Super Store Data Warehouse

Created By: Bruno Fernandes, Hugo Abelheira, Tiago Coelho

Agenda

- Context
- Dimensional Bus Matrix
- Data Dictionaries
- Dimensional Model
- ETL Design
- Queries & Results
- Visualization
- Advantages and Challenges
- Future Work
- Conclusion

Context

The subject of this project is based on a retail dataset — the *Sample Superstore* dataset — which contains detailed information on orders, customers, products, sales, shipping, and profits across various U.S. regions.

The goal of this project is to design and implement a data warehouse solution that enables multidimensional analysis of retail performance. By modeling the data using dimensional modeling techniques, we aim to facilitate efficient and insightful queries regarding business performance across time, geographical areas, and product categories.

Dimensional Bus Matrix

Data mart	Star	Dimension	Calendar	CalendarMonth	Customer	Shipping	Location	Product	Category	State	Region
Sales	Item		x		X		X	X			
	Orders		X		X	X	X				
	OrderM			X						X	
	ProductPerformance			X					X	X	
	ShippingBehavior					X			X		X
	ShippingBehaviorS					X			X	X	

Data Dictionaries - Dimensions

Name	Description	SCD	Version	1.0	Date	2025-03-31
Calendar	Stores information about transaction and shipping dates	Type 1	Hierarchy	Day < Month < Year		
Attribute	Description	Level	Key	Type	Size	Precision
calendar_id	Unique identifier for date	Date	PK	ID		
full_date	Full date (YYYY-MM-DD)	Date	UK	DATE		
year_id	Year Surrogate	Year	LK	INT		
year_number	Year of the date	Year	UK	INT		0
month_id	Month Surrogate	Month	LK	INT		
month_number	Month of the date	Month	UK	INT		0
month_name	Name of the month	Month		VARCHAR	15	
day_id	Day Surrogate	Day	LK	INT		
day_number	Day of the date	Day	UK	INT		0

Data Dictionaries - Fact Tables

Star	**Orders**	Version	1.0	Date	2025-03-31
Granularity	One order instance				
Dimensions					
OrderCalendar	Calendar				
ShippingCalendar	Calendar				
Customer	Customer				
Location	Location				
Shipping	Shipping				
Measures					
sales_order	Total price of the order (an order may contain different products)				
quantity_order	Total number of products in the order				
order_code	Identifier of the order (degenerate)				
lost_value_order	Difference between full price of the order and the discounted price				
profit_order	Total profit of the order				

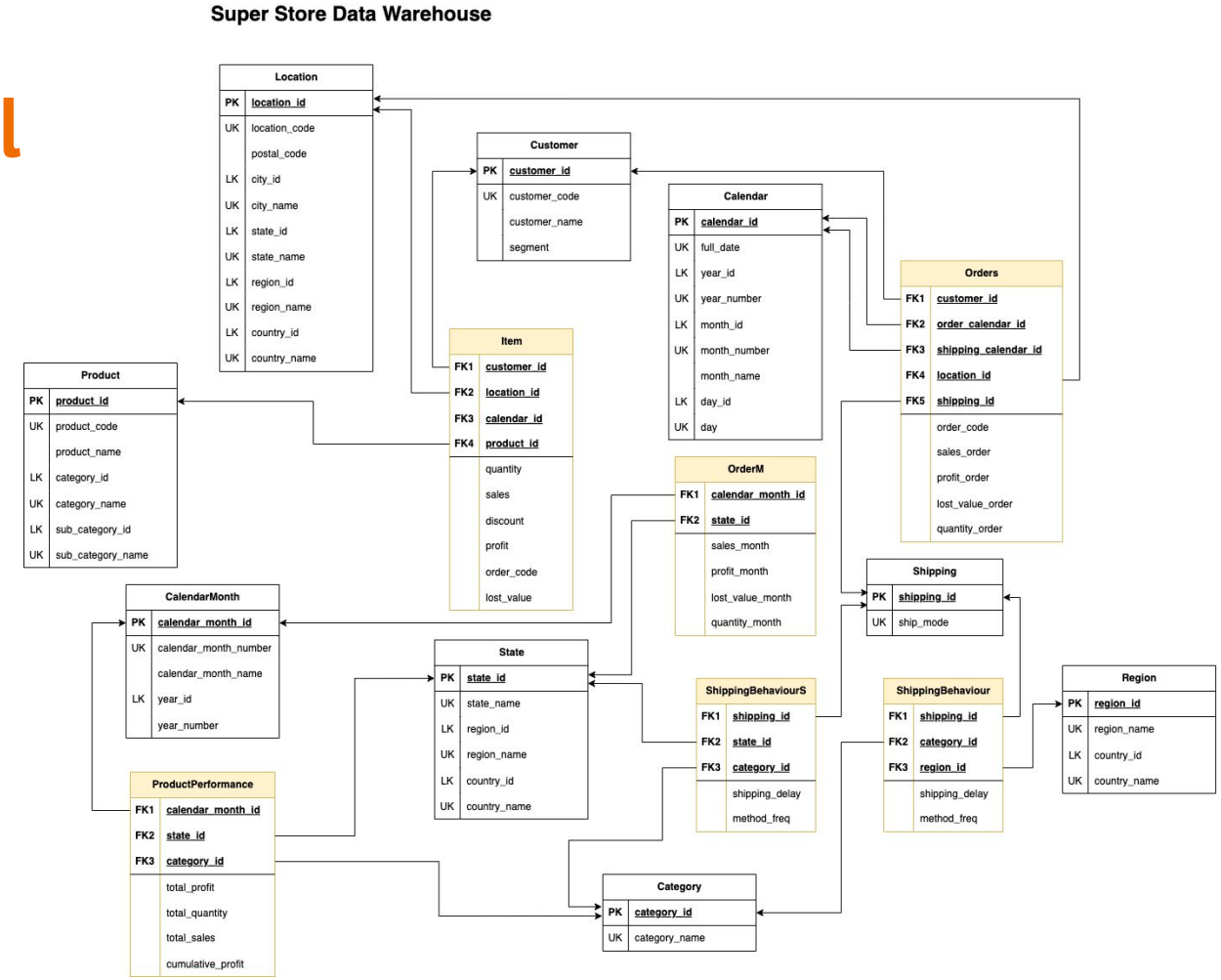
Dimensional Model

Fact Tables:

- Item
- Orders
- OrderM
- ShippingBehaviour
- ShippingBehaviourS
- ProductPerformance

Dimensions:

- Location
- Customer
- Calendar
- Product
- CalendarMonth
- State
- Shipping
- Category
- Region

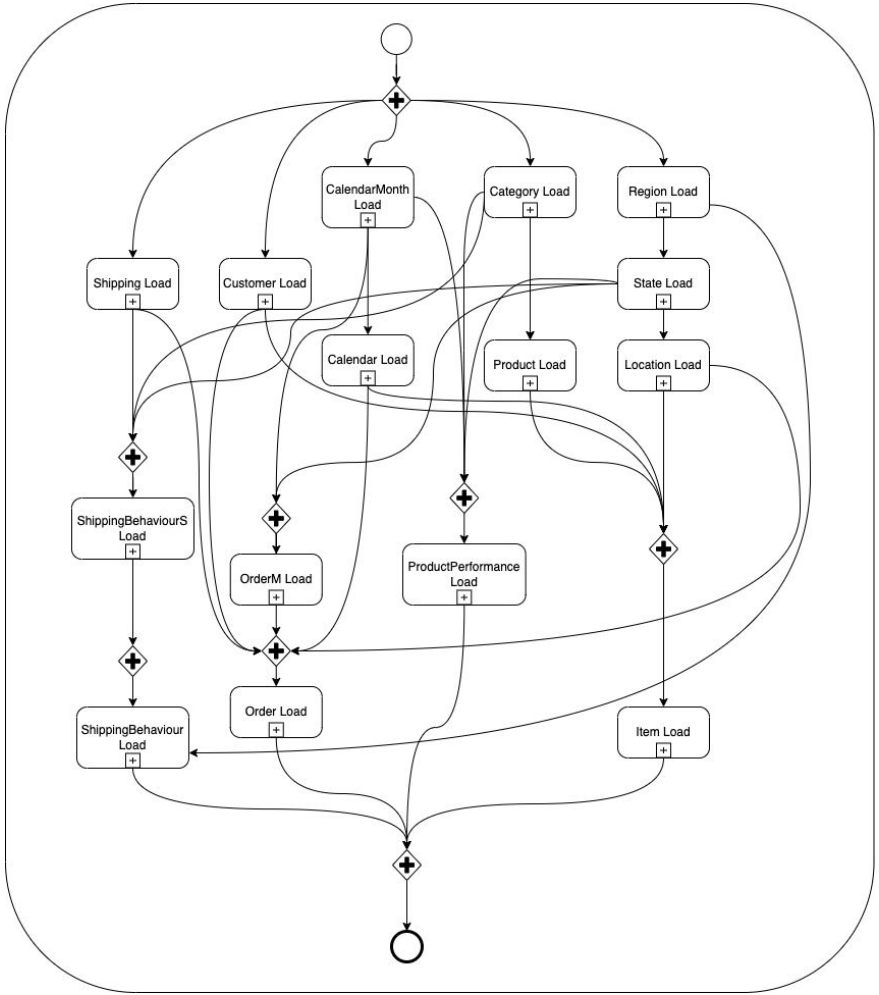


ETL Design - Overview

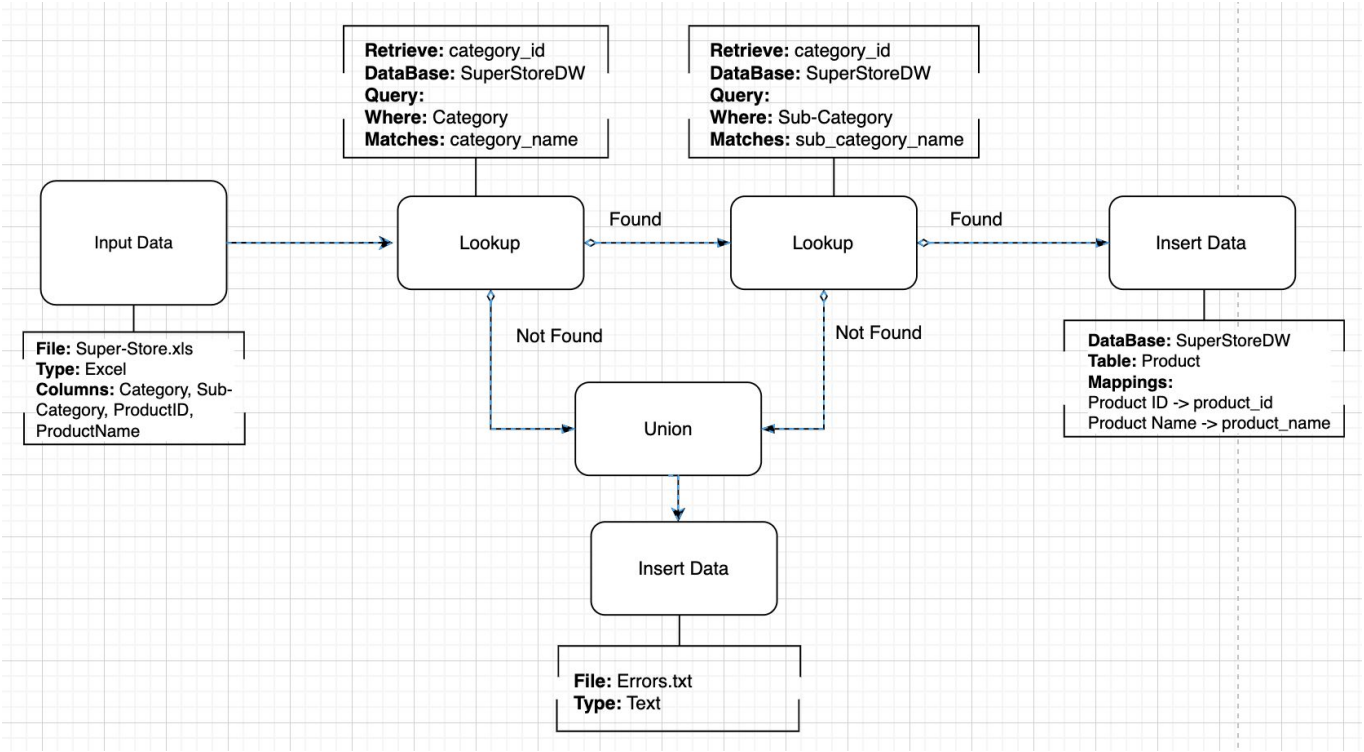
This diagram illustrates the load sequence for our Data Warehouse ETL.

Each box represents a table load, either dimensional (e.g., 'CalendarMonth Load') or fact (e.g., 'Order Load').

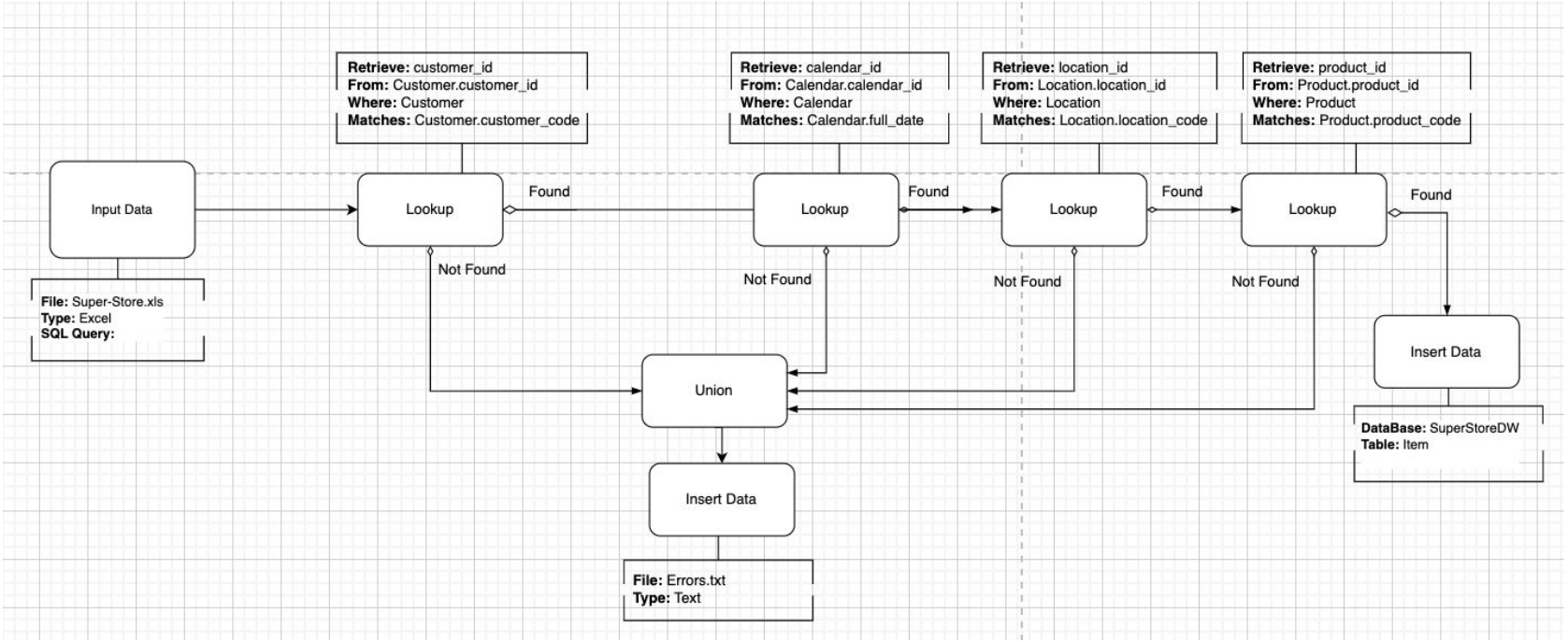
The arrows indicate dependency and the order in which tables are loaded, ensuring Data Warehouse integrity and efficiency for analysis.



ETL Design - Dimensions

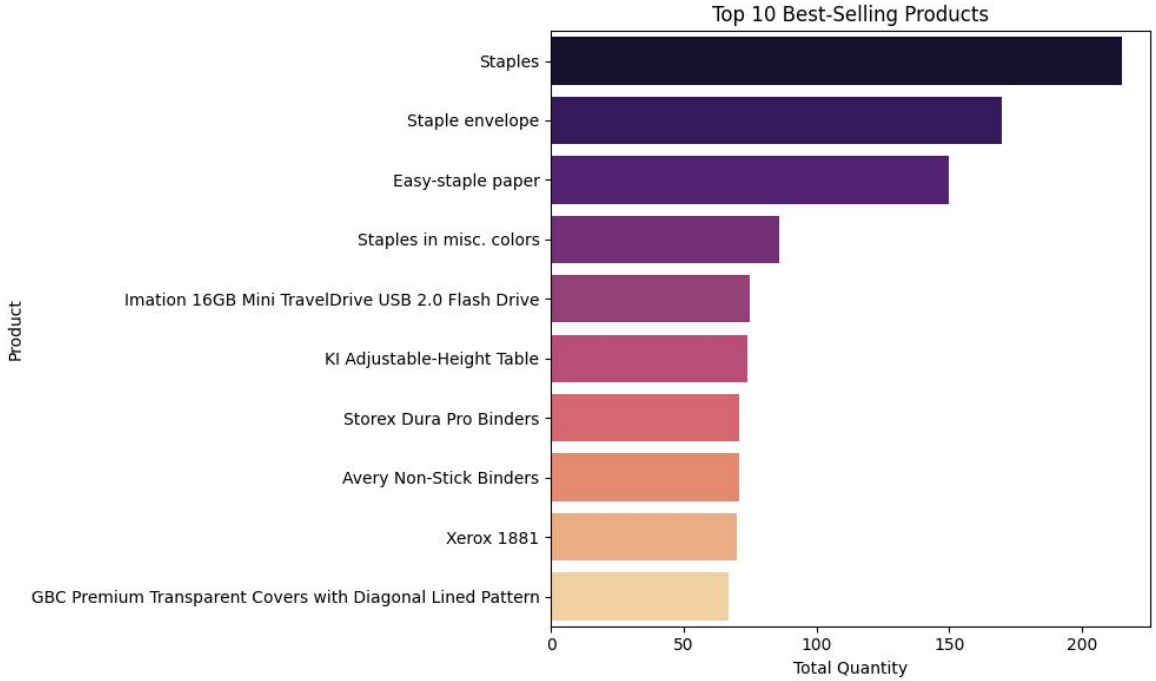


ETL Design - Fact Tables



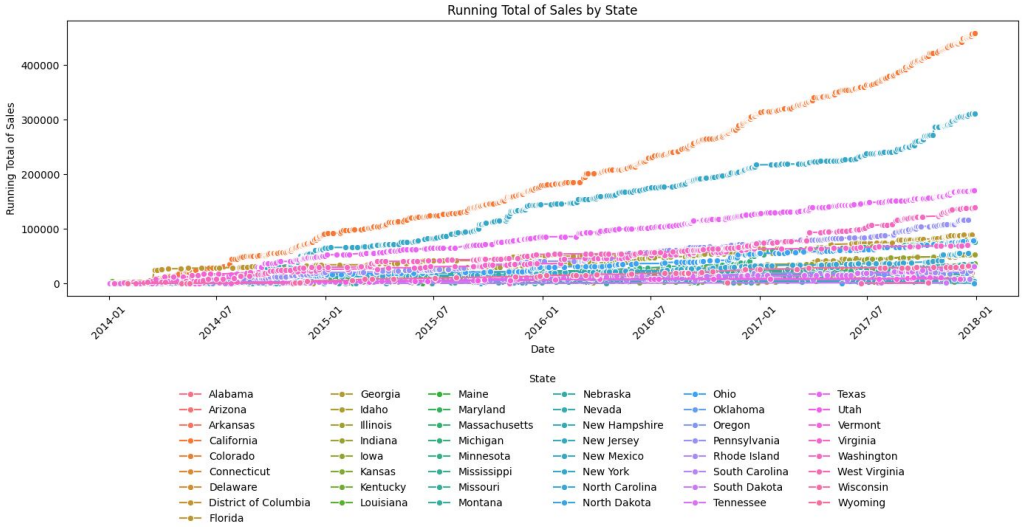
Queries & Results

```
SELECT
    p.product_name,
    SUM(i.quantity) AS total_quantity
FROM Item i
JOIN Product p ON i.product_id = p.product_id
GROUP BY p.product_name
ORDER BY total_quantity DESC
LIMIT 10;
```



Queries & Results

```
SELECT
    s.state_name,
    c.full_date,
    SUM(o.sales_order) OVER (PARTITION BY s.state_name
ORDER BY c.full_date) AS running_total
FROM Orders o
JOIN Location l ON o.location_id = l.location_id
JOIN State s ON l.state_id = s.state_id
JOIN Calendar c ON o.order_calendar_id = c.calendar_id
ORDER BY s.state_name, c.full_date;
```

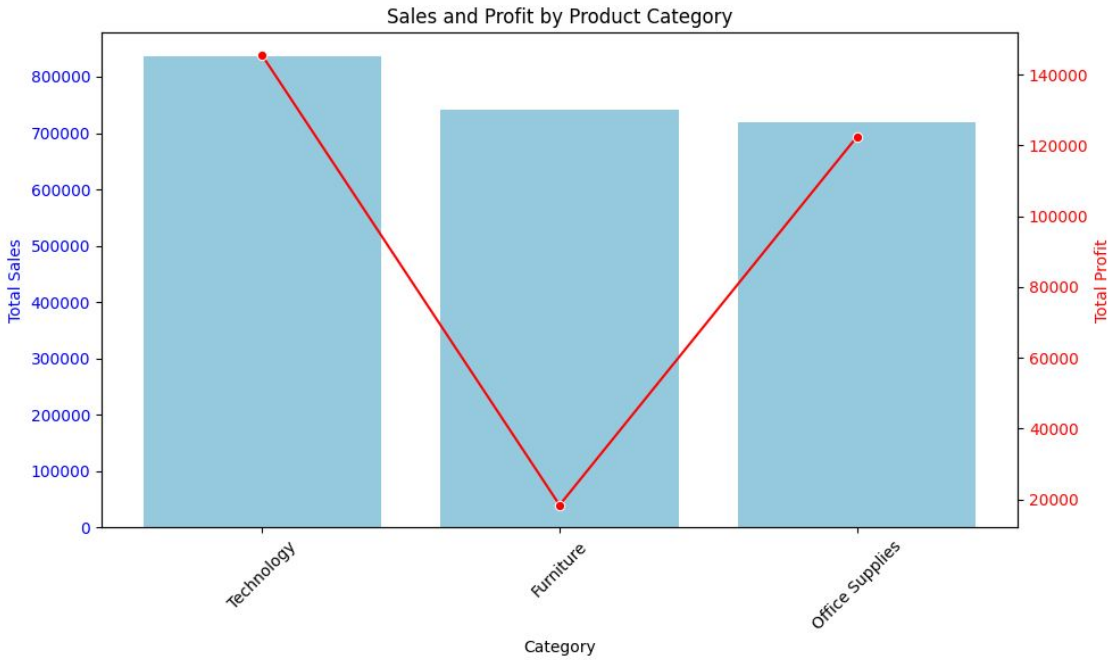


Queries & Results

```
SELECT
    sh.ship_mode,
    AVG(DATEDIFF(c2.full_date, c1.full_date)) AS average_delivery_time
FROM Orders o
JOIN Shipping sh ON o.shipping_id = sh.shipping_id
JOIN Calendar c1 ON o.order_calendar_id = c1.calendar_id
JOIN Calendar c2 ON o.shipping_calendar_id = c2.calendar_id
GROUP BY sh.ship_mode
ORDER BY average_delivery_time;
```

Queries & Results

```
SELECT
    cat.category_name,
    SUM(i.sales) AS total_sales,
    SUM(i.profit) AS total_profit,
    ROUND(AVG(i.discount), 2) AS average_discount
FROM Item i
JOIN Product p ON i.product_id = p.product_id
JOIN Category cat ON p.category_id = cat.category_id
GROUP BY cat.category_name
ORDER BY total_sales DESC;
```



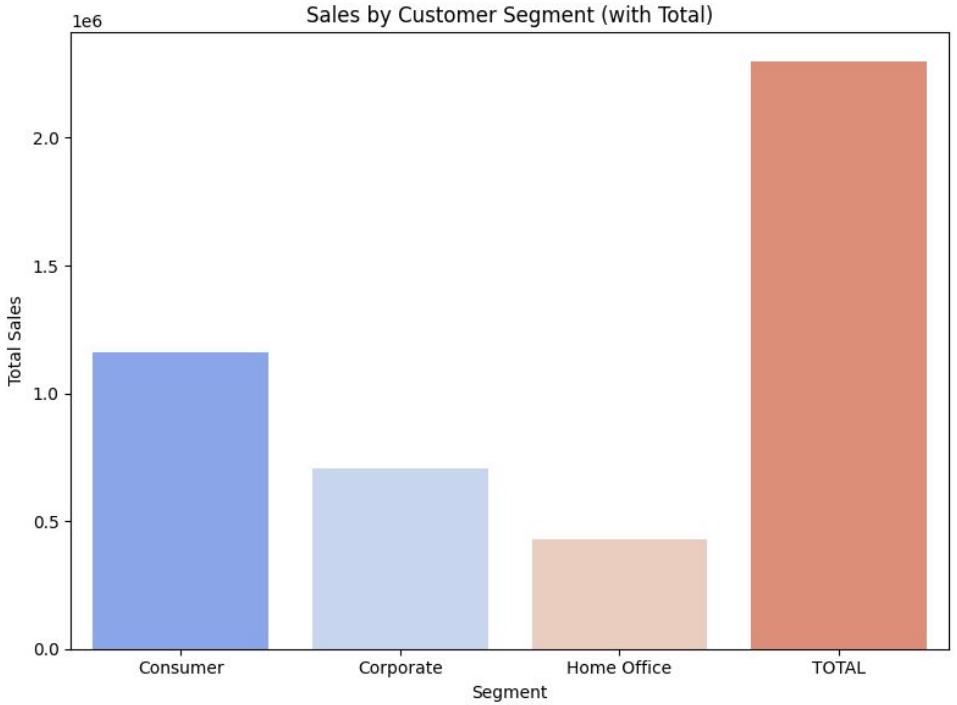
Queries & Results

```
SELECT
    cat.category_name,
    SUBSTRING_INDEX(
      GROUP_CONCAT(p.product_name ORDER BY i.sales DESC SEPARATOR ', '),
      ', ', 10) AS top_products, SUM(i.sales) AS total_sales
FROM Item i
JOIN Product p ON i.product_id = p.product_id
JOIN Category cat ON p.category_id = cat.category_id
GROUP BY cat.category_name;
```

	category_name	top_products	total_sales
0	Furniture	HON 5400 Series Task Chairs for Big and Tall, Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish, Chromcraft Bull-Nose Wood Oval Conference Tables & Bases, Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish, Sauder Forest Hills Library, Woodland Oak Finish, HON 5400 Series Task Chairs for Big and Tall, Bretford Rectangular Conference Table Tops	741999.98
1	Office Supplies	GBC Ibimaster 500 Manual ProClick Binding System, Ibico EPK-21 Electric Binding System, High Speed Automatic Electric Letter Opener, Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind, GBC DocuBind P400 Electric Binding System, Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind, High Speed Automatic Electric Letter Opener, Martin Yale Chadless Opener Electric Letter Opener, Ibico EPK-21 Electric Binding System, GBC DocuBind P400 Electric Binding System	719046.99
2	Technology	Cisco TelePresence System EX90 Videoconferencing Unit, Canon imageCLASS 2200 Advanced Copier, Canon imageCLASS 2200 Advanced Copier, Canon imageCLASS 2200 Advanced Copier, Canon imageCLASS 2200 Advanced Copier, 3D Systems Cube Printer, 2nd Generation, Magenta, HP Designjet T520 Inkjet Large Format Printer - 24" Color, Canon imageCLASS 2200 Advanced Copier	836154.10

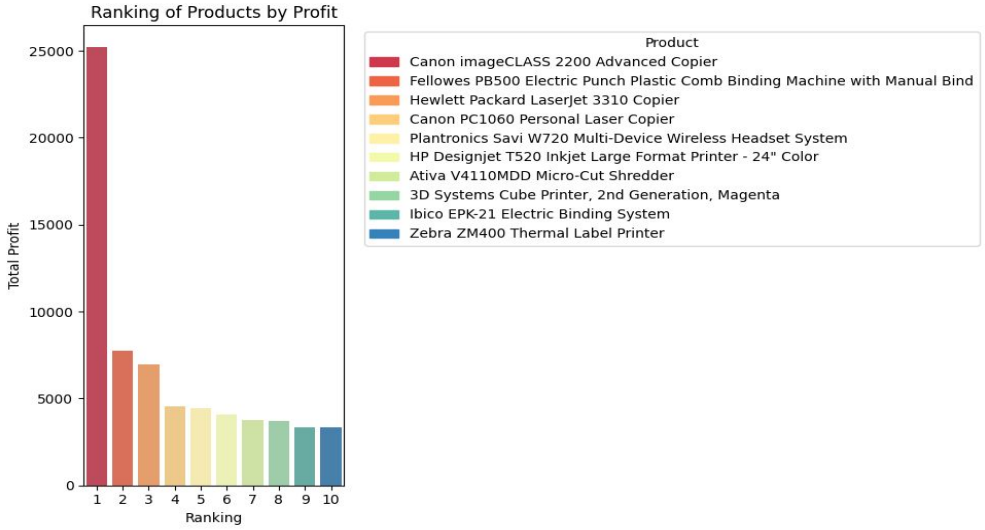
Queries & Results

```
SELECT

    IF(GROUPING(c.segment), 'TOTAL',
c.segment) AS segment, SUM(o.sales_order)
AS total_sales

FROM Orders o

JOIN Customer c ON o.customer_id =
c.customer_id

GROUP BY c.segment WITH ROLLUP;
```



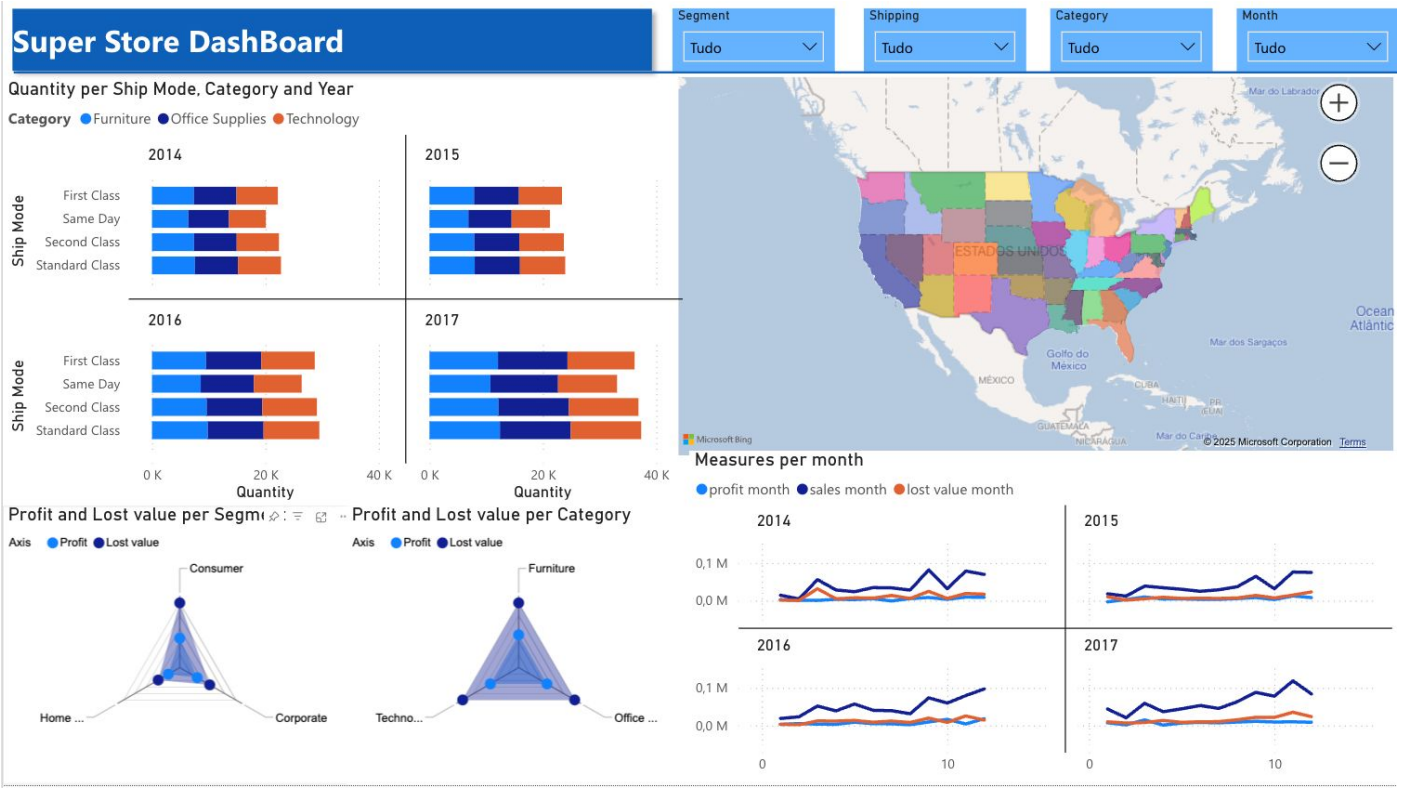
Queries & Results

```
SELECT
    product_name,  total_profit, RANK() OVER (ORDER BY total_profit DESC)
AS ranking
FROM (
    SELECT
        p.product_name, SUM(i.profit) AS total_profit
    FROM Item i
    JOIN Product p ON i.product_id = p.product_id
    GROUP BY p.product_name) AS sub
ORDER BY ranking
LIMIT 10;
```



Visualization

DashBoard



Advantages and Challenges

- Improved query performance compared to operational databases
- Enhanced data accessibility for users
- Support for complex analytical queries

- Data quality issues
- Complexity of the ETL process

Future Work

- The Super Store Data Warehouse serves as a strong foundation for future initiatives
- Potential for predictive analytics, automated ETL processes, and integration of diverse data sources to enhance decision-making
- Continuous evolution of data analysis with new tools and techniques

Conclusion

- Scalable Dimensional Model ensuring efficient data organization.
- Consistency & Documentation through the Bus Matrix and Data Dictionaries.
- Challenges Addressed in data quality and schema complexity.
- Future-Ready foundation for advanced analytics and business intelligence.

Super Store Data Warehouse

Created By: Bruno Fernandes, Hugo Abelheira, Tiago Coelho