideas. With all projects there can be improvements and with the results of ours there is definitely room for improvement. While we created a workflow in Octoparse that gathers information on grocery items and the respective details of these items, we could take it a step further and grab not only everyday household items, but also clothing and technology items as well. Opting for more data would cost a tremendous amount of time to scrape through the websites for data, opening another door for improvement in our project, runtime of our data scraper. While there is only a finite amount we could improve our runtime, utilizing the workflow components in a more efficient manner could save minutes, up to possibly hours of running time. Using retool would work in the same fashion despite having more data, more sorting would need to be done but queries would operate the same. All these improvements are definitely possible to achieve and could greatly enhance user experience with our application. As the current build is finalized, our team achieved

everything we set out to complete with our data scraper, developing not only our technical skills as computer scientists but additionally our competence in a group setting.

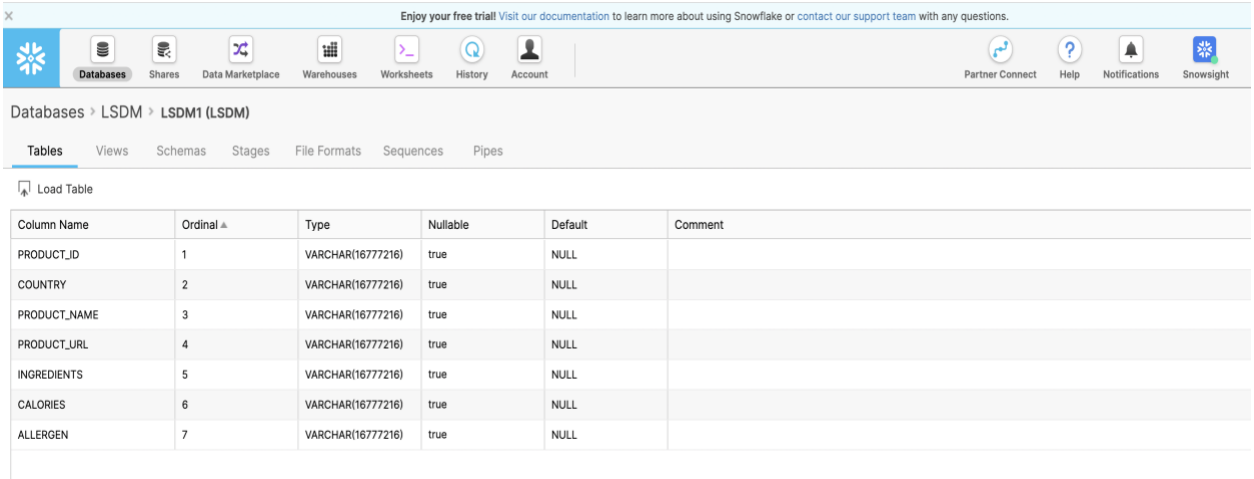
Demo of our project:

Please click on the following https link to access our project webpage:

https://lsdm.retool.com/embedded/public/b3de9b67-ffe0-4b48-a270-5b04ff37e3dc



Demo of an input by a user



Snowflake database

Future work:

Our plan is to implement a user personal account to save favorite products to remember its ingredients and avoid possible allergen. Moreover, we will allow a user input to add a new product to our database. Finally, expand our database to include more products from different countries.

Conclusion:

In our project, we developed a webservice to allow a user to search products from three countries (USA, UK, and AU) to see its ingredients, calories, and allergen. The project webpage is hosted by Retool application using SSL connection to our snowflake's database, and an HTTPS certificate for a secure browsing.

Team members contribution:

Omar Aedh: Data scraping, webpage development, database connection to Retool, and report.

Sean Adames: Webpage development, report, video recording, data collection and cleaning.

Resources:

1. https://www.octoparse.com/?gclid=CjwKCAjwvGUBhAzEiwASUMm4gGVotUZEcj8qxOr_DT3g10u-R4IFEaJ3BAtz1svdlT7BnFrJTXOLhoCsv8QAvD_BwE
2. https://realpython.com/beautiful-soup-web-scraper-python/
3. https://www.youtube.com/watch?v=XVv6mJpFOb0
4. https://www.youtube.com/watch?v=F6CEXNb54TI
5. https://docs.retool.com/docs/mongodb
6. https://docs.retool.com/docs/snowflake-integration
7. https://www.youtube.com/watch?v=lqFgt4_BS6o
8. https://www.youtube.com/watch?v=sv5niaHKo5g
9. https://docs.snowflake.com/en/user-guide/data-load-web-ui.html
10. https://www.youtube.com/watch?v=IjAflHMkuzk