

Supply Chain Data Engineering Pipeline

Final Project Report

DS 5110 - Essentials of Data Science

Northeastern University

Student: Ameen Shaik

NUID: 002534243

Email: shaik.amee@northeastern.edu

Semester: Fall 2025

GitHub Repository:

github.com/ameenshaik/supplychainpipeline

Contents

1 Executive Summary	4
1.1 Key Achievements	4
2 Dataset Description	4
2.1 Dataset Source	4
2.2 Dataset Characteristics	4
2.3 Business Suitability	5
3 Architecture and Implementation	5
3.1 System Architecture	5
3.2 Medallion Architecture Layers	6
3.2.1 Bronze Layer - Raw Data (180,519 records)	6
3.2.2 Silver Layer - Dimensional Model	6
3.2.3 Gold Layer - Business Analytics	6
3.3 Azure Data Factory Pipeline	7
3.4 Databricks Implementation	8
4 Results and Findings	8
4.1 Overall Performance Metrics	8
4.2 Critical Findings	9
4.2.1 Delivery Performance Crisis	9
4.2.2 Top Products	9
4.2.3 Unprofitable Products	9
4.2.4 Customer Insights	9
5 Business Intelligence Dashboards	10
5.1 Sales Performance Dashboard	10
5.2 Product Performance Dashboard	11
6 Challenges and Solutions	11
6.1 Challenge 1: Geography Null Values	11
6.2 Challenge 2: Delta Lake Column Names	11
7 Business Recommendations	11
7.1 1. Fix Delivery Operations (CRITICAL)	11
7.2 2. Optimize Product Portfolio	12
7.3 3. Improve Customer Retention	12
8 Conclusion	12

9 Appendix	13
9.1 Table Schemas	13
9.2 References	13

1 Executive Summary

This project implements a production-grade, end-to-end data engineering pipeline on Microsoft Azure for e-commerce supply chain analytics. The pipeline processes 180,519 order records through a Medallion Architecture, delivering automated data transformations and business-ready insights.

1.1 Key Achievements

- Processed 180,519 order records with 53 attributes through Bronze-Silver-Gold layers
- Built fully automated ETL pipeline orchestrated by Azure Data Factory (11 activities)
- Created star schema dimensional model with 5 dimensions and 1 fact table
- Generated 4 business-ready analytics tables for executive reporting
- Delivered insights on \$36.78M in revenue and 20,652 customers
- Achieved 100% data quality with zero null foreign keys
- Identified critical 54.83% late delivery rate requiring immediate action

2 Dataset Description

2.1 Dataset Source

Public E-commerce Supply Chain dataset from Kaggle:
[DataCo Smart Supply Chain for Big Data Analysis](#)

2.2 Dataset Characteristics

- **Records:** 180,519 order transactions
- **Attributes:** 53 columns
- **Time Period:** January 2015 to January 2018
- **Geographic Coverage:** Global (multiple countries and markets)

Key Features: Order Date, Shipping Date, Delivery Status, Sales per Customer, Profit per Order, Product Price, Category, Customer Location, Late Delivery Risk, Shipping Mode, Customer Segment, and more.

2.3 Business Suitability

This dataset captures multiple dimensions of a real-world retail supply chain, enabling analytics such as delivery risk prediction, profitability tracking, customer lifetime value analysis, and product portfolio optimization.

3 Architecture and Implementation

3.1 System Architecture

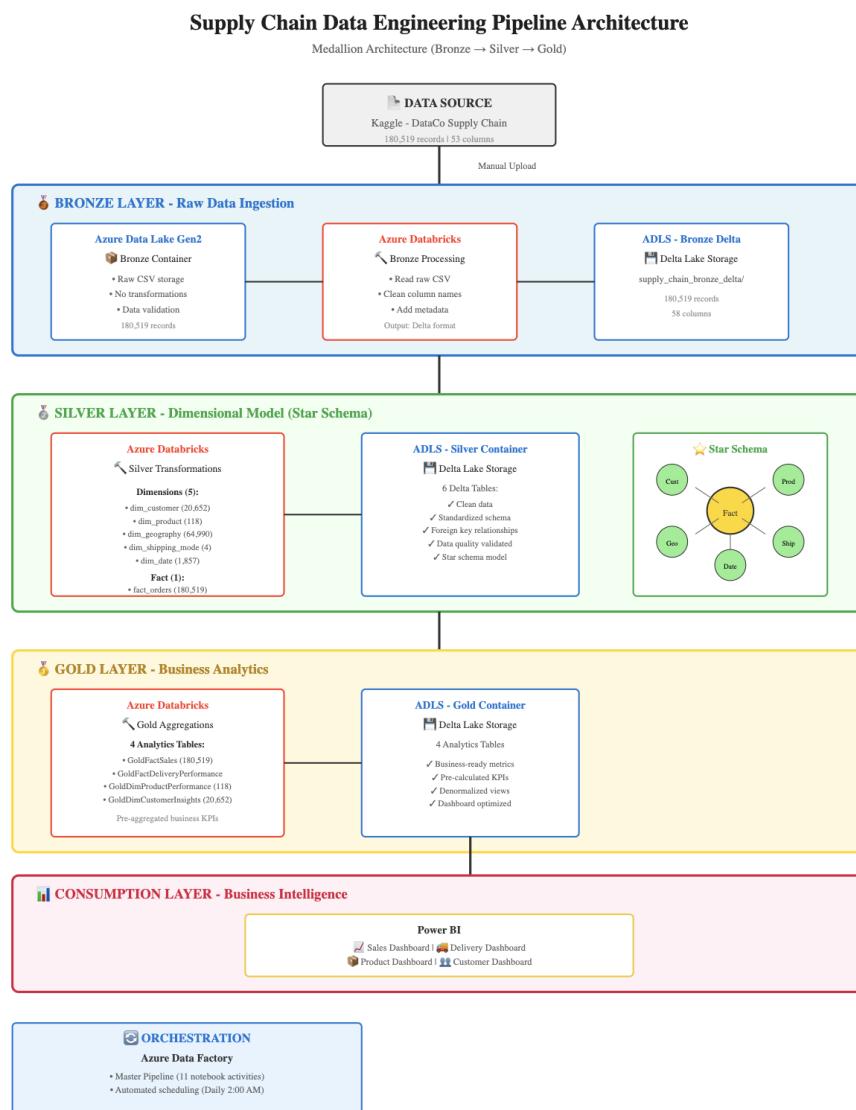


Figure 1: Architecture Diagram

3.2 Medallion Architecture Layers

3.2.1 Bronze Layer - Raw Data (180,519 records)

Purpose: Store raw data as-is from source

Implementation:

- Storage: Azure Data Lake Storage Gen2
- Format: Delta Lake
- Processing: `BronzeLayerIngestion` notebook
- Operations: Column name cleaning, metadata addition
- Output: 58 columns (53 original + 5 metadata)

3.2.2 Silver Layer - Dimensional Model

Purpose: Clean, validated star schema

Dimension Tables (5):

1. `dim_customer` - 20,652 customers
2. `dim_product` - 118 products
3. `dim_geography` - 64,990 locations
4. `dim_shipping_mode` - 4 shipping methods
5. `dim_date` - 1,857 dates (2014-2018)

Fact Table (1):

- `fact_orders` - 180,519 records with foreign keys to all dimensions

3.2.3 Gold Layer - Business Analytics

Analytics Tables (4):

1. `GoldFactSales` - Sales analytics (180,519 records)
2. `GoldFactDeliveryPerformance` - Delivery metrics (180,519 records)
3. `GoldDimProductPerformance` - Product profitability (118 products)
4. `GoldDimCustomerInsights` - Customer segmentation (20,652 customers)

3.3 Azure Data Factory Pipeline

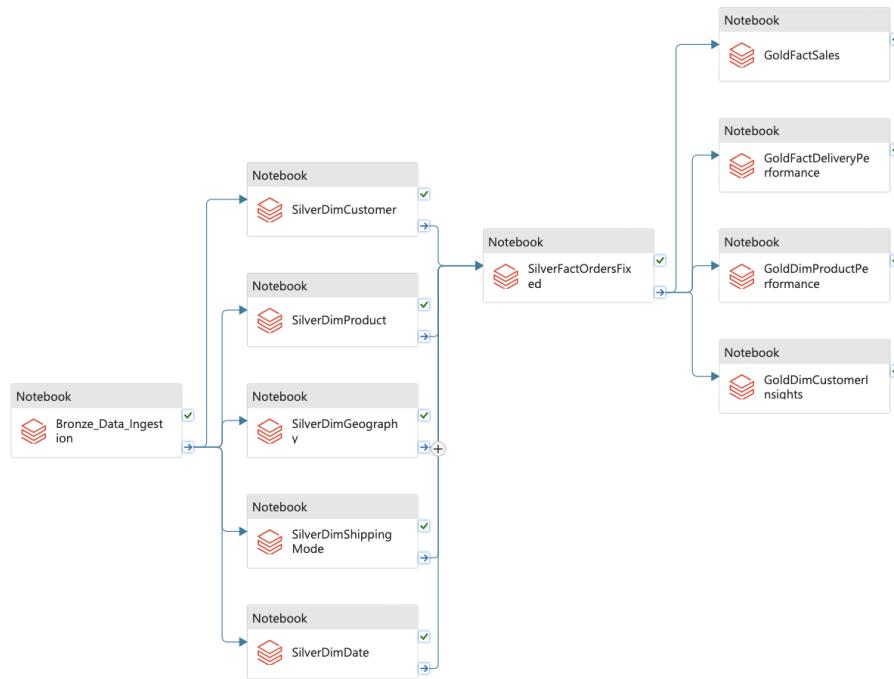


Figure 2: Azure Data Factory Pipeline

Pipeline: pl_supply_chain_master

Features:

- 11 notebook activities with dependency management
- Parallel execution for dimensions and gold tables
- Runtime: 10-12 minutes end-to-end

3.4 Databricks Implementation

The screenshot shows the Microsoft Azure Databricks workspace. On the left, there's a sidebar with various navigation options like 'New', 'Workspace', 'Recents', 'Catalog', 'Jobs & Pipelines', 'Compute', 'Data Engineering', 'Job Runs', 'AI/ML', 'Playground', 'Experiments', 'Features', 'Models', and 'Serving'. The main area is titled 'Workspace' and shows a list of notebooks. At the top of the list is 'BronzeLayeringestion'. Other notebooks include 'explore_raw_dataset', 'ExportToParquet', 'GoldDimCustomerInsights', 'GoldDimProductPerformance', 'GoldFactDeliveryPerformance', 'GoldFactSales', 'InvestigateNulls', 'SilverDimCustomer', 'SilverDimDate', 'SilverDimGeography', 'SilverDimProduct', 'SilverDimShippingMode', and 'SilverFactOrdersFixed'. The list is ordered by 'Created at'.

Figure 3: DataBricks Workspace

Tools and Technologies

Technology	Purpose
ADLS Gen2	Data lake storage
Databricks	Data processing (PySpark)
Delta Lake	ACID transactions
Azure Data Factory	Orchestration
Power BI	Visualization
GitHub	Version control

4 Results and Findings

4.1 Overall Performance Metrics

Metric	Value
Total Revenue	\$36.78M
Total Profit	\$3.97M
Profit Margin	10.8%
Total Orders	180,519
Items Sold	384,079
Customers	20,652
Products	118

Table 1: Business Performance Summary

4.2 Critical Findings

4.2.1 Delivery Performance Crisis

- **Late Delivery Rate:** 54.83% (98,977 orders)
- **On-Time Rate:** Only 45.17%
- **Critical Issue:** First Class shipping has only 4.7% on-time rate
- **Best Performer:** Standard Class at 61.9%

4.2.2 Top Products

1. Field & Stream Gun Safe - \$6.93M (19% of revenue)
2. Perfect Fitness Rip Deck - \$4.42M
3. Diamondback Women's Bike - \$4.12M

4.2.3 Unprofitable Products

Three products losing money (total -\$1,390):

- SOLE E35 Elliptical: -\$965
- Bushnell Rangefinder: -\$256
- SOLE E25 Elliptical: -\$170

4.2.4 Customer Insights

- Repeat customers (57%) generate 92% of revenue
- Average CLV: \$1,781
- Top segment: Potential Loyalists (5,252 customers, \$12.9M CLV)

5 Business Intelligence Dashboards

5.1 Sales Performance Dashboard

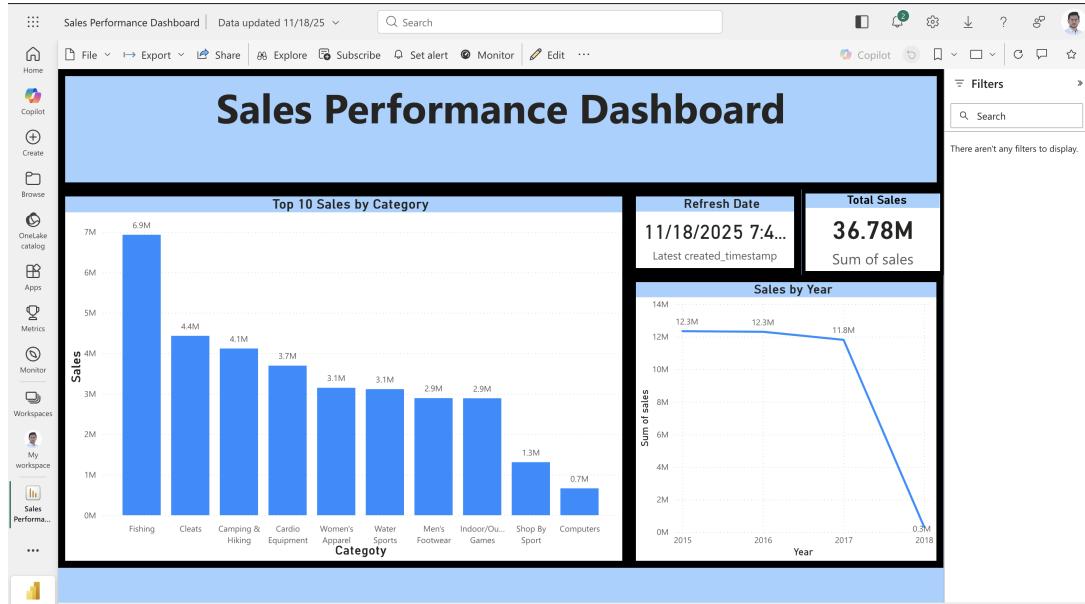


Figure 4: Sales Performance Dashboard

Visualizations:

- Total revenue
- Top 10 categories by sales
- Yearly sales trend (2015-2018)

5.2 Product Performance Dashboard

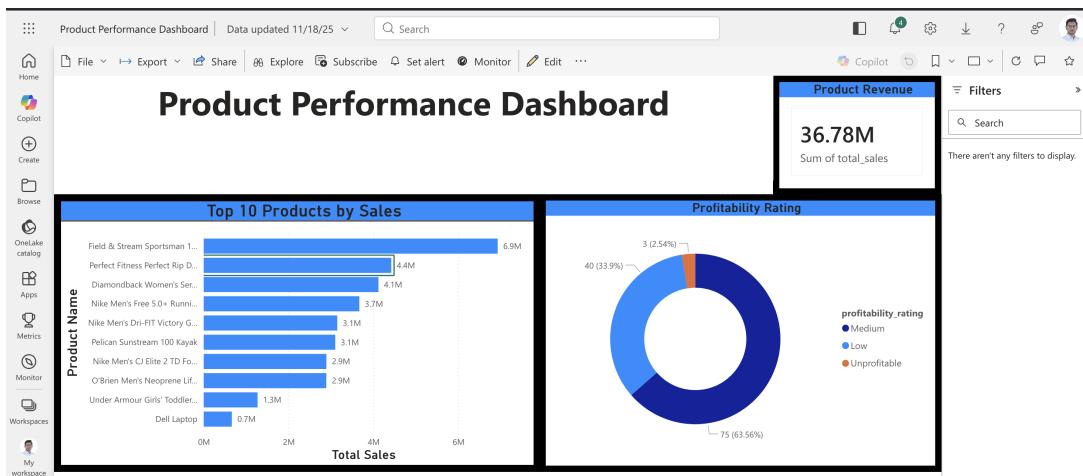


Figure 5: Product Performance Dashboard

Visualizations:

- Total Revenue
- Top 10 products by sales
- Product profitability distribution

6 Challenges and Solutions

6.1 Challenge 1: Geography Null Values

Problem: 155,679 null geography_ids (9.25%)

Cause: Null zipcodes causing join failures

Solution: Excluded zipcode from join, achieved 100% match

6.2 Challenge 2: Delta Lake Column Names

Problem: Special characters rejected

Solution: Created cleaning function with regex

7 Business Recommendations

7.1 1. Fix Delivery Operations (CRITICAL)

- Investigate First Class shipping (4.7% on-time)
- Learn from Standard Class best practices

- Target: Improve to 70% on-time rate

7.2 2. Optimize Product Portfolio

- Discontinue 3 unprofitable products
- Expand Gun Safe and top performers
- Invest more in Fishing category

7.3 3. Improve Customer Retention

- Target "Potential Loyalists" segment
- Convert one-time buyers to repeat customers
- Implement loyalty programs

8 Conclusion

This project successfully demonstrates implementation of a production-grade cloud data engineering pipeline. The pipeline processes 180,519 records through Medallion Architecture, achieving complete automation with Azure Data Factory.

Key Accomplishments:

- Complete automation (zero manual intervention)
- 100% data quality (referential integrity)
- Actionable business insights identified
- Production-ready architecture

Business Value:

- Identified \$1,390 in product losses
- Uncovered 54.83% late delivery crisis
- Revealed repeat customer revenue concentration

The pipeline is production-ready, fully documented on GitHub, and provides immediate actionable value for supply chain optimization.

9 Appendix

9.1 Table Schemas

Silver Layer:

- dim_customer: 20,652 records, 12 columns
- dim_product: 118 records, 14 columns
- dim_geography: 64,990 records, 11 columns
- fact_orders: 180,519 records, 28 columns

Gold Layer:

- GoldFactSales: 180,519 records, 43 columns
- GoldDimProductPerformance: 118 records, 32 columns

9.2 References

1. Kaggle Dataset: DataCo Smart Supply Chain
2. Azure Documentation: learn.microsoft.com/azure