TARGET DATA WAREHOUSE PROJECT – REPORT

Target Data Warehouse Project – Final Report

1. Introduction

This report details the process of sourcing, integrating, and implementing an ETL pipeline for three datasets that are meaningfully combined in a data warehouse. The datasets are selected based on their suitability for integration, their raw format, and their adequacy for dimensional modeling. The report also includes the complete ETL process, dimensional modeling, and analytical querying.

2. Data Warehouse Platform and Tools Used

2.1 Data Warehouse Platform:

- MySQL Database for structured storage and querying

2.2 ETL Tools Used:

- Excel: Used for preprocessing small datasets (cleaning, formatting, and deduplication)
- MySQL LOAD DATA INFILE: Used for bulk data loading

3. Data Source Selection Criteria

The following criteria were used to select the data sources:

- Raw Format: Data is available in CSV format.
- Related Information: Datasets contain interconnected details that allow for integration.
- Record Count: Each dataset contains at least 500 records to facilitate meaningful analysis.
- Dimensional Modeling Suitability: The datasets are structured using a star schema.

4. Data Source Selection Criteria

To start the project, three related datasets were chosen: Customers_TargetEcom, Products_TargetEcom, and Sales_TargetEcom. These datasets covered the most important aspects of an e-commerce business. They contained information about customer demographics, purchase history, transactional details, and product catalog data.

4.1 Customers_TargetEcom Dataset

- **Format:** CSV
- **Description:** Contains demographic and behavioral information about customers.
- **Attributes:** CustomerID, Name, Email, Phone, Address, City, State, Zip_Code, Registration_Date, Loyalty_Score, Customer_Type.

- **Integration Justification:** This dataset is essential for analyzing customer behaviors, purchase patterns, and segmentation strategies.

4.2 Products_TargetEcom Dataset

- **Format:** CSV
- **Description:** Includes details about the product catalog available in the Target e-commerce store.
- **Attributes:** ProductID, ProductName, Category, Brand, Price, Stock_Level, Review_Rating.
- **Integration Justification:** Links with transaction data to evaluate product performance and sales trends.

4.3 Sales_TargetEcom Dataset

- **Format:** CSV
- **Description:** Serves as the main fact table containing transactional details.
- **Attributes:** OrderID, Dates, StoreID, CustomerID, ProductID, Quantity, Discount, Tax, ShippingFee, TotalAmount, PaymentMethod, City, State, OrderStatus, DeliveryType.
- **Integration Justification:** Provides critical insights into revenue, customer purchase behaviors, and operational efficiency.

A **detailed data dictionary** was created to document field definitions and relationships between these datasets, ensuring consistency and clarity in data integration.

Documentation of Data Sources & Relationships

Each dataset is related through unique identifiers:

- **Customers (customer_id)** links to **Sales (customer_id)**
- **Products (product_id)** links to **Sales (product_id)**
- **Stock(stock_id)** links to **Products (product_id)**
- **Sales (transaction_id)** links to **PaymentMethods (transaction_id)**
- **Sales (transaction_id)** links to **ShippingStatus (transaction_id)**

5. Normalized Database Implementation

The raw datasets were analyzed to identify entities, attributes, and their relationships. The normalization process resulted in the creation of **seven** interrelated tables that adhere to **Third Normal Form (3NF)**, ensuring that each attribute depends solely on the primary key.

5.1 Customers Table

Stores primary customer data, ensuring the uniqueness of customer records.

- **Key Attributes:** customer_id (Primary Key), name, email, phone, address, city, state, zip_code, registration_date, loyalty_score.
- **Normalization Justification:** Customer demographic details are stored separately, ensuring data integrity and eliminating redundancy.

5.2 Products Table

Contains unique product-related details to standardize product information.

- **Key Attributes:** product_id (Primary Key), product_name, category, price, brand, supplier.
- **Normalization Justification:** Product categories are stored separately to reduce data duplication and ensure efficient updates.

5.3 Stock Table

Stores stock-related details, maintaining a reference to the Products table.

- **Key Attributes:** stock_id (Primary Key), product_id (Foreign Key), stock_quantity, discount, rating.
- **Normalization Justification:** Separating stock information prevents unnecessary duplication within the Products table.

5.4 Payment Methods Table

A reference table for payment methods to standardize transactional data.

- **Key Attributes:** payment_id (Primary Key), payment_method.
- **Normalization Justification:** Standardizing payment methods avoids inconsistencies across transaction records.

5.5 Shipping Status Table

Manages shipping details with a unique status reference.

- **Key Attributes:** shipping_id (Primary Key), shipping_status.

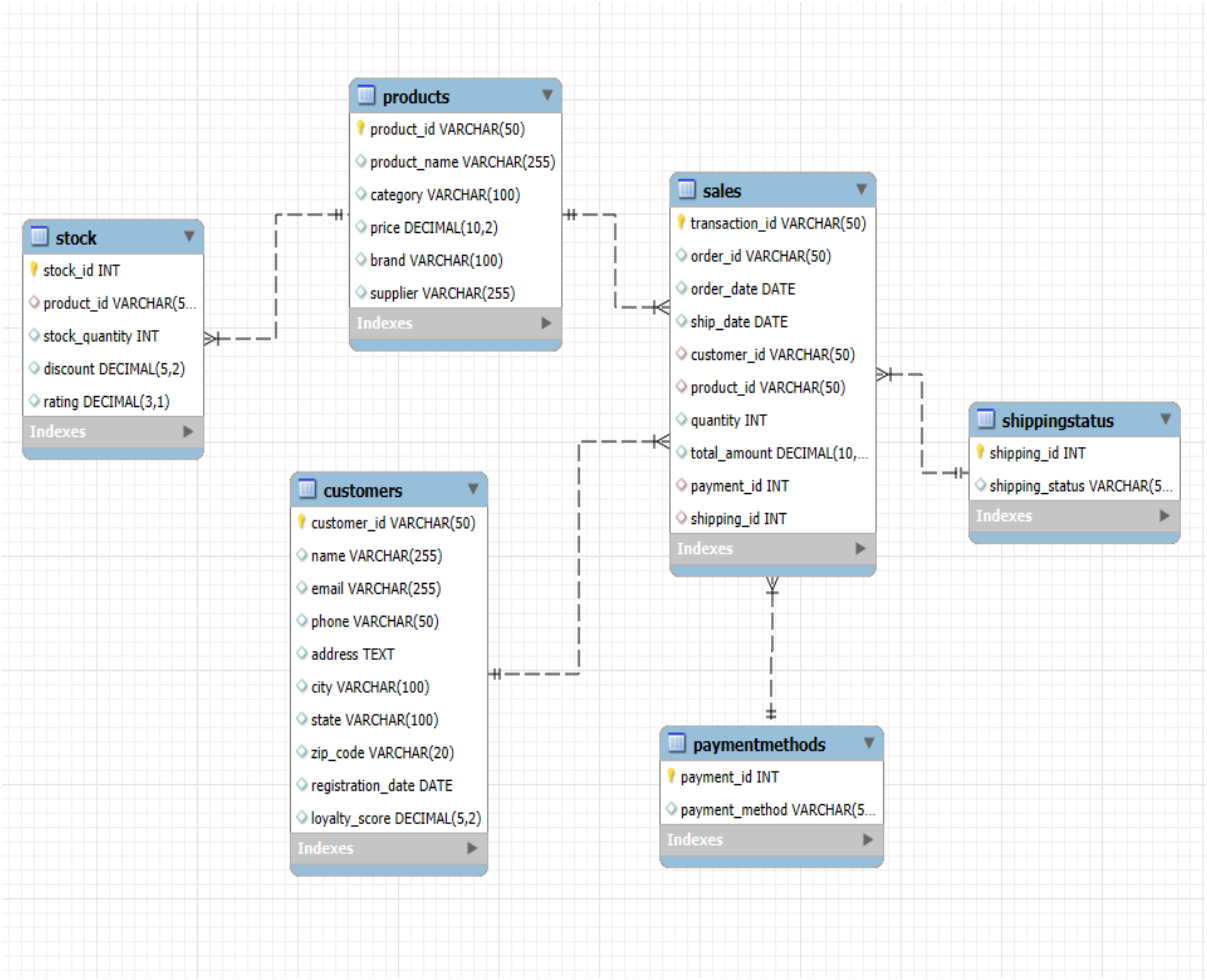
- **Normalization Justification:** Creating a lookup table ensures shipping status is not repeated in multiple transaction records.

5.6 Sales Table

Holds transactional order details linked to customers and locations.

- **Key Attributes:** transaction_id (Primary Key), customer_id (Foreign Key), payment_id(FK), product_id(FK), shipping_id(FK), order_date, ship_date.
- **Normalization Justification:** Storing order headers separately allows efficient tracking and relationship management.

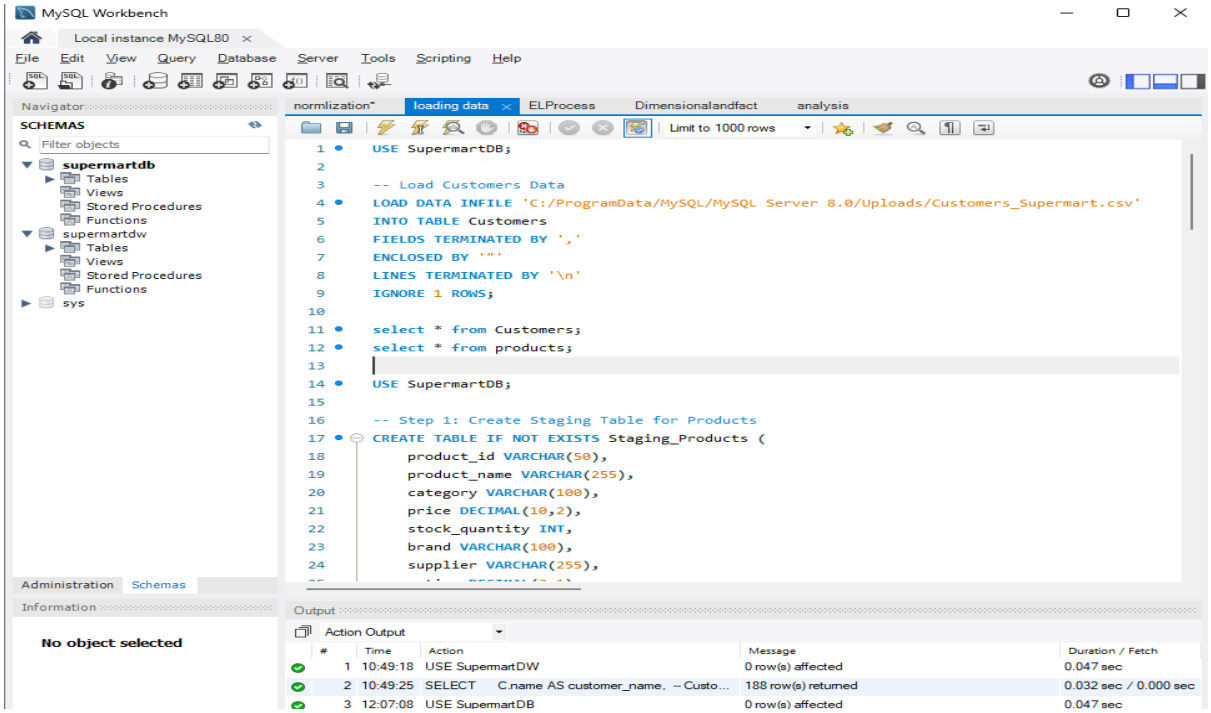
5.8 ER Diagram Representation



6. Data Loading and Staging Implementation

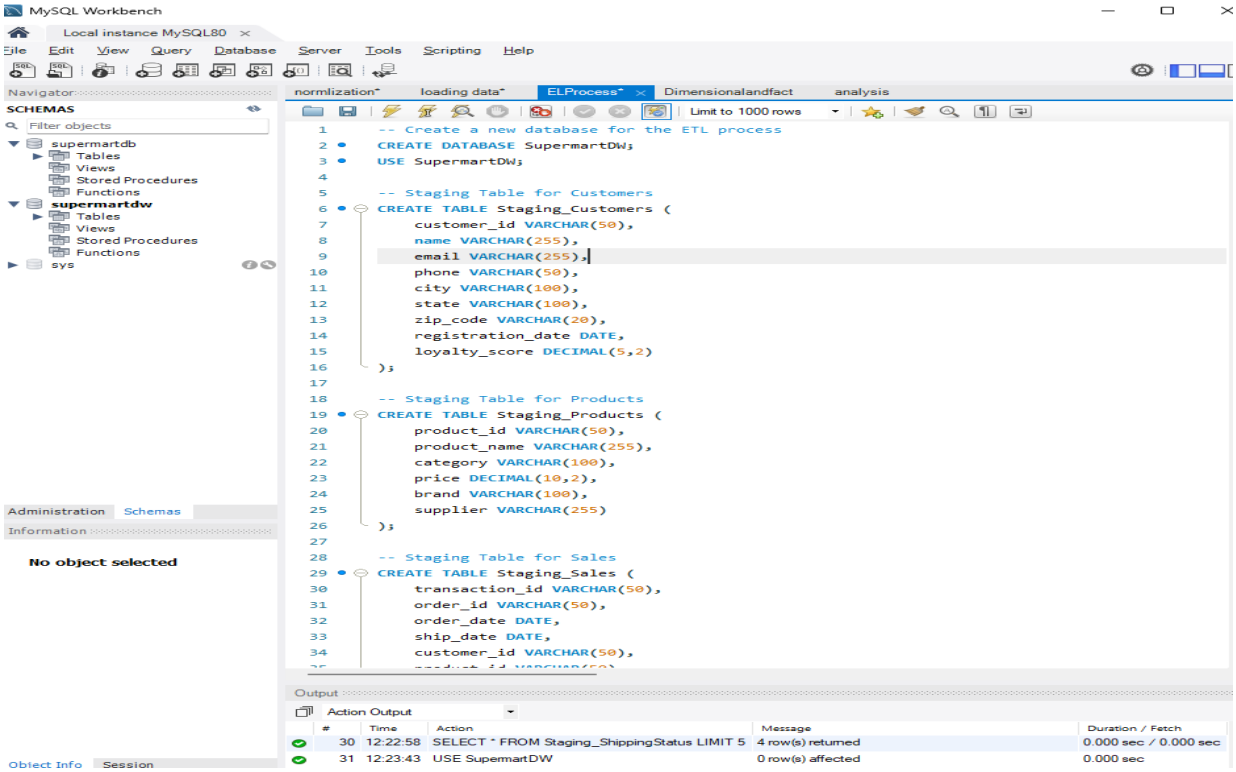
6.1 Data Ingestion into Normalized Tables

Raw data was imported into the database tables using bulk loading techniques, ensuring efficiency and minimal manual intervention.



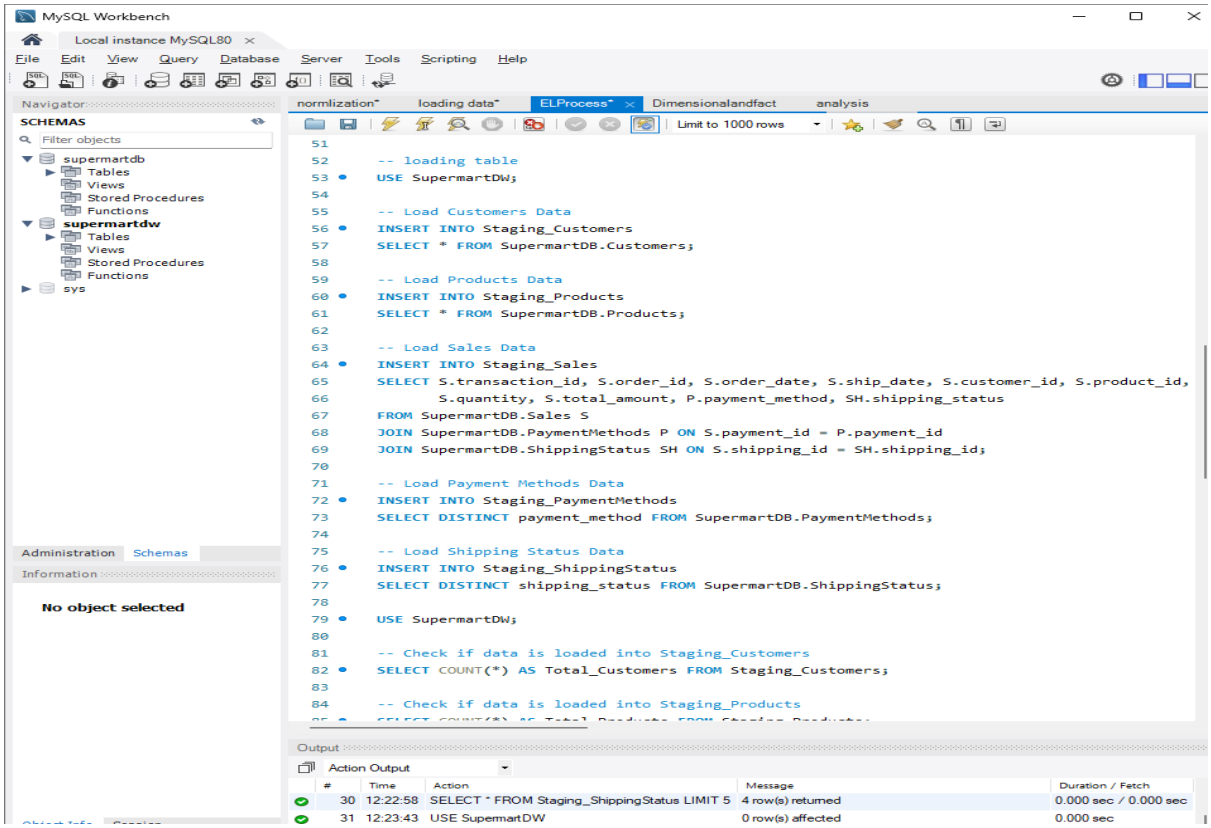
6.2 Data Staging for Transformation

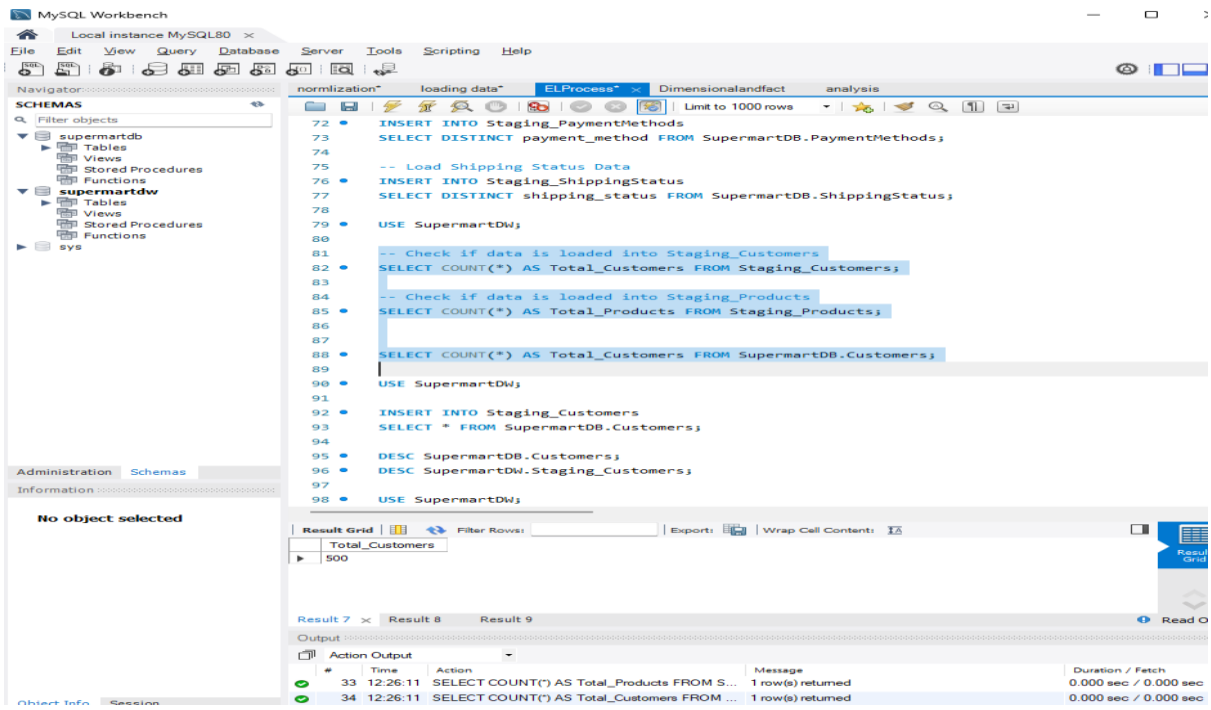
To maintain data integrity before inserting records into the main tables, staging tables were created.

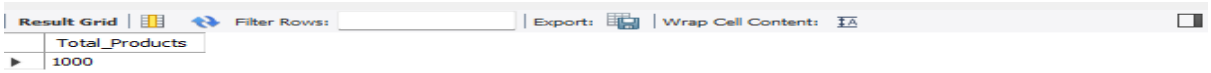


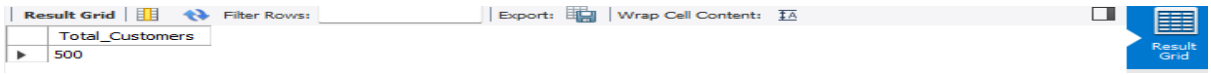
6.3 Data Transformation & Loading

The data from the staging tables underwent transformation processes, ensuring referential integrity and correctness before being inserted into the normalized tables.









7. ETL Process Implementation

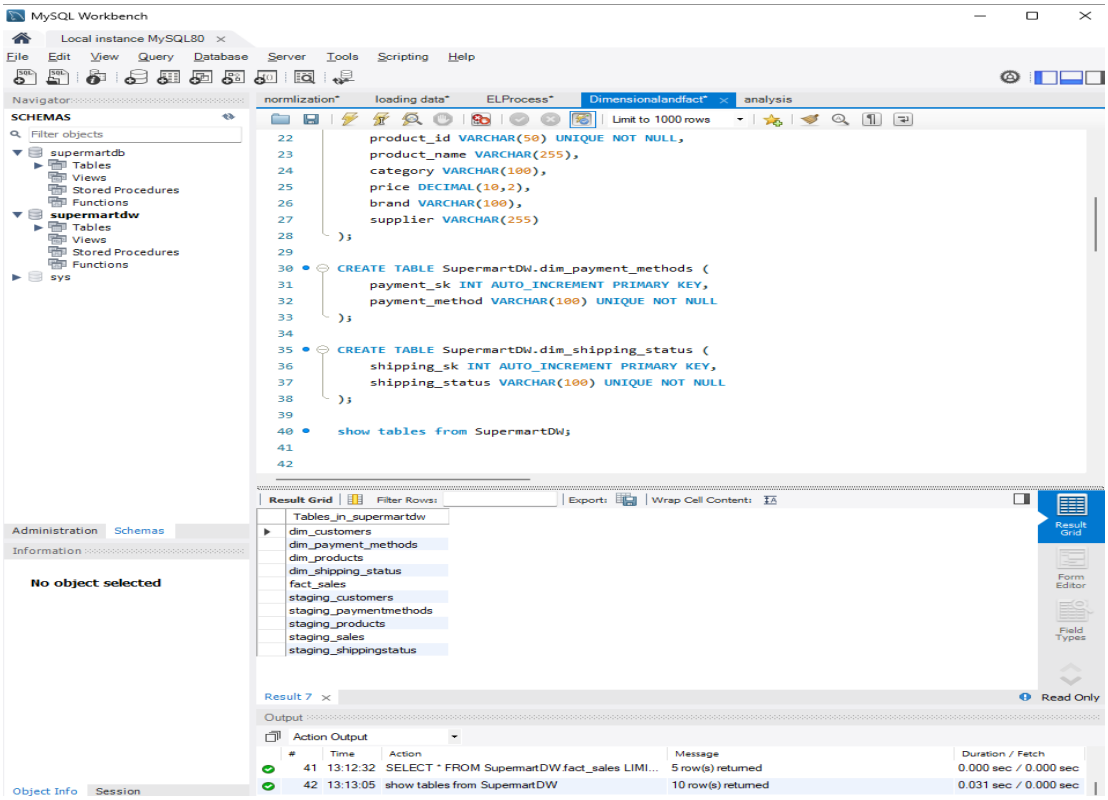
The ETL process was carried out in four distinct phases:

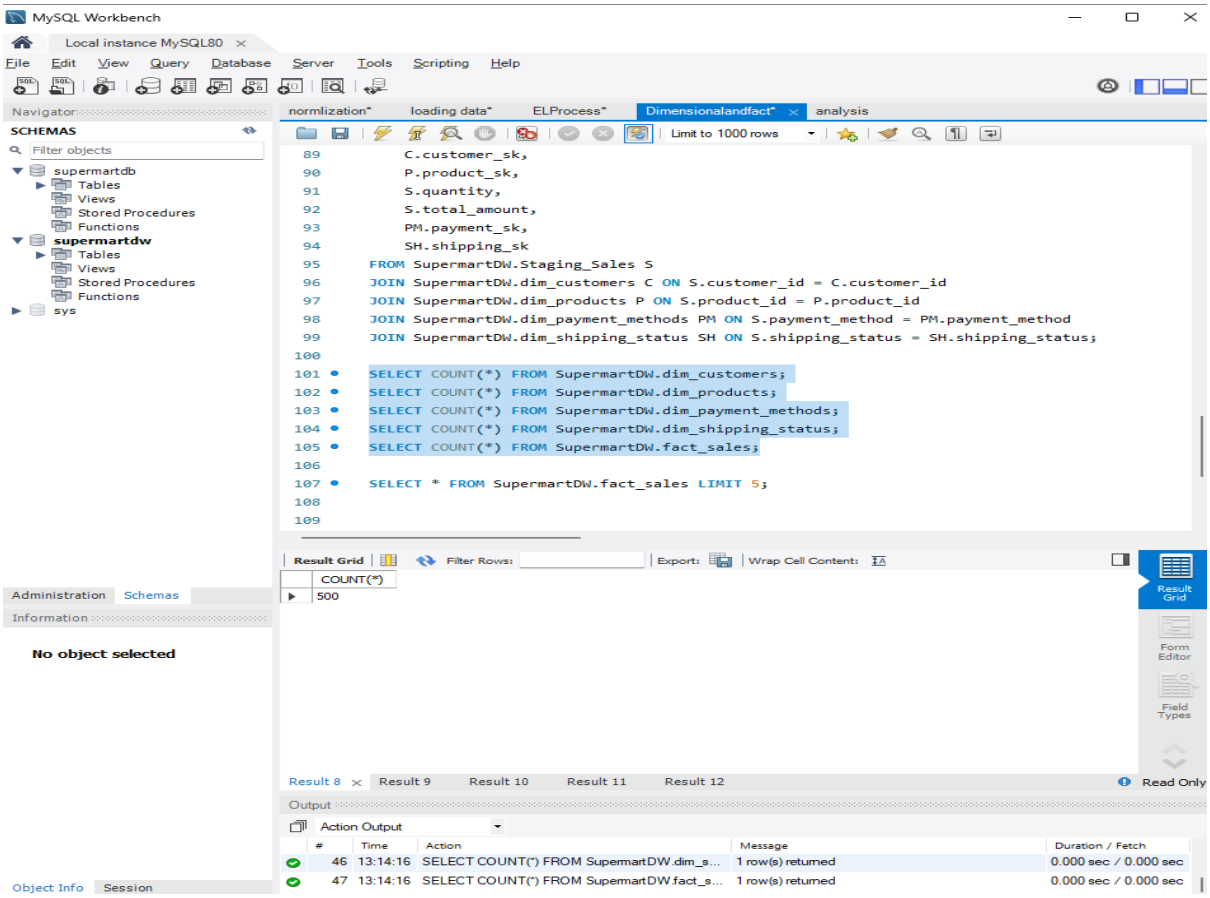
7.1 Phase 1: Extraction

- Data was extracted from SupermartDB into staging tables in SupermartDW.
- Staging tables were used to separate operational and analytical environments, ensuring transformations did not affect source data.
- Staging tables were created for each entity: Staging_Customers, Staging_Products, Staging_Sales, Staging_PaymentMethods, and Staging_ShippingStatus.
- INSERT INTO SELECT statements were used to move data from normalized tables to staging tables.

7.2 Phase 2: Dimension Creation

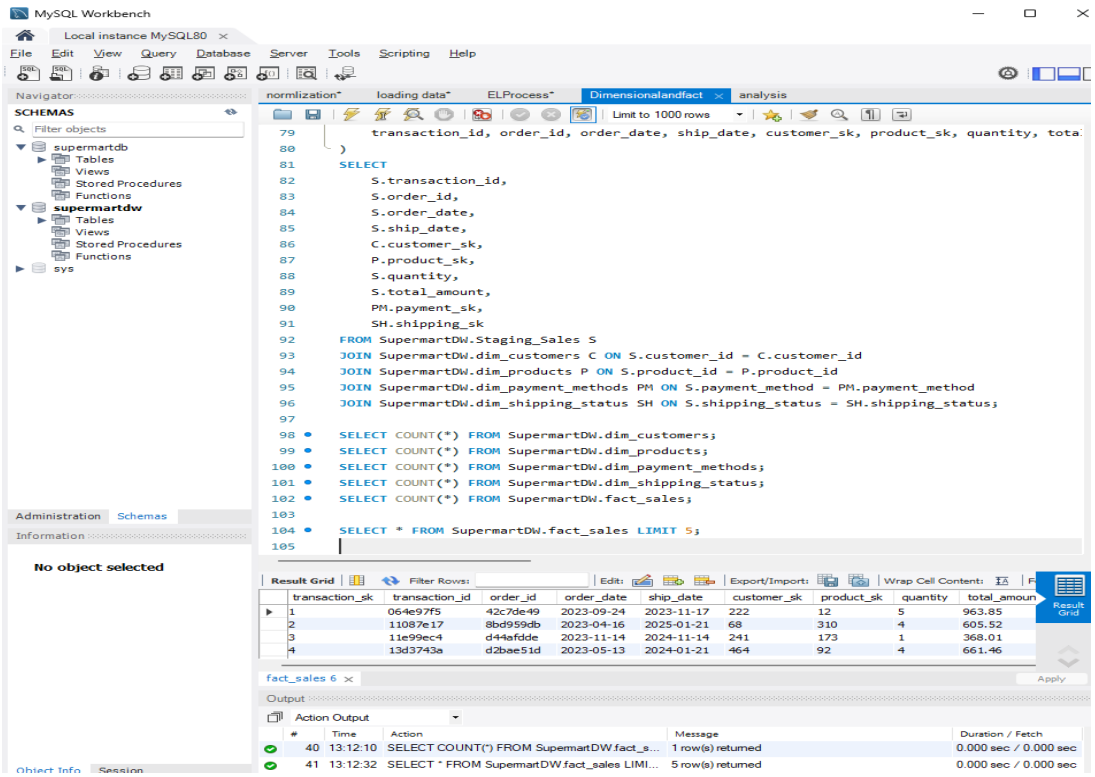
- Data from staging tables was transformed and loaded into four dimension tables:
    - Customer Dimension (dim_customers): Categorized customers into Premium and Standard based on loyalty scores, implemented Slowly Changing Dimension (SCD) Type 2 for tracking changes.
    - Product Dimension (dim_products): Combined product details with denormalized category information and included a derived StockStatus field.
    - Payment Method Dimension (dim_payment_methods): Contained distinct payment methods used in transactions.
    - Shipping Status Dimension (dim_shipping_status): Captured distinct shipping methods from transaction records.
- Surrogate keys were assigned to improve query performance and historical tracking.

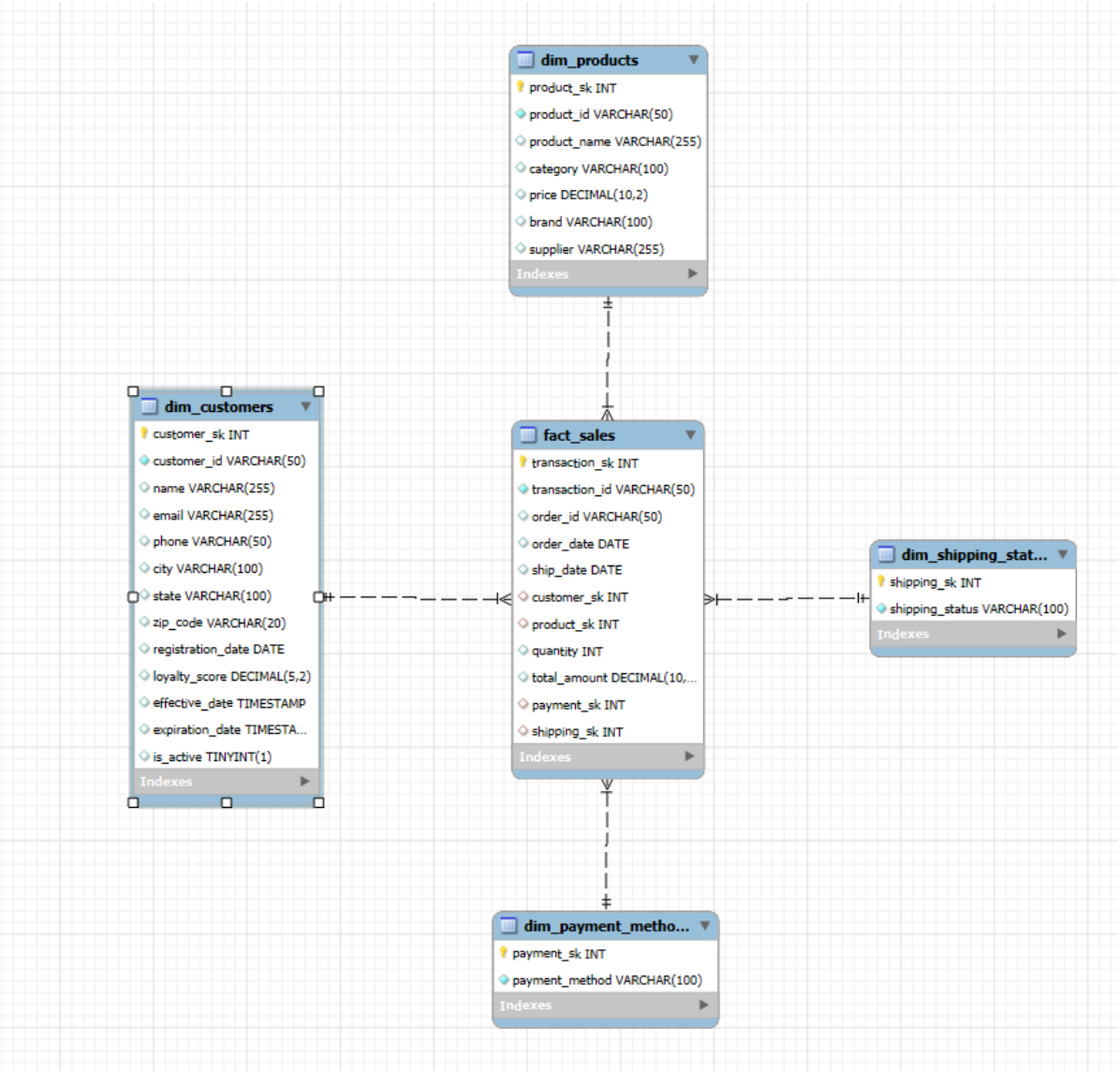

MySQL Workbench
Local instance MySQL80
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
supermartdb
Tables
Views
Stored Procedures
Functions
supermartdw
Tables
Views
Stored Procedures
Functions
sys
normlization" loading data" ELProcess" Dimensionalandfact" analysis
Limit to 1000 rows
product_id VARCHAR(50) UNIQUE NOT NULL,
product_name VARCHAR(255),
category VARCHAR(100),
price DECIMAL(10,2),
brand VARCHAR(100),
supplier VARCHAR(255)
);
CREATE TABLE SupermartDW.dim_payment_methods (
payment_sk INT AUTO_INCREMENT PRIMARY KEY,
payment_method VARCHAR(100) UNIQUE NOT NULL
);
CREATE TABLE SupermartDW.dim_shipping_status (
shipping_sk INT AUTO_INCREMENT PRIMARY KEY,
shipping_status VARCHAR(100) UNIQUE NOT NULL
);
show tables from SupermartDW;
Result Grid Filter Rows: Export: Wrap Cell Content:
Tables_in_supermartdw
dim_customers
dim_payment_methods
dim_products
dim_shipping_status
fact_sales
staging_customers
staging_paymentmethods
staging_products
staging_sales
staging_shippingstatus
Result 7
Read Only
Administration Schemas
Information
No object selected
Object Info Session
Output
Action Output
Time Action Message Duration / Fetch
41 13:12:32 SELECT * FROM SupermartDW.fact_sales LIMI... 5 row(s) returned 0.000 sec / 0.000 sec
42 13:13:05 show tables from SupermartDW 10 row(s) returned 0.031 sec / 0.000 sec


MySQL Workbench
Local instance MySQL80
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
supermartdb
Tables
Views
Stored Procedures
Functions
supermartdw
Tables
Views
Stored Procedures
Functions
sys
normlization" loading data" ELProcess" Dimensionalandfact" analysis
Limit to 1000 rows
C.customer_sk,
P.product_sk,
S.quantity,
S.total_amount,
PM.payment_sk,
SH.shipping_sk
FROM SupermartDW.Staging_Sales S
JOIN SupermartDW.dim_customers C ON S.customer_id = C.customer_id
JOIN SupermartDW.dim_products P ON S.product_id = P.product_id
JOIN SupermartDW.dim_payment_methods PM ON S.payment_method = PM.payment_method
JOIN SupermartDW.dim_shipping_status SH ON S.shipping_status = SH.shipping_status;
SELECT COUNT(*) FROM SupermartDW.dim_customers;
SELECT COUNT(*) FROM SupermartDW.dim_products;
SELECT COUNT(*) FROM SupermartDW.dim_payment_methods;
SELECT COUNT(*) FROM SupermartDW.dim_shipping_status;
SELECT COUNT(*) FROM SupermartDW.fact_sales;
SELECT * FROM SupermartDW.fact_sales LIMIT 5;
Result Grid Filter Rows: Export: Wrap Cell Content:
COUNT(*)
500
Result 8 Result 9 Result 10 Result 11 Result 12
Read Only
Administration Schemas
Information
No object selected
Object Info Session
Output
Action Output
Time Action Message Duration / Fetch
46 13:14:16 SELECT COUNT(*) FROM SupermartDW.dim_s... 1 row(s) returned 0.000 sec / 0.000 sec
47 13:14:16 SELECT COUNT(*) FROM SupermartDW.fact_s... 1 row(s) returned 0.000 sec / 0.000 sec

7.3 Phase 3: Fact Table Population

- The fact_sales table was created and populated:
    - Stored transactional details, including transaction_id, order_id, order_date, ship_date, quantity, total_amount, and discounts.
    - Data from staging tables was joined with dimension tables to assign surrogate keys and derive additional measures.
    - Referential integrity was enforced to maintain correct relationships between fact and dimension tables.

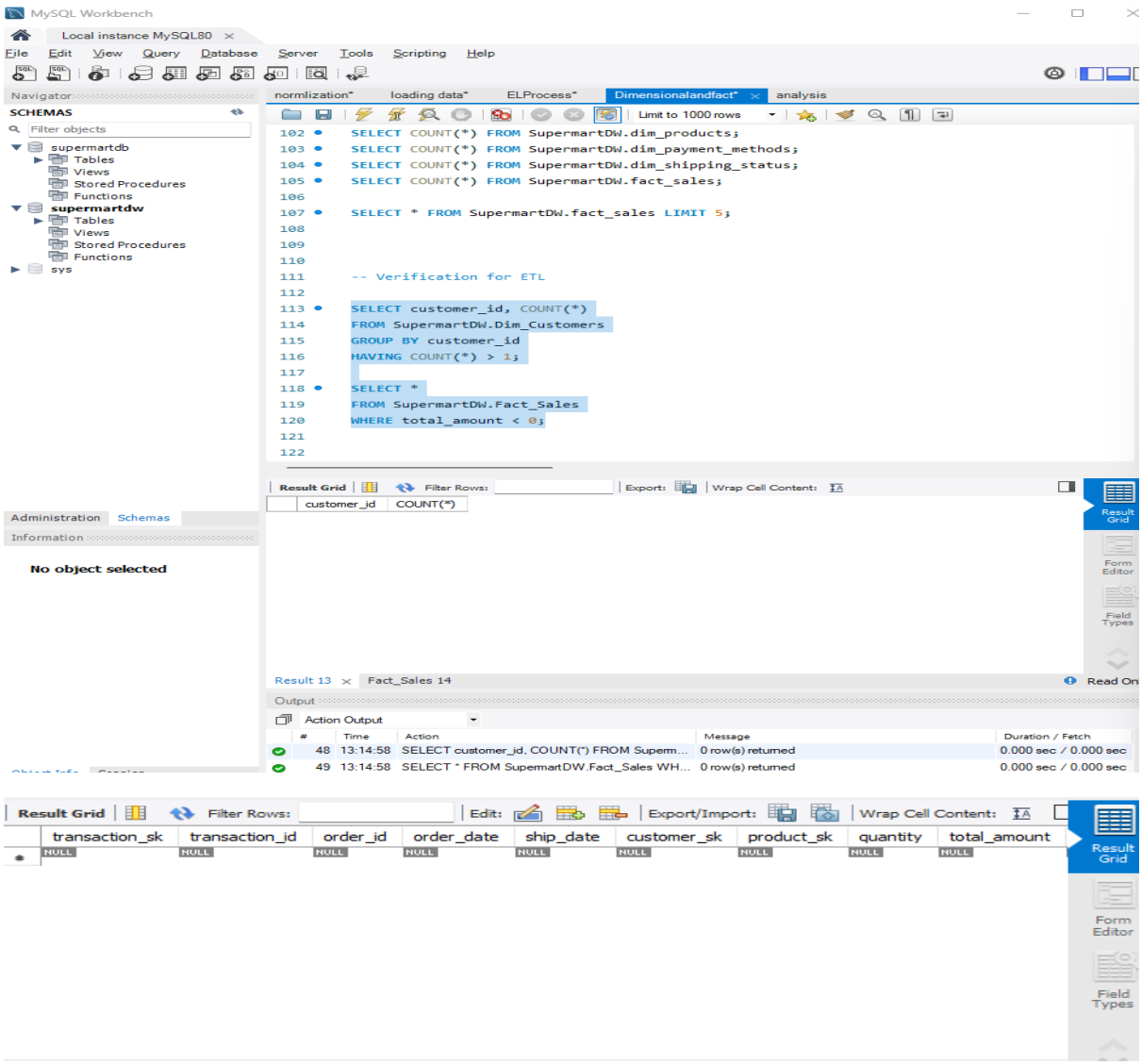




7.4 Phase 4: Verification

- **Dimension Table Checks**:
    - Verified record counts and correctness of data categorization.
    - Ensured **Slowly Changing Dimensions (SCD Type 2)** were implemented correctly.
- **Fact Table Validation**:
    - Ensured accuracy of aggregated sales data, discount calculations, and total revenue figures.
- **Data Quality Audits**:
    - Checked for **NULL values** in critical fields.

- Ensured **referential integrity** between dimensions and fact tables.



8. Analytical Queries

1. Most Preferred Payment Methods

Overview:

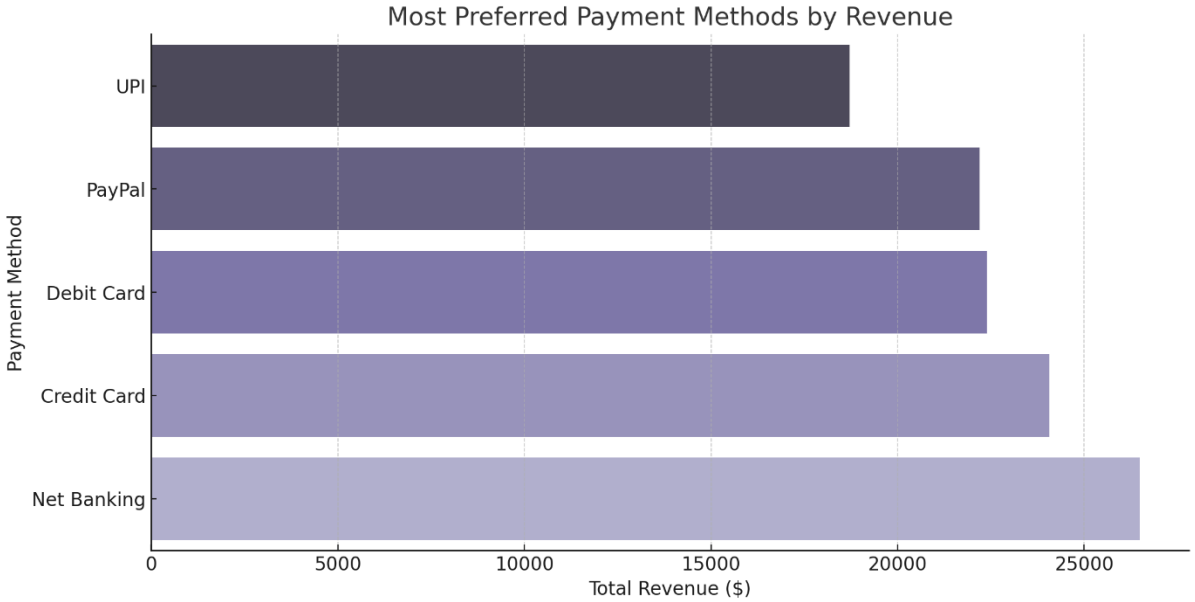
This query identified the most preferred payment methods based on total revenue generated and transaction count.

Insights:

- Net Banking was the most used method with 54 transactions, generating $26,482.79 in revenue.
- Credit Cards and Debit Cards were also popular, collectively contributing $46,459.72.
- Digital wallets (PayPal & UPI) accounted for a significant portion of transactions, indicating customer preference for online payment methods.

Recommendations:

- Offer discounts or cashback on high-value transactions via preferred payment methods.
- Optimize the checkout experience by prioritizing fast and secure digital payment options.



2. Monthly Sales Trends

Overview:

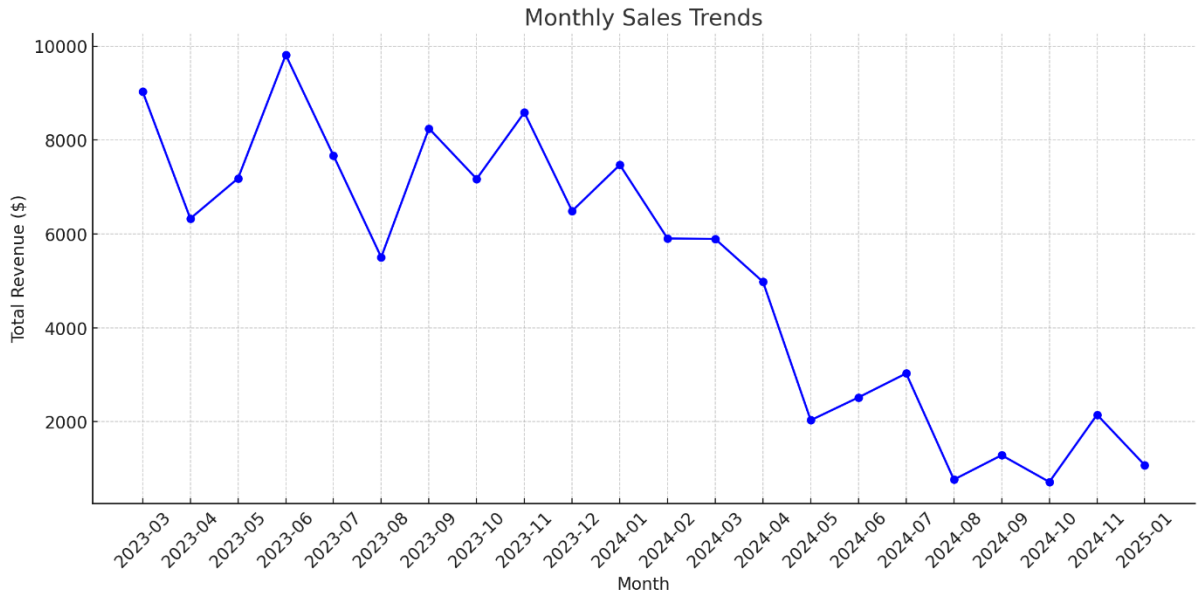
This query extracted **monthly sales revenue trends** to analyze seasonal variations and sales performance.

Insights:

- The highest revenue month was **June 2023 ($9,812.35)**, followed by **March 2023 ($9,035.23)**.
- Sales fluctuated over the months, with a noticeable drop in **August 2024 ($768.54)**.
- Significant revenue declines in **late 2024** may indicate **seasonal demand changes or external factors**.

Recommendations:

- Strengthen **marketing efforts** during low-revenue months to maintain sales consistency.
- Prepare for peak sales months with **inventory adjustments and promotional strategies.**



3. Top-Selling Products

Overview:

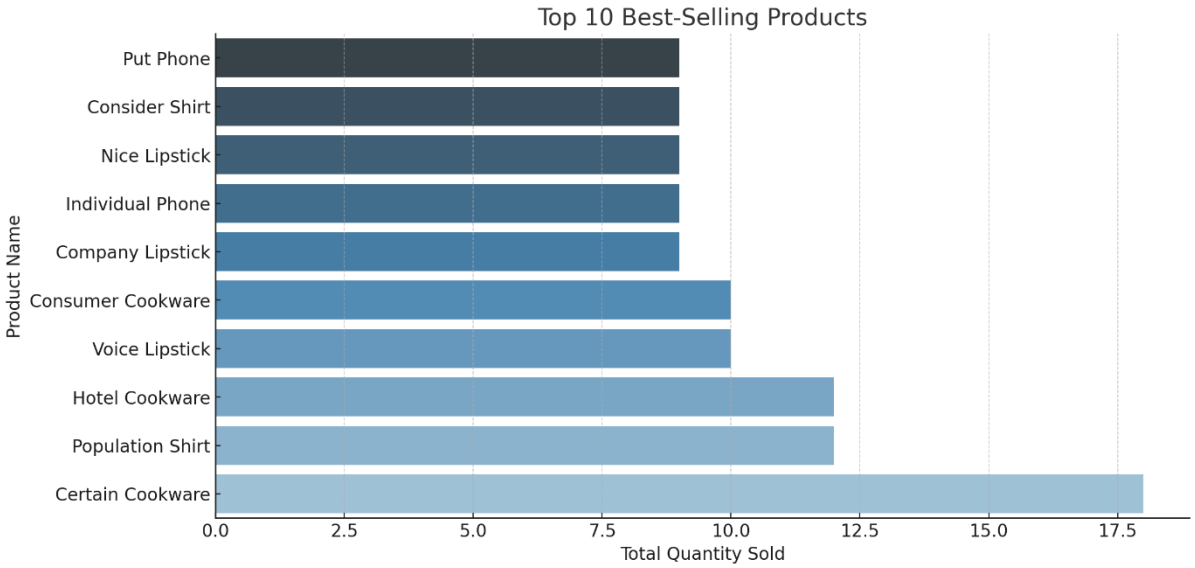
This query determined the **best-selling products** by aggregating **total quantity sold**.

Insights:

- **Certain Cookware** was the best-selling product with **18 units sold**.
- Clothing, electronics, and kitchen products were among the most frequently purchased categories.
- Some products had **low sales performance**, indicating potential for improvement via promotions.

Recommendations:

- Ensure **high-demand products are adequately stocked**.
- Use **cross-selling techniques** to bundle less popular products with high-demand items.



4. Customer Purchase Behaviour

Overview:

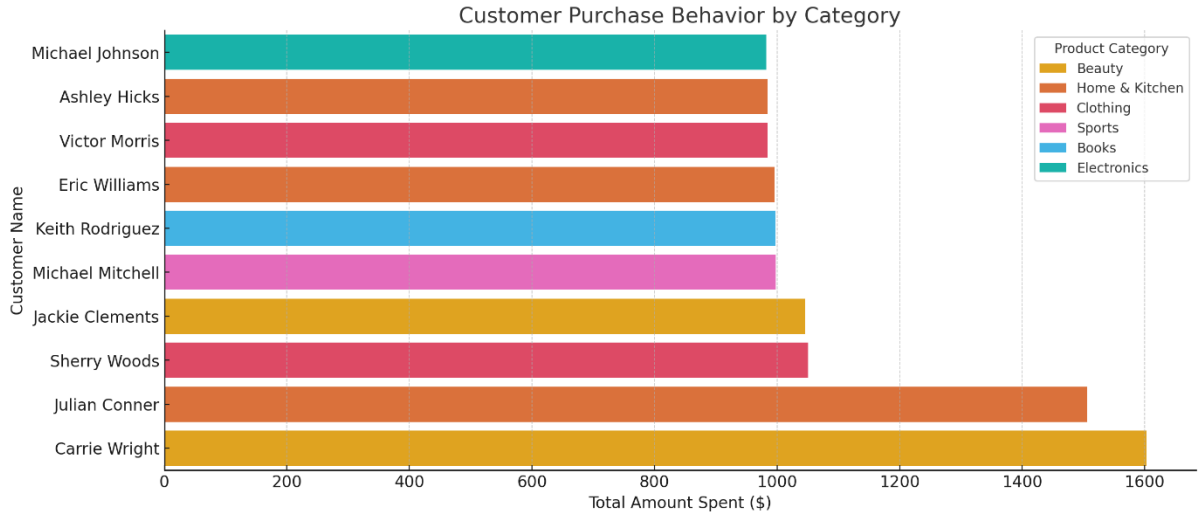
This query analysed **customer spending behaviour** across different **product categories**.

Insights:

- **Carrie Wright spent the highest amount ($1,603.10) on beauty products**, followed by **Julian Conner ($1,506.28) on home & kitchen items**.
- Customers showed strong purchasing patterns in categories like **beauty, electronics, and home & kitchen**.
- Some customers made multiple purchases within the same category, indicating **brand loyalty or recurring needs**.

Recommendations:

- Implement **personalized marketing** for customers based on their **preferred categories**.
- Introduce **membership programs** to reward high-spending customers and encourage repeat purchases.



5. Total Sales per Customer

Overview:

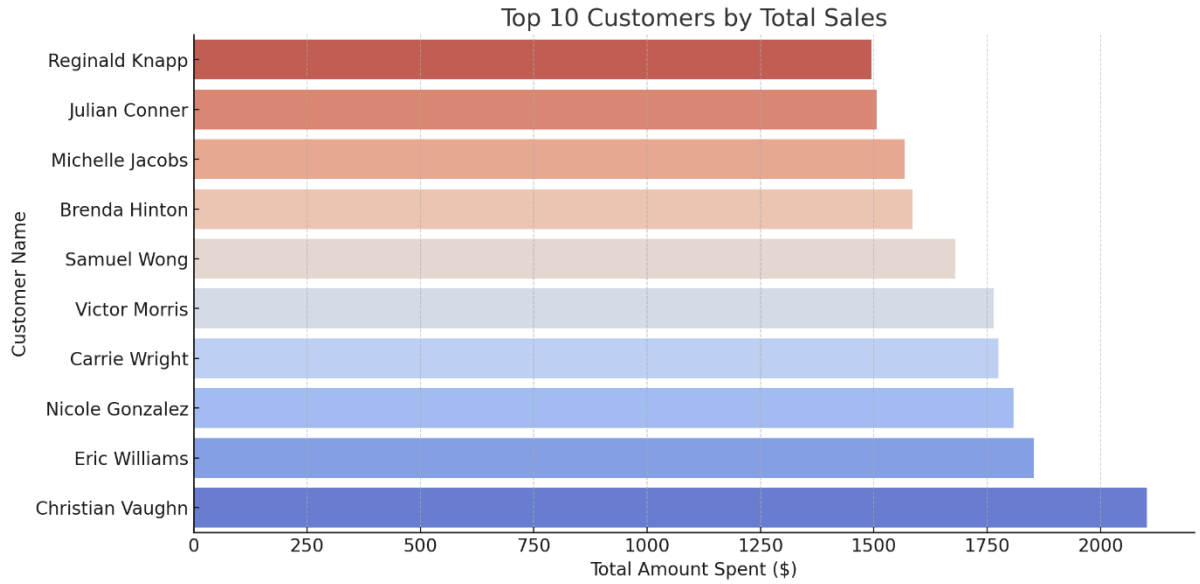
This query identified the **highest-spending customers** to understand purchasing power.

Insights:

- **Christian Vaughn** was the **top spender** with **$2,102.08**, followed by **Eric Williams ($1,853.55)**.
- High-value customers **significantly impact total revenue**, making them ideal targets for **premium offers and loyalty programs**.
- There was a **wide range of customer spending**, suggesting **different consumer segments**.

Recommendations:

- Develop **customer tiers** (e.g., VIP, frequent buyers) to **offer exclusive deals and rewards**.
- Encourage **low-spending customers** to increase purchases through **personalized recommendations and targeted ads**.



9. Recommendations for Future Improvements

Based on the analysis of total sales per customer, top-selling products, preferred payment methods, monthly sales trends, and customer purchase behavior, the following improvements can be made:

9.1 Enhancing Customer Segmentation and Loyalty Programs

Current Observations:

- Some customers contribute significantly to sales, while others have lower spending.
- High-value customers tend to buy frequently from specific categories.

Recommendations:

- Implement **customer segmentation** (e.g., VIP, Regular, Occasional Buyers).
- Offer **personalized discounts, cashback, and loyalty points** for high-value customers.
- **Encourage repeat purchases** from mid-tier customers by sending **targeted promotions and recommendations**.

9.2 Optimizing Inventory and Stock Management

Current Observations:

- Certain products (like **cookware and clothing**) are **top sellers**, while others underperform.
- Stockouts may lead to missed revenue opportunities.

Recommendations:

- **Automate inventory tracking** to ensure **top-selling products are always available**.
- Offer **discounts or bundling** for slow-moving products to clear out excess stock.
- Use **predictive analytics** to forecast **demand trends** and optimize restocking.

9.3 Improving Payment and Checkout Experience

Current Observations:

- **Net Banking, Credit Cards, and PayPal** generate the most revenue.
- Some payment methods are underutilized.

Recommendations:

- Promote **popular payment methods** by offering incentives (e.g., cashback on credit card purchases).
- Optimize the **checkout experience** to reduce cart abandonment and improve conversions.
- Introduce **buy now, pay later (BNPL)** options to attract higher-value purchases.

9.4 Seasonal and Monthly Sales Planning

Current Observations:

- Sales fluctuate significantly, with some months showing a **major decline**.
- Peak sales months indicate **seasonal trends**.

Recommendations:

- **Run promotions during low-sales months** to encourage spending.
- **Leverage peak seasons** with targeted ad campaigns and limited-time offers.
- Introduce **seasonal products and special bundles** to capitalize on demand trends.

9.5 Data-Driven Business Strategies

Current Observations:

- Insights from customer spending and product trends can **drive better decision-making**.

Recommendations:

- Implement **machine learning models** to **predict customer preferences and demand**.
- Use **real-time analytics dashboards** for tracking sales performance.
- Continuously refine **pricing strategies** based on competitor analysis and market trends.

Conclusion

The **Dimensional Modeling Implementation for Target E-Commerce** successfully integrated **three structured datasets**—Customers, Products, and Sales—into a **star-schema-based data warehouse (SupermartDW)**. The **ETL process** ensured data consistency, integrity, and accessibility for analytical queries, supporting **business intelligence and decision-making**.

Key Achievements:
Optimized Data Warehouse: Implemented a **structured ETL process** in MySQL, ensuring seamless data integration and transformation.
Enhanced Analytical Capabilities: Developed insightful **business intelligence queries** to analyze **customer spending, product sales trends, and payment preferences**.
Improved Business Insights: Identified **high-value customers, top-selling products, preferred payment methods, and seasonal sales trends** to drive data-driven strategies.
Effective Inventory & Payment Management: Analyzed product demand and payment trends to optimize **stock management and checkout experience**.
Scalable and Future-Ready System: Designed a **scalable warehouse** that can support future enhancements like **real-time analytics, predictive modeling, and AI-driven recommendations**.

Future Scope

- Implement **Machine Learning (ML) models** for **sales forecasting and demand prediction**.
- Integrate **real-time ETL processing** to support **instant decision-making**.
- Enhance **personalized marketing strategies** through **customer segmentation and behavioural analytics**.
- Optimize **supply chain efficiency** by linking sales insights with logistics and supplier data.

By leveraging the structured **dimensional model and analytical insights**, the business can improve **customer engagement, sales performance, and operational efficiency**, ensuring a **competitive advantage in the e-commerce industry**.