Data Warehouse Project – Milestone 2

Guided by: Prof. Joseph Kinn
Team Alpha: Sakshi Singh, Shivani Pandeti, Amulya Walimbe, Reeya Patra
Course: IST 722
Date: 10/03/25

1. Project Overview

The objective of this milestone was to design, implement, and validate a complete **ETL (Extract, Transform, Load)** data pipeline using **Snowflake** as the data warehouse platform and **Power BI** for visualization.

The pipeline converts raw data from the *Global Superstore* dataset into a structured **star schema** supporting analytical queries for multiple business processes:

1. Sales Analysis
2. Shipping Performance
3. Customer Monthly Performance
4. Product Monthly Performance

All processes are fully automated through SQL scripts executed in Snowflake, ensuring data integrity, consistency, and reusability.

2. System Architecture

The architecture follows a **three-layer Snowflake design**, ensuring separation of concerns between ingestion, transformation, and presentation.

Layer	Schema	Description
Raw Stage	RAW_STAGE	Stores raw data exactly as imported from CSV. Acts as the landing zone for all extracts.
Transform	TRANSFORM	Performs data cleansing, standardization, key generation, and business logic transformations. Houses intermediate and aggregated fact and dimension tables.
Data Warehouse (Presentation Layer)	DW	Contains the final validated tables used for reporting and visualization in Power BI.

All scripts were executed using the compute warehouse: **SUPERSTORE_WH**
and the central database: **GLOBAL_SUPERSTORE_DW**

3. ETL Implementation Workflow

Step 1 – Environment Setup (01_CREATE_DATABASE_SCHEMA.sql)

The first step involved creating the Snowflake environment including:

- Compute Warehouse: SUPERSTORE_WH
- Database: GLOBAL_SUPERSTORE_DW
- Schemas: RAW_STAGE, TRANSFORM, and DW

Step 2 – Data Extraction & Staging (02_STAGE_LOAD.sql)

The raw dataset Global_Superstore2.csv was uploaded into the Snowflake internal stage and loaded into a staging table named:

RAW_STAGE.STG_GLOBAL_SUPERSTORE

Purpose:
To store the unprocessed dataset with proper data types and structure.

Key Columns Loaded:
Order ID, Customer ID, Product ID, Country, Category, Sales, Quantity, Profit, Discount, Ship Date, Order Date, and Shipping Cost.

Outcome:
Raw data successfully extracted into Snowflake with **51,290 rows** and verified column mappings.

Step 3 – Data Transformation (03_DIMENSIONS_BUILD.sql & 04_FACTS_BUILD.sql)

Transformation logic was executed in the TRANSFORM schema to clean, standardize, and structure the dataset into **dimension** and **fact** tables.

3.1 Dimension Tables

Five core dimensions were created:

Dimension	Key Attributes	Purpose
DIM_DATE	DateKey, FullDate, Month, Quarter, Year	Provides time hierarchy for analysis
DIM_CUSTOMER	CustomerKey, CustomerID, CustomerName, Segment	Identifies unique customers and segments
DIM_PRODUCT	ProductKey, ProductID, ProductName, Category, Sub-Category	Captures product-level information
DIM_GEOGRAPHY	GeoKey, Country, State, City, Region	Defines the market and regional structure

DIM_SHIPPING	ShipKey, ShipMode, OrderPriority	Classifies delivery types and priorities

Key Transformation Techniques:

- Use of **DISTINCT** for uniqueness.
- **MD5(CONCAT())** used to generate surrogate keys.
- Derived date components using YEAR(), MONTH(), DAY(), DAYNAME().
- Segmentation for customers and hierarchical grouping for products.

3.2 Fact Tables

Four fact tables were developed to support multiple business processes:

Fact Table	Granularity	Business Process	Key Measures
FACT_SALES	Per Order Line	Sales Analysis	Sales, Quantity, Discount, Profit
FACT_SHIPPING	Per Shipment	Shipping Performance	Shipping Cost, Days to Ship, Delay Flag
FACT_CUSTOMERMONTHLY	Per Customer per Month	Customer Performance	Monthly Sales, Orders, Profit Margin
FACT_PRODUCTMONTHLY	Per Product per Month	Product Performance	Product Sales, Quantity, Monthly Profit Margin

Transform Logic Highlights:

- Aggregate data using GROUP BY for monthly performance tables.
- Calculate delay days using DATEDIFF('day', order_date, ship_date).
- Generate surrogate keys linking each fact to its dimensions.

Step 4 – Validation (05_VALIDATION_CHECKS.sql)

Comprehensive validation queries were executed to ensure the accuracy of transformations before loading the final warehouse schema.

Validation Performed:

1. Record count comparisons between facts and staging.
2. Null and duplicate checks on dimension keys.
3. Verification of referential integrity between facts and dimensions.
4. Range and sanity checks for measures (Profit, Discount, Days to Ship).

Step 5 – Final Load (06_LOAD_DW.sql)

The final step in the pipeline was to **load all validated dimension and fact tables** into the DW schema.
This marks the transition from the transformation layer to the presentation layer, ready for analytical consumption.

Result:
9 Tables successfully loaded (5 Dimensions + 4 Facts).
All row counts matched between TRANSFORM and DW schemas

4. Star Schema Design

Final data model in Snowflake follows a **star schema** structure:

Conformed Dimensions:
DIM_DATE, DIM_CUSTOMER, DIM_PRODUCT, DIM_GEOGRAPHY, DIM_SHIPPING

Fact Tables:
FACT_SALES, FACT_SHIPPING, FACT_CUSTOMERMONTHLY, FACT_PRODUCTMONTHLY

Each fact table is linked to dimensions through surrogate keys.
This design enables scalable analytical queries and supports multiple business processes from a single integrated model.

5. Business Processes Implemented

Business Process	Fact Table	Key Insights Enabled
Sales Analysis	FACT_SALES	Revenue, Profit, Margin, Category performance
Shipping Analysis	FACT_SHIPPING	Cost per delivery, Delay patterns, Mode efficiency
Customer Monthly Analysis	FACT_CUSTOMERMONTHLY	Customer retention, Profitability by segment
Product Monthly Analysis	FACT_PRODUCTMONTHLY	Product trends, Profit by sub-category

Each process is powered by standardized dimension keys, ensuring unified reporting across all analyses.

6. Validation Screenshot

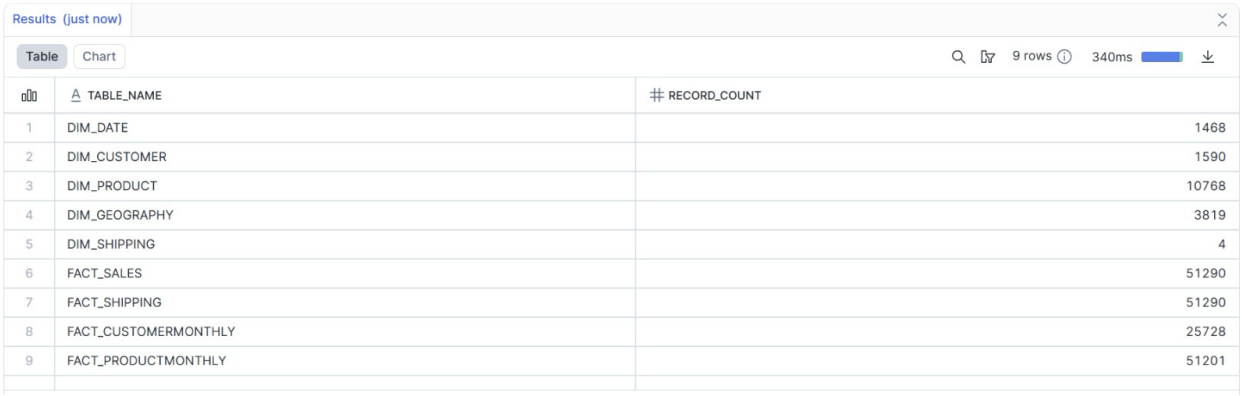
Results (just now)

Table | Chart | 9 rows | 340ms

	TABLE_NAME	RECORD_COUNT
1	DIM_DATE	1468
2	DIM_CUSTOMER	1590
3	DIM_PRODUCT	10768
4	DIM_GEOGRAPHY	3819
5	DIM_SHIPPING	4
6	FACT_SALES	51290
7	FACT_SHIPPING	51290
8	FACT_CUSTOMERMONTHLY	25728
9	FACT_PRODUCTMONTHLY	51201

7. Conclusion

This milestone successfully demonstrates a complete **end-to-end ETL pipeline**:

- Data extracted from raw CSV into Snowflake.
- Cleaned, standardized, and transformed into structured star schema.
- Validated across multiple business processes.
- Ready to be loaded into a dedicated analytical schema (DW) for Power BI visualization.

All four business processes — **Sales**, **Shipping**, **Customer**, and **Product** — were implemented, validated, and integrated into the warehouse environment.
The result is a robust and reusable data warehouse model supporting accurate business insights and scalable analytics.