Final Discussion Summary

Michael Wacey

For Professor Adewale Akinfaderin

SEAS 8515: Data Engineering for AI

June 1, 2024

George Washington University

# Introduction

Two years ago, a company was founded to address the need for better insight into on-shelf inventory in retail locations. They do this by engaging brand agents to go on missions into retail locations. The missions can include counting stock, rearranging products, retrieving recalled products, and anything else the client needs done. The data that is collected along with supporting pictures is provided back to the client.

The next phase is to ingrate the mission data with point-of-sale data to give manufacturers, distributors, and retailers a better understanding of where products are. This data will be made available to any organization that needs it. An example of what we plan to do is to run weekly product category missions to every Wal Mart in the US. This would generate about 20,000 pictures a week that need to be analyzed by an machine learning algorithm that can identify and count products.

# Description of the Business Problem

It is important that everyone in the consumer goods supply chain knows where every product is at every point in time. With this knowledge they can optimize the distribution and merchandising of products to optimize the profitability of manufacturers, distributors, and retailers. Manufacturers know how much product leaves their shipping docks, because of scanners retailers can report point of sale data in real time, everything in between (in transit, in warehouse, in backroom, on shelf) is currently a black hole of information.

Manufacturers do not have real-time data on where their products are in the supply chain. Distributors get just in time requests to distribute products and to deliver products to retailers. Retailers place orders to stock shelves but rarely get notified of delays or substitutions.

The current goal is to address merchandising in stores so that manufacturers and distributors have better insight into what is going on in stores. In the future this can be expanded to include distributors trucking and warehousing information.

It is important to note that some participants in the industry have solved parts of the problem for specific needs. These tend to be the larger players who have more influence over the supply chain. Wal Mart is one retailer that has significant influence on the supply chain and can do this. However, even Wal Mart has compliance issues when it comes to stocking shelves based on agreements with manufacturers.

# Data Overview

The primary data sets that will be used.

* Manufacturers distribution – the amount of product that a distributor has sent out by date. It can include the destination but does not usually.
* Planograms – agreement between a manufacturer and retailer on how products will be shelved. It can be as simple as one facing on the third shelf up from the floor to as complex as a detailed picture that shows shelf tags and more.
* On Shelf Inventory – A count of the number of a product on the shelf in a retail location.
* Point of Sale (POS) data – the products sold by a retail location. Often includes the data and time and the other products bought at the same time.
* Manufacturers – companies that produce consumer goods products.
* Distributors – companies that transport and warehouse consumer goods.
* Retailers – companies that sell consumer goods either through their own retail locations or franchise retail locations
* Consumer Goods Products – Products that are sold in retail locations
* Retail Locations – Physical locations that receive and sell consumer goods products.

| Data | Type | Source | Volume | Variety | Velocity | Veracity |
| --- | --- | --- | --- | --- | --- | --- |
| Manufacturers Distribution | Transactions | Manufacturers | 100,000 + per day | Some follow our standard, but most are unique to the manufacturer. Can be CSV or XLSX | Batch | High |
| Planograms | Transactions | Manufacturers | 1,000 per month | Generally, pictures and textual description. | Batch | Medium |
| On Shelf Inventory | Transactions | Internal | 1,000,000 per month | Collected in our application to our database | As data is collected | High |
| Point of Sale (POS) Data | Transactions | Multiple data brokers | Up to one hundred million per day. | Each data broker provides different information. | High | High |
| Manufacturers | Master | Manufacturers, D&B, Neilson | Low | High | Low | Low |
| Distributors | Master | Distributors, D&B, Neilson | Low | High | Low | Low |
| Retailers | Master | Retailers, D&B, Neilson | Low | High | Low | Low |
| Consumer Goods Products | Master | Manufacturers, Neilson | Low | High | Low | Low |
| Retail Locations | Master | Retailers, Neilson | Low | High | Low | Low |

# Data Engineering Methods

Given that most data we receive is in batch and has a great variety, we will load all the data into a data lake. From the data lake, we will process the data into a highly normalized data warehouse. The processing from data lake to data warehouse will:

* Standardize structures
* Integrate data values – for example, retailers, manufacturers, and Neilson have different codes for the same retail locations. This is true for almost all of the data we have.
* Address data quality
* Address duplicates
* Other issues as needed

Even though we will be receiving data in batches, it will be processed as near real time as possible. Within the data lake files will be decomposed into transactions. The transactions will then be processed through to the data warehouse. The goal set is that any data loaded to our internal systems or received directly to our data lake will be processed into the data warehouse in less than three minutes.

The data lake uses Hadoop for storage and multiple technologies for processing. Data can come in through Secure FTP, our defined API, and a Spark installation that feeds data into the Data Lake. Processing within the data lake is primarily Java but we are looking at alternative tools. We anticipate in the future that Data Scientists will access the Data Lake using tools like Python.

The data warehouse is a Microsoft SQL Server database in third normal form. We primarily use C# to move data from the data lake to the data warehouse. Within the data warehouse we primarily use SQL to process the data. Every table in the data warehouse has a single integer primary key that is a surrogate, and nulls are not allowed in any columns. If a column can have three or more well defined values, we will create a categorical table for it.

Reporting and analytics are allowed from the data warehouse, but they are not encouraged, the data marts are the primary reporting and analytics data stores. There are multiple Power BI models built against the data warehouse. Examples are in store inventory, point of sale, product profitability. We seem to be settled on creating a new Power BI model once or so a month.

The last stage is the data marts. Each mart is built to support a specific set of analytical queries. Most of them mimic the ones that we create for the data warehouse. The data mart ones are easier for end users to use and provide significant performance improvements. Each data mart is a snowflake built in Microsoft SQL Server. They are all populated through SQL Stored Procedures. In the future we will make use of Microsoft Analysis Services and Snowflake.

# Application of Concepts to the Problem

A lot of the data we receive is pictures of products on a shelf. It currently takes human intervention to analyze the pictures to identify the shelf level, the number of facings, the number of products, the price label, any shelf tags or promotions, or such. Our brand agents collect this information while taking the photos and our command center reviews it for correctness. We are currently putting in place machine learning analytics to automate this process. Automating in store is complicated because of the variety of devices that brand agents have and lack of internet connectivity in may retail stores.

Automating it after we receive the pictures requires the brand agents to still collect the data and only serves as a validation step. Our first step will be to process the photos and compare to the values provided by the brand agent. Then the command center will only have to review exceptions.

We plan to use the capability of modern phones to execute Machine Learning algorithms on the phone. We are concerned that the models may be too large, but we will be able to use model size reduction techniques. Recent cameras have lidar or lidar like capabilities to produce point clouds. We plan to investigate the use of these to help with product identification and product counting.

# Conclusion

In the two years the company has been in business, we have made significant progress in several areas such as signing up brand agents, clients, and building out a robust mobile application. But we have just started to build out the infrastructure for data, reporting, analytics, machine learning, and AI. While we can continue to build analytics as we have, we are looking to use LLM as the way to make analytics queries. It will mean that the LLM will need to understand out data marts and data warehouse. We have a long way to go and there are many decisions to be made. The knowledge and experience from this class will deeply inform how we move forward to ensure that we are taking advantage of the most modern and proven technologies and approaches.

References

CHEBBI, Imen & Boulila, Wadii & Farah, Imed. (2015). Big Data: Concepts, Challenges and Applications. 9330. 10.1007/978-3-319-24306-1\_62.

**IBM.** (2024, May). Why data governance is essential for enterprise AI. [IBM Blog]. https://www.ibm.com/blog/why-data-governance-is-essential-for-enterprise-ai/

**Google Cloud.** (2024, May 26). MLOps: Continuous delivery and automation pipelines in machine learning. [cloud.google.com]. Retrieved from <https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>

Gómez-Caicedo, M. I., Gaitán-Angulo, M., Bacca-Acosta, J., Briñez Torres, C. Y., & Cubillos Díaz, J. (2022). Business analytics approach to artificial intelligence. *Frontiers in Artificial Intelligence*, 5, 974180. <https://www.frontiersin.org/articles/10.3389/frai.2022.974180/full>

Jenkins, T. (2004). The First Language - A Case for Python? *Innovation in Teaching and Learning in Information and Computer Sciences*, *3*(2), 1–9. doi:10.11120/ital.2004.03020004

Mui, Michael & Holler, Anne. (2019). Evolving Michelangelo Model Representation for Flexibility at Scale.

Zhang, Y., Sun, Y., & Liu, X. (2022, May 10). Rapid regression detection in software deployments through sequential testing [arXiv preprint arXiv:2205.14762]. <https://arxiv.org/abs/2205.14762>