LinkedIn Laura Sanchez

AN APPROACH TO A CLOUD ARCHITECTURE DESIGN FOR FRESHDEAL!

LAURA MARGARITA SÁNCHEZ PADILLA

CANADA

2025

TABLE OF CONTENT

ABSTRACT

This paper presents the design and implementation strategy for a cloud-based data architecture that supports Fresh Deal, a digital marketplace connecting agricultural producers directly with buyers. The goal is to optimize the agricultural supply chain through scalable, reliable, and cost-efficient data engineering practices. The architecture integrates both batch and streaming ingestion pipelines to process diverse datasets, including e-commerce transactions, product listings, geolocation, web tracking, and customer records. Data is organized into a Lakehouse structure, with Bronze, Silver, and Gold layers ensuring traceability, quality, and analytics readiness. This design demonstrates how cloud tools can transform raw agricultural data into actionable intelligence, empowering decision-making, improving transparency, and supporting farm-to-table delivery models.

1. INTRODUCTION

This article explores the motivation, data strategy, and architecture behind Fresh Deal, showing how thoughtful data engineering can transform the agricultural supply chain into a more equitable and efficient ecosystem.

The process to create this article was divided into several parts. First, it was important to understand the business being created so we could identify the data sources and sinks. In other words: what data the cloud architecture will process, where it resides, its nature, the best way to process it, and how it will be used once processed.

After that, the architecture design phase began. This is when we decide which tools are best to move data from its raw form to the final form the company needs for insights and forecasting. In this stage, we selected tools from the Microsoft Azure cloud ecosystem.

Next, it was necessary to design the pipelines that will be used across the architecture. At this point it’s crucial to choose approaches that minimize cost, run efficiently, and reduce the risk of errors and duplicated data. Once a robust pipeline design is in place, we plan for failure scenarios. For this project we chose a retry with exponential backoff strategy because it allows pipelines time to recover without overwhelming the system.

By following this structured approach, Fresh Deal demonstrates how data engineering can be applied strategically to solve real-world challenges.

2. ABOUT FRESHDEAL!

Fresh Deal! is a centralized online marketplace where agricultural sellers post their products, and buyers can browse and purchase them. The platform also includes a third-party logistics system that handles product delivery after purchase confirmation, making the entire process seamless and efficient.

In an increasingly data-driven world, the agriculture sector still faces challenges in transparency, accessibility, and efficiency. Small and medium-scale farmers often struggle to reach end consumers, relying heavily on intermediaries that reduce their profits and delay the distribution of fresh goods. At the same time, buyers demand more traceable and efficient supply chains.

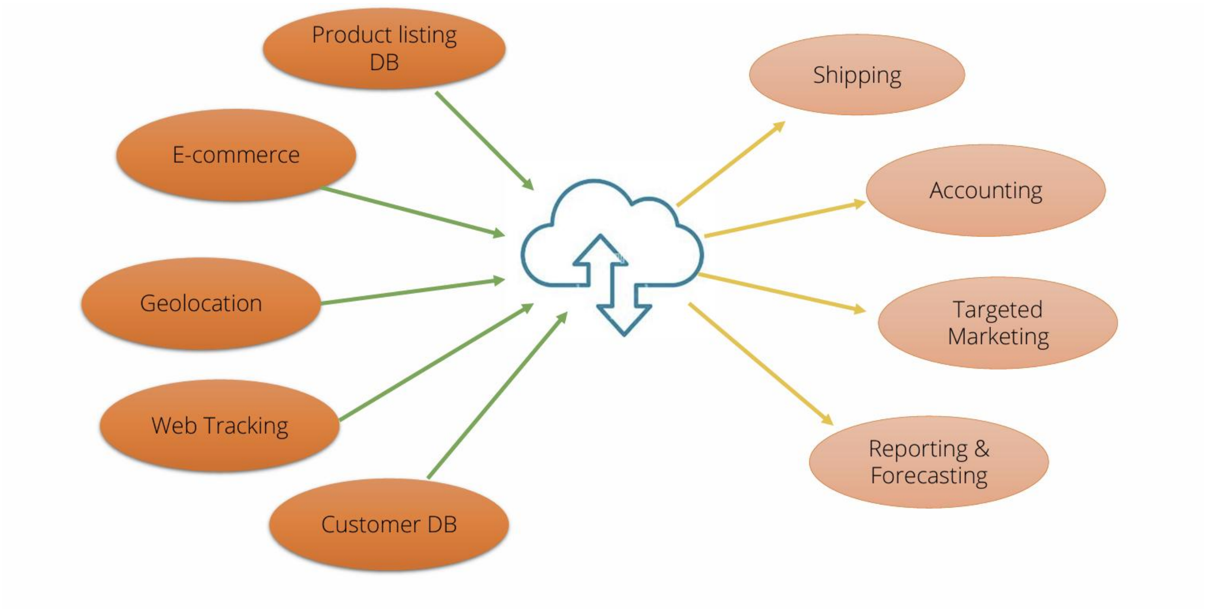
Fresh Deal! emerges as a solution to this disconnection. A digital platform designed to connect agricultural producers directly with buyers through a seamless, technology-driven marketplace. By combining e-commerce functionality, real-time logistics, and intelligent data pipelines, Fresh Deal empowers farmers, enhances food freshness, and simplifies the farm-to-table journey.

3. MISSION AND OBJECTIVES

Build a scalable and reliable data architecture to optimize the sales process and empower farmers to reach end users directly

- Enable direct farm-to-table deliveries to preserve product freshness.
- Enhance transparency and traceability of agricultural goods.
- Support data-driven decision-making through advanced data collection and analytics.

4. DATA DIAGRAM



4.1. Data Sources

After running Discovery and brainstorm sessions, five datasets were considered to be ingested.

- **Product listing database:** Contains the core information of the suppliers or sellers on the product they want to sell. It is structured and will be ingested as batch. It will be useful for product catalog, inventory tracking, pricing analysis.
- **E-commerce data:** Records of events such as purchases, order fulfillments, payments. It is structured and will be ingested as streaming. This data is useful for accounting, sales analysis and revenue reporting.
- **Geolocation data:** Information on customers' location. Its nature is structured and semistructured and will be ingested as batch. Knowing customers' location help to process the shippings effectively.
- **Web tracking data:** Data on user's behaviour on the web such as page views, clicks, sesión duration, and navigation paths. Its nature is semistructured and the ingestión type is batch. It is useful to dentify user behavior analytics.
- **Customer database:** Contains customer information such as identity, contact details, preference, purchase history and behaviour. It's structured and semistructured. The Ingestion type is batch. It will be useful for marketing (segmentation, retention and personalization)

4.2. Data Sink

The prupose of processing this data through a cloud architecture is to be able to find different insights for the bussines. In that order aggregated data will go to dashboards, machine learning and will be available for accounting, shipping and marketing teams.

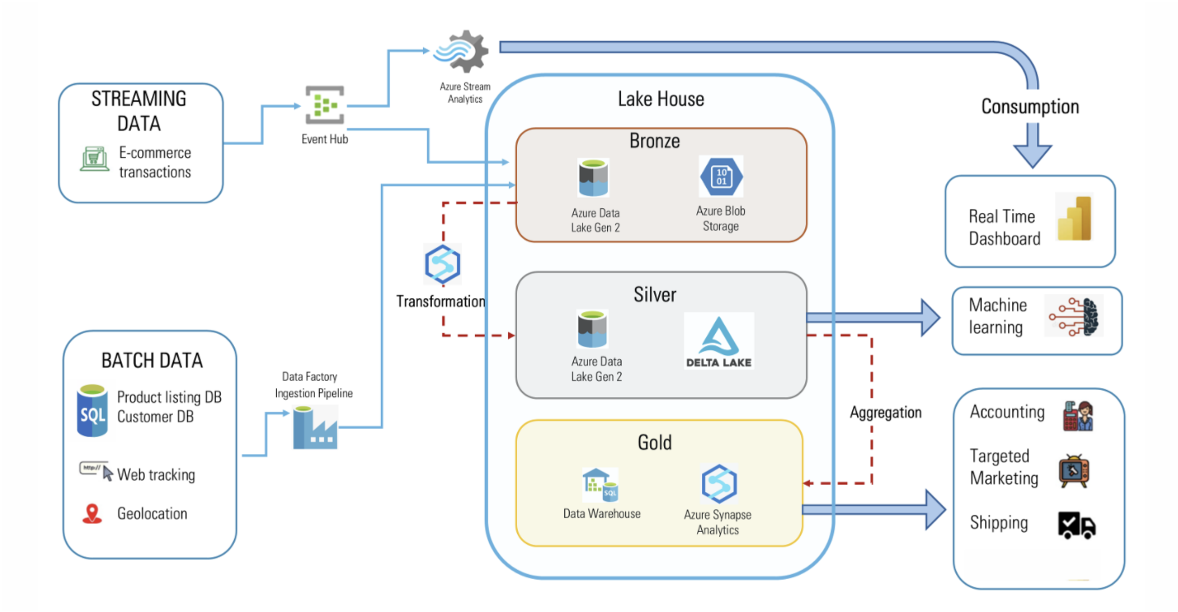
- **Shipping:** Processed order fulfillment data is delivered to manage logistics (assigning carriers, package tracking, and customer notification). Its purpose is to enable timely and accurate order fulfillment.
- **Accounting:** financial, transactional and operational data is processed for financial reporting, auditing and compliance. It helps to drive profitability analysis, cost control and forecasting.
- **Targeted marketing:** process of delivering personalized campaigns, ads, and content to specific customer segments based on their behavior, preferences, location, and purchase history. It will have a big impact in improving customers engagement.
- **Real time dashboard and forecasting:** Visual interface that presents key metrics, trends, and insights in real time or on demand. It will rovide real-time visibility into KPIs and trends to make data-driven decisions.

5. CLOUD ARCHITECTURE DIAGRAM

In the data architecture designed for this project, the goal was to implement the most suitable tools to process each dataset depending on its size, nature, and purpose.

To begin with, on the left side of the architecture are the data sources to be ingested. These were divided into two categories. For e-commerce transaction data, ingestion was done as streaming data because the aim was to obtain continuous, real-time updates. This data was collected through Azure Event Hubs and then processed in Azure Stream Analytics to create a more structured database, which in turn enabled the development of a real-time dashboard with live data and updates.

On the other hand, datasets ingested in batch mode were processed through Azure Data Factory.



5.1. Bronze Layer

After the process described earlier, the data enters the first layer of the lakehouse architecture: the Bronze Layer.

In this stage, the data is stored in its raw, unprocessed form, exactly as it was ingested from the source systems. This layer serves as a secure, immutable record of the original data, ensuring that no information is lost or altered before transformation.

We use Azure Data Lake Storage Gen2 as the main storage system because it is highly scalable, secure, and integrates seamlessly with other Azure services. For very large datasets, such as web tracking logs or clickstream data, we leverage Azure Blob Storage to optimize cost efficiency, since Blob Storage is well-suited for massive, infrequently accessed files.

The Bronze Layer ensures data traceability, compliance, and reproducibility for future analysis or reprocessing.

5.2. Silver Layer

From the Bronze Layer, the raw data undergoes cleaning, transformation, and enrichment using Azure Synapse Analytics. Once curated, it is stored in the Silver Layer, which contains validated, structured, and enriched datasets ready for advanced analytics.

We store this curated data in Azure Data Lake Storage Gen2 integrated with Delta Lake. Delta Lake provides ACID transactions, schema enforcement, and the ability to perform time travel queries (retrieve previous versions of the data), which is extremely valuable for auditability and reproducibility.

This layer is ideal for teams such as Machine Learning or Advanced Analytics to build models, run experiments, and generate predictions.

5.3. Gold Layer

Once the data is fully curated, it moves to the aggregation and presentation phase within Azure Synapse Analytics.

Here, we join, aggregate, and calculate KPIs to create business-ready datasets that are optimized for reporting. This processed and highly structured data is stored in a dedicated Data Warehouse, which supports fast query performance for large volumes of historical data.

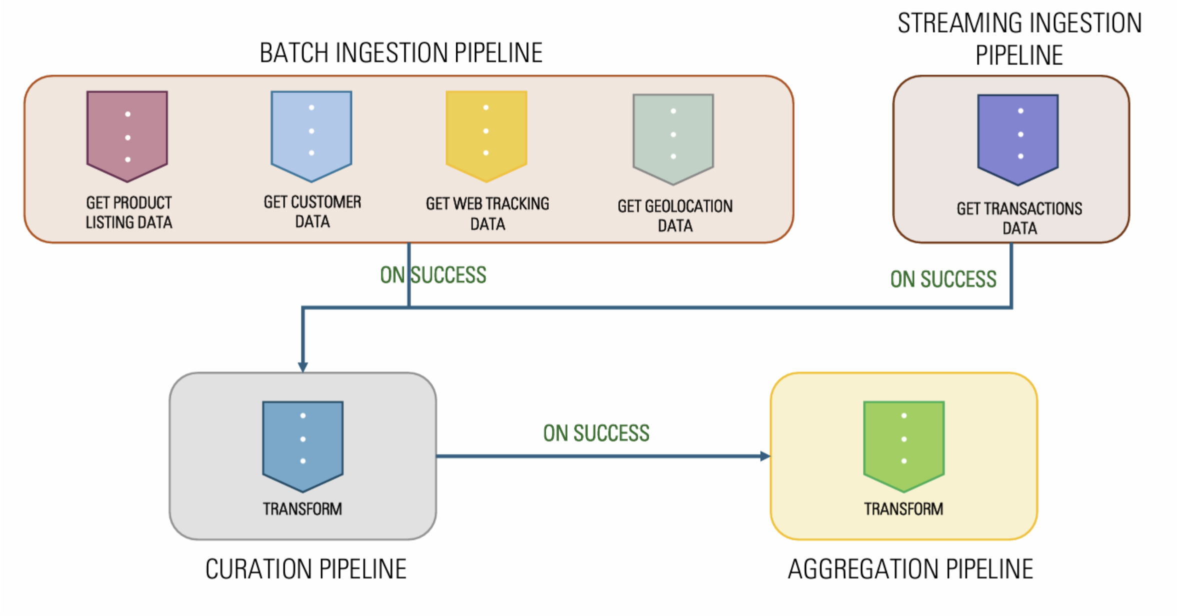
This layer enables decision-makers and analysts to access ready-to-use datasets without needing to perform complex transformations themselves.

5.4. Consumption

Finally, data from the Gold Layer is made available for visualization, business intelligence, and operational reporting. This is done using Power BI, which connects directly to the Data Warehouse or curated datasets in the lakehouse.

Different teams, such as marketing, operations, finance, and executive leadership can access dashboards tailored to their specific needs. Each team sees only the data relevant to their work, ensuring security and compliance.

6. PIPELINE DESIGN

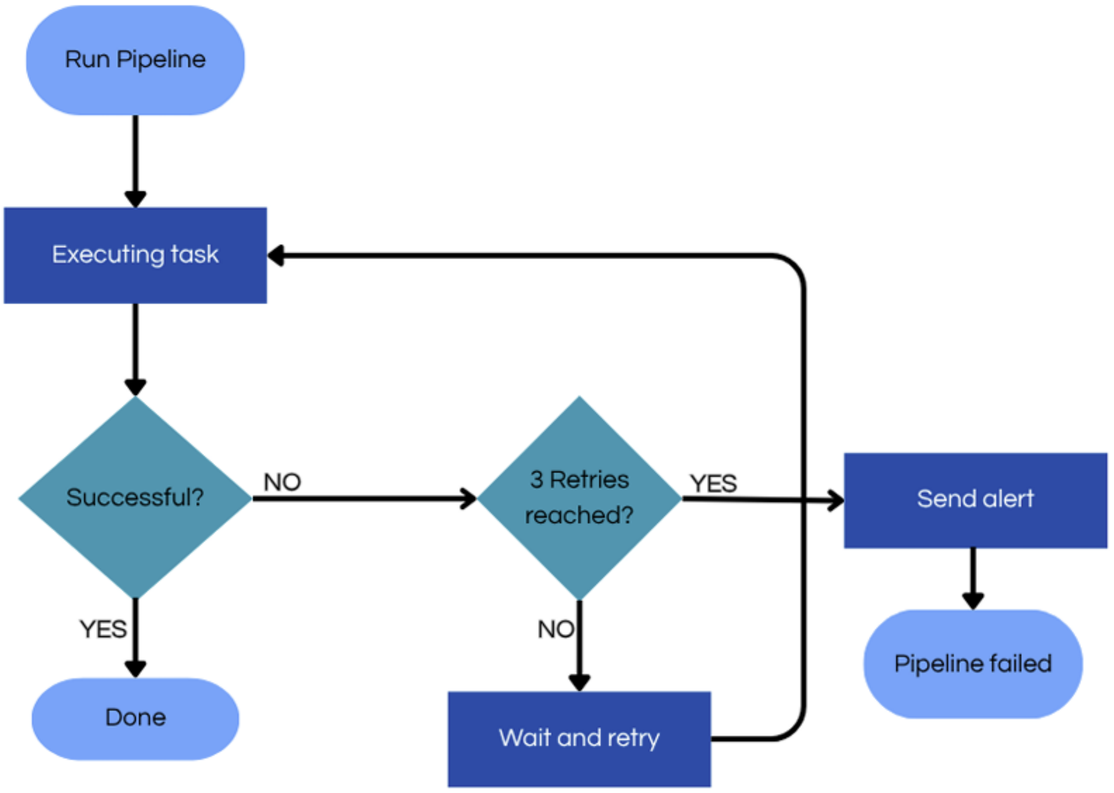


The design is a master pipeline. It separates ingestion (how data gets into the system) from curation (cleaning/enriching) and aggregation (creating analytics-ready datasets).

- **Batch Ingestion Pipeline** collects large or periodic datasets (product listings, customer records, web tracking snapshots, geolocation snapshots).
- **Streaming Ingestion Pipeline** handles real-time event streams (e-commerce transactions).
- Both ingestion paths feed a **Curation Pipeline** that transforms raw data into cleaned, standardized, versioned data (Silver).
- A final **Aggregation Pipeline** produces aggregated, business-ready tables (Gold) for reporting and BI.

Every pipeline will run after the previous one runned successfully. This separation improves scalability, traceability and and saving costs.

6.1. Pipeline Failure Strategy



Even with a Good pipeline disign, failures can still happen. For this reason is important to create a pipeline failure strategy. In this Project the strategy chosen was retry logic with exponential back off. In other words it means that in case something goes wrong it is recommended to retry to run the pipeline 3 times, the first time it will take 30 seconds wait, second time 1 minute and the third time 2 minutes. If after this three times the pipeline still not running successfuly, an alert will be send to the admin or data engeneering team

7. CONCLUSIONS

The proposed cloud architecture for Fresh Deal successfully addresses the complexity and diversity of data in the agricultural supply chain by combining real-time and batch processing capabilities. The Lakehouse model ensures that data is captured in its raw form for compliance, curated for advanced analytics, and aggregated for fast, business-ready consumption. Microsoft Azure's ecosystem offers a seamless integration of tools that optimize both performance and cost, while maintaining flexibility to handle varying data volumes and structures.

By separating ingestion, curation, and aggregation processes, the design achieves scalability, reduces operational risk, and ensures that each stage of the data lifecycle is optimized for its specific purpose. The inclusion of a pipeline failure recovery mechanism enhances system reliability, while the integration with Power BI democratizes data access across teams.

Ultimately, this architecture empowers Fresh Deal to deliver on its mission: connecting farmers directly to consumers, improving transparency, and enabling data-driven decision-making that benefits all stakeholders in the supply chain. The model can be adapted to similar marketplace platforms seeking to modernize their data infrastructure and leverage analytics for competitive advantage.