GLOBAL SUPERSTORE DATA WAREHOUSE PROJECT REPORT

Course: IST 722 – Data Warehousing
Instructor: Prof. Joseph Kinn
Team Alpha:

- Sakshi Singh
- Shivani Pandeti
- Amulya Walimbe
- Reeya Patra

Date: 1 December 2025

Executive Summary

This project presents the design and implementation of a complete **enterprise-grade Data Warehouse and Business Intelligence (BI) system** built on the Global Superstore dataset.

The goal was to transform inconsistent raw CSV data into a clean, structured analytics platform using **Snowflake** for ETL and **Power BI** for visualization.

The project follows a **three-layer data architecture** - RAW_STAGE, TRANSFORM, and DW, supporting a scalable and maintainable ETL pipeline.

A Kimball-based **star schema** was developed with conformed dimensions for Date, Customer, Product, Geography, and Shipping, and four fact tables covering Sales, Shipping, Customer Monthly, and Product Monthly analytics.

Four interactive Power BI dashboards deliver insights into **Sales performance**, **Shipping efficiency**, **Customer behavior**, and **Product profitability**, enabling stakeholders to evaluate revenue patterns, delivery delays, market trends, and operational bottlenecks.

The final solution provides an accurate, fast, and stable analytical foundation that supports informed decision-making and prepares the organization for future expansion into predictive analytics and automated reporting.

1. Introduction

Modern retail organizations generate large volumes of transactional data but often lack the structured infrastructure needed for reliable analysis. Raw data stored in flat files is difficult to query, inconsistent across attributes, and unsuitable for multi-dimensional reporting.

This project addresses these limitations by designing a full **Data Warehouse (DW)** for the Global Superstore dataset, enabling high-quality analytics across key business areas.
The solution includes:

- A structured Snowflake ETL pipeline
- A star schema data model
- Fact and dimension tables for analytics
- Power BI dashboards for business reporting

The outcome is a robust system capable of analyzing trends in revenue, customer behavior, logistics performance, and product profitability.

2. Problem Statement

Before building the warehouse, the organization faced several operational issues:

- Raw CSV files with inconsistent formatting, duplicated geographic entries, irregular date values, and unstandardized product metadata.
- No centralized repository of cleansed historical data.
- Reporting processes that were manual, repetitive, and error prone.
- Limited visibility into customer retention, shipping delays, and category profitability.
- No data model to support drill-down analytics or cross-filtering.

The data warehouse was built to solve these pain points by providing a unified, validated, and analytics-ready foundation.

3. Dataset Description

The Global Superstore dataset consists of **51,000+ rows and 24 fields**, containing order-level details across multiple countries and product groups.

Key attributes

- Order & Ship Dates
- Customer ID, Name, Segment
- Product ID, Category, Sub-Category
- Sales, Profit, Quantity, Discount
- Shipping Cost, Ship Mode
- City, State, Country, Region

Data Quality Observations

The raw CSV data was **not analysis-ready**:

- Text fields had inconsistent casing (city, region, segment)
- Duplicate geography combinations existed
- Uneven date formatting
- Whitespace and inconsistent spacing
- No unique surrogate keys
- Profit and discount values required validation

These issues required careful preprocessing in the ETL pipeline.

4. Data Architecture & Design

We implemented a **three-layer Snowflake architecture**, which ensures scalability, clean governance, and modularity.

4.1 Architecture Diagram

RAW_STAGE Layer

- Stores raw CSV exactly as imported
- Zero transformations
- Ensures immutability and traceability

TRANSFORM Layer

- Performs cleaning, standardization, key generation
- Applies business logic (profit margin, days-to-ship, AOV)
- Builds dimension and fact tables

DW Layer (Presentation Layer)

- Final optimized tables used by Power BI
- Contains conformed dimensions and analytics-ready facts

This architecture separates ingestion, transformation, and consumption, making debugging and scaling easier.

5. Schema Design & Modeling

5.1 Star Schema Design

The warehouse follows **Kimball's dimensional modeling principles**, with conformed dimensions shared across facts.

Dimensions

- **DIM_DATE**
- **DIM_CUSTOMER**
- **DIM_PRODUCT**
- **DIM_GEOGRAPHY**
- **DIM_SHIPPING**

Facts

- **FACT_SALES** (Order Line Level)
- **FACT_SHIPPING** (Shipment Level)
- **FACT_CUSTOMER_MONTHLY** (Customer × Month × Region)
- **FACT_PRODUCT_MONTHLY** (Product × Month × Region)

5.2 Grain Definitions

- Sales: *One row per order × product line*
- Shipping: *One row per shipment*
- Monthly tables: *One row per Customer/Product × Month × Region*

5.3 Surrogate Keys

- MD5-based keys for Geography and Shipping
- Numeric YYYYMMDD keys for Date Dimension
- Conformed dimensions ensure cross-dashboard consistency

This schema enables fast querying, and governance ready data structure.

6. ETL Pipeline Implementation

6.1 Extract

- Loaded CSV into Snowflake internal stage
- Created RAW_STAGE.STG_GLOBAL_SUPERSTORE
- Validated row counts (~51,290 rows)
- Ensured datatype alignment

6.2 Transform

Data Cleaning

- Trimmed whitespace
- Standardized text casing
- Corrected inconsistent region/segment labels
- Re-validated dates

Business Rules

- **Days-to-Ship:** DATEDIFF (OrderDate, ShipDate)
- **Profit Margin:** Profit / Sales
- **Late vs On-Time:** Days-to-Ship > 5
- **AOV:** Avg (Sales per Order)

Key Creation

- Surrogate Keys using MD5 hashing
- Monthly aggregation keys (YYYYMM)

Aggregation Logic

Created monthly facts for:

- Customer-level revenue & orders
- Product-level revenue, profit, and quantity

6.3 Load

- Loaded final DIM and FACT tables into DW schema
- Conducted row count reconciliation
- Verified referential integrity
- Connected DW directly to Power BI

7. Detailed Fact & Dimension Tables

7.1 Dimension Tables

DIM_DATE

- DateKey, FullDate, Year, Month, Quarter, Day, DayOfWeek

DIM_CUSTOMER

- CustomerKey, CustomerID, Name, Segment

DIM_PRODUCT

- ProductKey, ProductID, ProductName, Category, Sub-Category

DIM_GEOGRAPHY

- GeoKey, City, State, Region, Country

DIM_SHIPPING

- ShipKey, ShipMode, Shipping Priority

7.2 Fact Tables

FACT_SALES

- OrderDateKey, CustomerKey, ProductKey, GeoKey, ShipKey
- SalesAmount, ProfitAmount, Quantity, Discount

FACT_SHIPPING

- OrderDateKey, ShipDateKey, CustomerKey
- DaysToShip, ShippingCost, LateFlag

FACT_CUSTOMER_MONTHLY

- MonthKey, CustomerKey, GeoKey
- TotalSales, TotalProfit, OrderCount, ProfitMargin

FACT_PRODUCT_MONTHLY

- MonthKey, ProductKey, GeoKey
- MonthlySales, MonthlyProfit, Quantity, Margin

8. Data Validation

Validation Activities

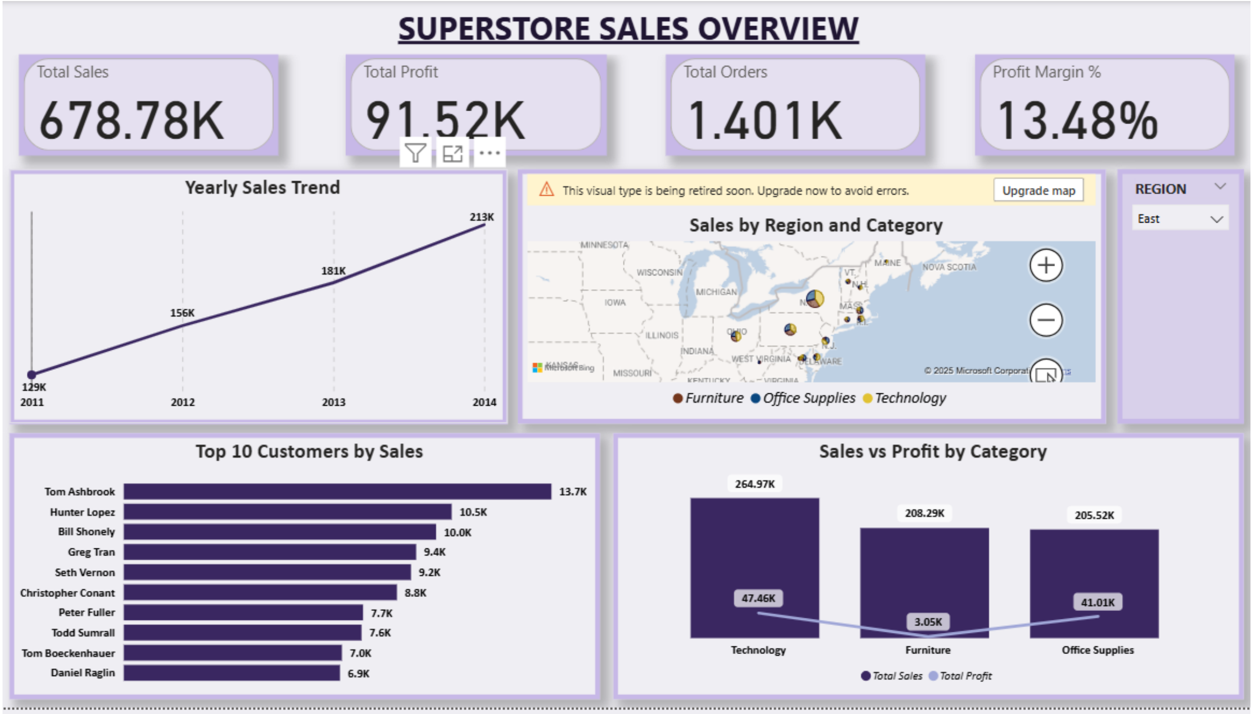
- Row count validation
- Duplicate detection in customer & geography fields
- Null handling for date & key fields
- Checking dimension-fact key alignment
- Outlier checks for Profit, Discount, Shipping Cost
- SLA validation (Late vs On-Time)

These steps ensured correctness and trustability of reporting data.

9. Business Intelligence Layer (Power BI)

Four dashboards were designed for stakeholders.

9.1 Sales Performance Dashboard



SUPERSTORE SALES OVERVIEW

Total Sales 678.78K

Total Profit 91.52K

Total Orders 1.401K

Profit Margin % 13.48%

⚠ This visual type is being retired soon. Upgrade now to avoid errors. Upgrade map

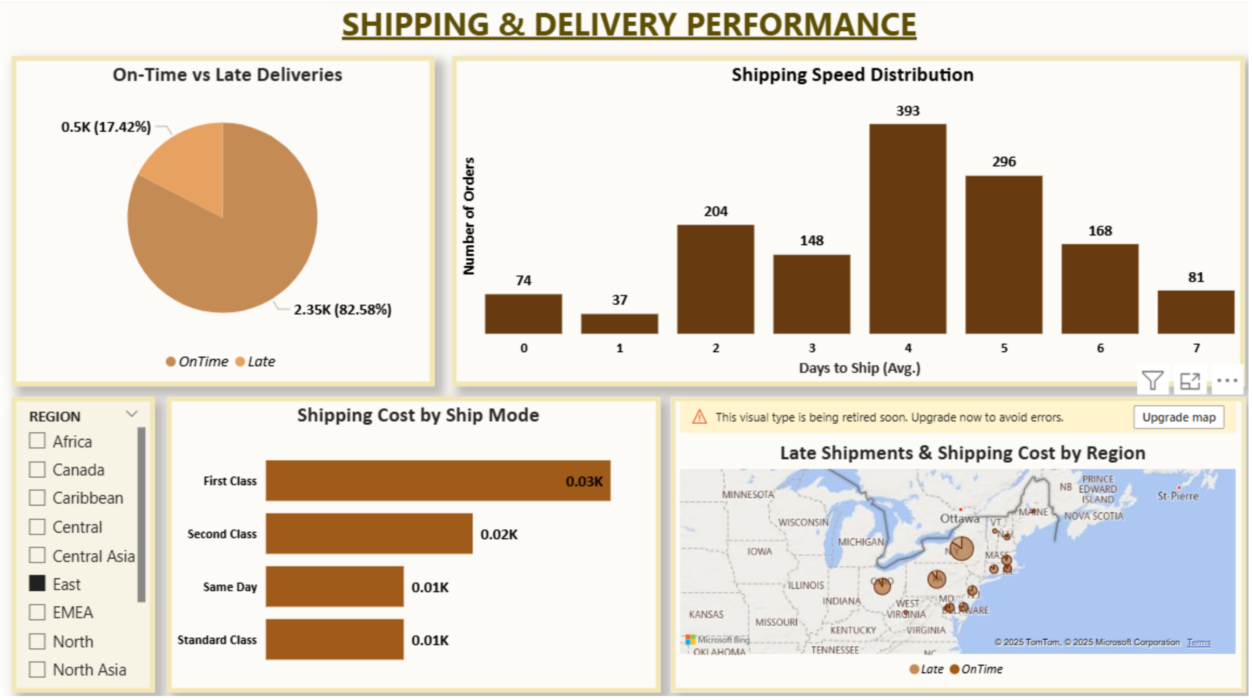
REGION: East

- Revenue, Profit, Profit Margin, Order Count
- YOY Sales Trend
- Sales by Category, Region
- Top 10 Customers

Insights

- Technology is the highest revenue category
- East region dominates performance
- Q4 consistently produces peak sales

9.2 Shipping Performance Dashboard



REGION

- ☐ Africa
- ☐ Canada
- ☐ Caribbean
- ☐ Central
- ☐ Central Asia
- ■ East
- ☐ EMEA
- ☐ North
- ☐ North Asia

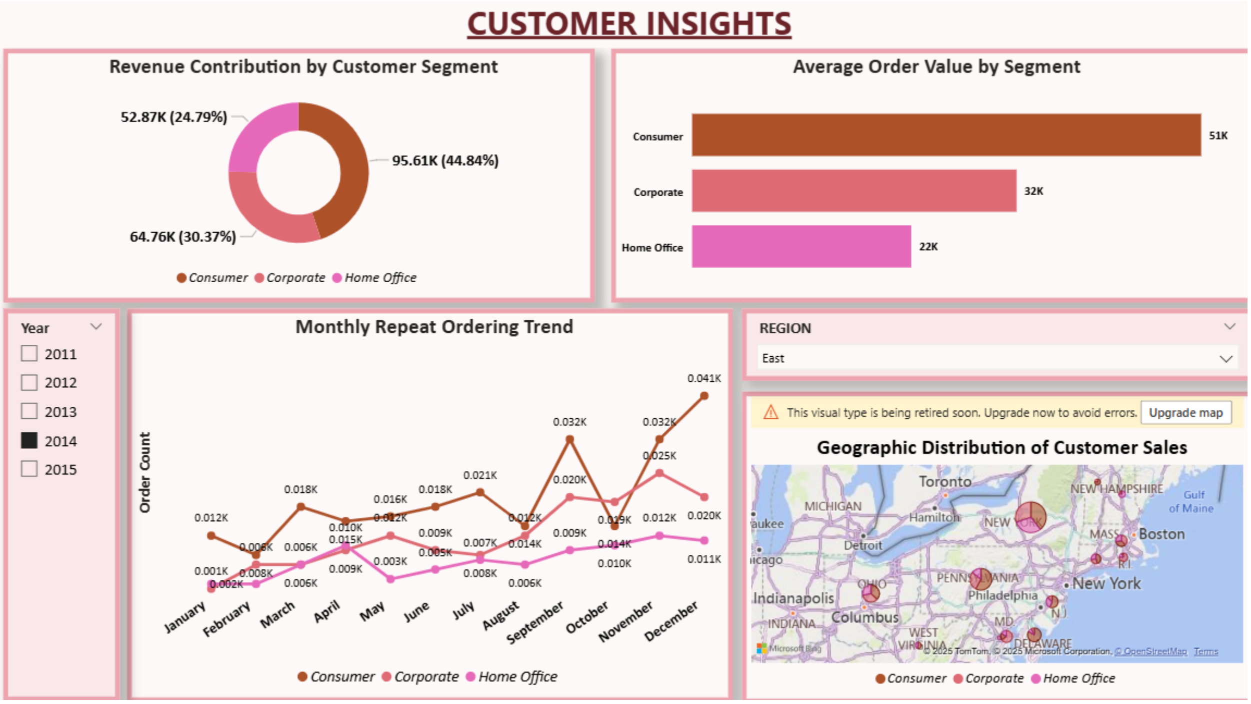
⚠ This visual type is being retired soon. Upgrade now to avoid errors.

- Days-to-Ship trend
- Shipping Cost by mode
- On-time vs Late Shipping
- Delay Heat Map

Insights

- 17% shipments arrive late
- Standard Class has slowest delivery
- APAC & EMEA regions have highest delay rates

9.3 Customer Insights Dashboard



CUSTOMER INSIGHTS

Year

□ 2011
□ 2012
□ 2013
■ 2014
□ 2015

REGION

East

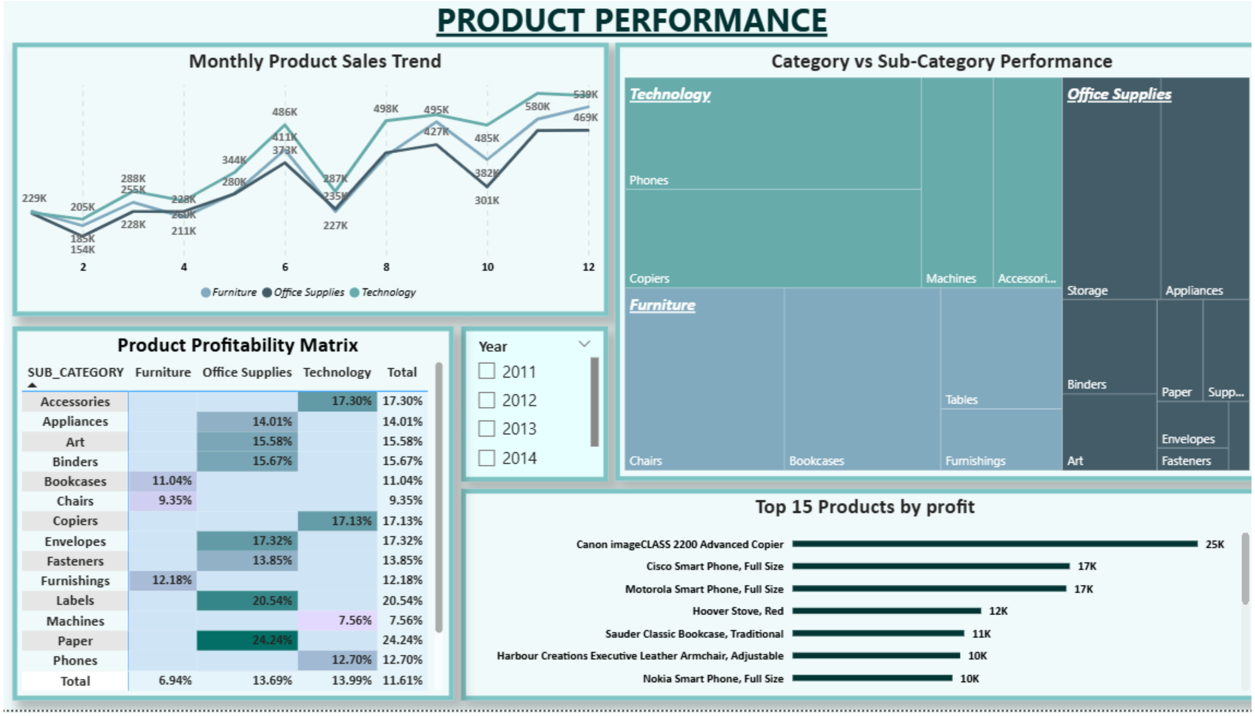
⚠ This visual type is being retired soon. Upgrade now to avoid errors.

- Pareto (80/20) customer revenue distribution
- Segment-wise AOV
- Monthly repeat customer trend

Insights

- Top 20% of customers generate ~50% revenue
- Corporate customers show strongest profitability
- Repeat customers rising consistently

9.4 Product & Category Performance Dashboard


PRODUCT PERFORMANCE
Monthly Product Sales Trend
229K
205K
185K
154K
288K
255K
228K
228K
260K
211K
344K
280K
486K
411K
373K
287K
235K
227K
498K
495K
427K
485K
382K
301K
580K
539K
469K
2
4
6
8
10
12
● Furniture ● Office Supplies ● Technology
Category vs Sub-Category Performance
Technology
Phones
Copiers
Machines
Accessori...
Office Supplies
Storage
Appliances
Binders
Paper
Supp...
Envelopes
Art
Fasteners
Furniture
Chairs
Bookcases
Tables
Furnishings
Year
□ 2011
□ 2012
□ 2013
□ 2014
Top 15 Products by profit
Canon imageCLASS 2200 Advanced Copier 25K
Cisco Smart Phone, Full Size 17K
Motorola Smart Phone, Full Size 17K
Hoover Stove, Red 12K
Sauder Classic Bookcase, Traditional 11K
Harbour Creations Executive Leather Armchair, Adjustable 10K
Nokia Smart Phone, Full Size 10K

Product Profitability Matrix

SUB_CATEGORY	Furniture	Office Supplies	Technology	Total
Accessories			17.30%	17.30%
Appliances		14.01%		14.01%
Art		15.58%		15.58%
Binders		15.67%		15.67%
Bookcases	11.04%			11.04%
Chairs	9.35%			9.35%
Copiers			17.13%	17.13%
Envelopes		17.32%		17.32%
Fasteners		13.85%		13.85%
Furnishings	12.18%			12.18%
Labels		20.54%		20.54%
Machines			7.56%	7.56%
Paper		24.24%		24.24%
Phones			12.70%	12.70%
Total	6.94%	13.69%	13.99%	11.61%

- Category profitability
- Sub-category ranking
- Monthly sales trend
- Discount vs Margin analysis

Insights

- Phones & Chairs are top performing
- Binders & Furnishings underperform
- High discounts strongly reduce margins

10. Key Insights & Discussion

- Revenue is heavily concentrated in Technology and consumer markets.
- Shipping delays are a notable operational issue, especially internationally.
- Customer loyalty drives consistent revenue growth.
- Product portfolio efficiency varies significantly by sub-category.
- High discounts correlate with lower profitability.

These findings can guide decisions related to pricing, inventory planning, regional strategy, and logistics optimization.

11. Challenges & Limitations

- Raw data quality inconsistencies required substantial cleaning.
- Designing correct grain for four fact tables needed multiple iterations.
- MD5 surrogate key strategy introduced debugging complexity.
- Power BI feature changes (e.g., missing multi-row card) limited visual options.
- High-cardinality fields impacted report performance.

12. Reflection

This project helped us understand how real-world data systems are built beyond classroom theory. We learned the importance of:

- Proper data cleaning before modeling
- Separating staging, transformation, and warehouse layers
- Designing facts with precise grain
- Maintaining conformed dimensions
- Implementing business rules in ETL, not BI
- Debugging and validating an entire ETL pipeline

13. Future Enhancements

- Automate ETL using Snow pipe + scheduled tasks
- Implement incremental loads
- Add Inventory and Supplier fact tables
- Add forecasting and predictive analytics
- Implement Row-Level Security (RLS) in Power BI
- Add anomaly detection and data quality monitoring dashboards

14. Conclusion

This project demonstrates the successful development of a complete end-to-end data warehouse and BI ecosystem. From cleaning raw CSV data to building a structured Snowflake warehouse and interactive Power BI dashboards, we delivered a solution that provides clear insights into sales, customers, products, and logistics.

The system is scalable, accurate, and aligned with modern data engineering best practices, laying the foundation for advanced analytics and future automation.