

|  |
| --- |
| Business Template  **GLOBAL SUPERSTORE** |
| **Logo / Image** |

Contents

[1 Business Description 3](#_Toc106125647)

[1.1 Business background 3](#_Toc106125648)

[1.2 Problems because of poor data management 3](#_Toc106125649)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc106125650)

[2 Dimensions of a Business 3](#_Toc106125651)

[3 Logical Scheme 3](#_Toc106125652)

[4 Data Flow 3](#_Toc106125653)

[5 Fact Table Partitioning Strategy 3](#_Toc106125654)

# 

# Business Description

## Business background

During the COVID lockdown on-line shops selling office furniture, supplies and laptops/notepads/mobile phones and other technological products for home offices saw a surge in demand. Even after the lockdown was over, the demand for this kind of products has remained high, however the competition is fierce. Global Store is an on-line retail store providing office furniture, office supplies and technological products to a wide range of customers throughout the world. It has become increasingly important to analyze the demand pattern of the different type of customers, located in different parts of the world, to adjust the stock to be able to provide the customers with the products they need, while minimizing costs.

## Problems because of poor data management

Poor data management hinders the optimal product stock, leading to product unavailability in some regions and product surplus in others. The product unavailability in one region leads to longer delivery time, and that causes customer dissatisfaction. The product surplus can lead to increased costs of storing the products in the warehouse. To prevent these problems in the future, the sales data of 2022 and 2023 business years was collected.

## Benefits from implementing a Data Warehouse

Implementing a data warehouse for Global Superstore should address the problems described before. Implementing a data warehouse can answer, for example, the following questions:

* What customer segment generates the highest sales?
* What are the yearly total sales and profit figure in each market?
* What are the best months for sales for each customer segment?
* Which employee generated most of the sales each year?
* For each product category how is the demand distributed across geographical regions, and also across different market units?
* For each employee how is the demand distributed across geographical regions?
* Is there a correlation between certain product category demand and the sector that the customer is in?
* Is there a correlation between certain product demand and customer location?
* Which sector’s orders are mostly critical and in what months of the year?
* Are there certain cities where the demand is outstanding? For which products?
* In which regions the demand for same day delivery is the highest?
* Does order priority depend on the geographical region or customer segment?

## DATASETS DESCRIPTION

The first dataset contains the following information about sales to the consumer customer segment in 2022 and 2023.

Product Information:

Product Nr: The unique identifier of the product.

Product name: The name of the product that contains the name of the manufacturer as well.

Category Name: The category of the product (Furniture, Technology, Office supplies).

Sub-category Name: The sub-category of the product.

Employee information:

Employee FirstName: The first name of the employee dealing with the order.

Employee LastName: The last name of the employee dealing with the order.

Employee Nr: The identifier of the employee dealing with the order.

Sales Information:

Order Number: The unique identifier of the order.

Order Date: The date of the order.

Ship date: The date of shipment.

Sales: The actual selling amount in US $.

Quantity: The number of units sold.

Discount: The discount rate of the order.

Profit: The profit amount realized by Global superstore on the order.

Shipping cost: Cost associated with the delivery.

Customer Information:

Customer Nr: The unique identifier of the customer.

Customer name: The name of the customer.

Address: The address of the customer (street name and number)

Gender: Gender of the customer (M/F)

Age group: the age interval that the customer belongs to.

Email: The email address of the customer.

Segment: All the customers in this dataset are of consumer segment.

Geographical information:

City: The city of the delivery.

State: The state of the delivery.

Country: The country of the delivery.

Postal Code: The postal code of the delivery.

Additional Attributes:

Ship mode: The type of shipment (Same day, First class, Second class, Standard class).

Market: The market distribution unit (US, Canada, EU, EMEA, APAC, Africa, LATAM)

Order priority: The priority of the order delivery (Critical, High, Medium, Low).

The second dataset contains the following information about sales to the corporate and home office customer segment in 2022 and 2023.

Product Information:

Product ID: The unique identifier of the product.

Product name: The name of the product that contains the name of the manufacturer as well.

Category: The category of the product (Furniture, Technology, Office supplies).

Sub-category: The sub-category of the product.

Employee information:

Employee Name: The name (first and last) of the employee dealing with the order.

Employee ID: The unique identifier of the employee dealing with the order.

Sales Information:

Order ID: The unique identifier of the order.

Order Date: The date of the order.

Ship date: The date of shipment.

Sales: The actual selling amount in US $.

Quantity: The number of units sold.

Discount: The discount rate of the order.

Profit: The profit amount realized by Global superstore on the order.

Shipping cost: Cost associated with the delivery.

Customer Information:

Customer ID: The unique identifier of the customer.

Tax number: The tax number of the customer (firm).

Sector: The sector (industry) in which the customer operates (banking, manufacturing, insurance, hospitality)

Contact person: The name of the contact person of then customer.

Shipping address: The address of the shipment (street name and number).

Email: The email address of the customer (contact person).

Segment: The customers in this dataset are corporate or home office.

Geographical information:

City: The city of the delivery.

State: The state of the delivery.

Country: The country of the delivery.

Postal Code: The postal code of the delivery.

Region: The region of the delivery.

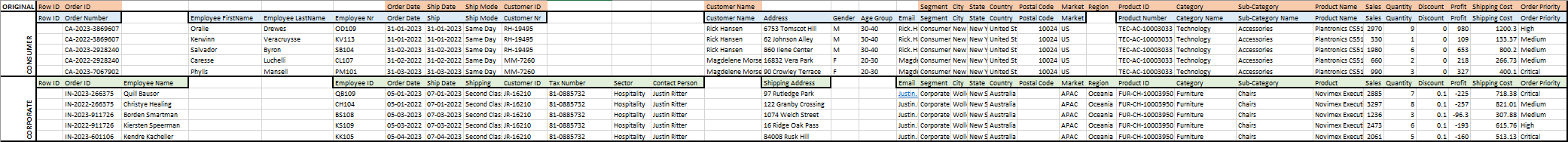
Additional Attributes:

Ship mode: The type of shipment (Same day, First class, Second class, Standard class).

Market: The market distribution unit (US, Canada, EU, EMEA, APAC, Africa, LATAM)

Order priority: The priority of the order delivery (Critical, High, Medium, Low).

The differences between the first data source compared with the second data source: the first data source has Customer Name column, and the employee name is split in two columns: Employee First Name, Employee Last Name; The second data source has Tax number, Sector and contact person, for the first data set there is customer name, gender and age group column; For the geographical location the first data source has a region field.



## GRAIN / DIM / FACT

The grain of my dataset is: order, order date, product, employee, customer, shipping address, market, order priority and ship mode, that means that this is the level of information that is represented by a single record in the fact table.

Dimensions:

Dim Employee

Shows data of the employee

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Employee ID | The primary key of this table | Bigint |
| Employee NR | The employee ID of the employee dealing with the order. | Text |
| employee FirstName | The first name of the employee dealing with the order. | Text |
| employee LastName | The last name of the employee dealing with the order. | Text |

Example with filled data

|  |  |  |  |
| --- | --- | --- | --- |
| Employee ID | Employee NR | employee FirstName | employee LastName |
| 1 | OD109 | Oralie | Drewes |

Dim Customer

Shows data of the customer

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Customer ID | The primary key of this table | Bigint |
| Customer NR | The unique identifier of the customer. | Text |
| Tax Number | The tax number of the customer (firm). | Text |
| Sector | The sector (industry) in which the customer operates (banking, manufacturing, insurance, hospitality) | Text |
| Contact person | The name of the contact person of then customer. | Text |
| Customer Name | The name of the customer if the customer is of consumer type | Text |
| Email | The email address of the customer. | Text |
| Gender | The gender of the customer (M/F) | Varchar(1) |
| Age Group | the age interval that the customer belongs to. | Text |
| Segment | The customer segment can be : consumer, home office or corporate | Text |

Example with filled data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer ID | Customer NR | Tax Number | Sector | Contact person | Shipping address | Gender | Age Group | Email | Segment |
| 1 | JR-16210 | 81-0885732 | Hospitality | Justin Ritter | 97 Rutledge Park | M | 30-40 | Justin.Ritter@email.com | Corporate |

Dim Address

Shows the address to where the order has been sent to

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Address ID | The primary key of this table | Bigint |
| Shipping address | The address of the shipment (street name and number). | Text |
| City | The city of the delivery. | Text |
| State | The state of the delivery. | Text |
| Country | The country of the delivery. | Text |
| Postal Code | The postal code of the delivery. | Int |
| Region | The region of the delivery. | Text |

Example with filled data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Address ID | Shipping Address | City | State | Country | Postal Code | Region |
| 1 | 97 Rutledge Park | New York City | New York | United States | 10024 | US |

Dim Product

Shows the product that the customer has bought

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Product ID | The primary key of this table | Bigint |
| Product NR | The unique identifier of the product. | Text |
| Product name | The name of the product that contains the name of the manufacturer as well. | Text |
| Category | The category of the product (Furniture, Technology, Office supplies). | Text |
| Sub-category | The sub-category of the product. | Text |

Example with filled data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Product ID | Product NR | Product name | Category | Sub-category |
| 1 | FUR-CH-10003950 | Novimex Executive Leather Armchair, Black | Furniture | Chairs |

Dim Order

Shows the Order specifications

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Order ID | The primary key of this table | Bigint |
| Order Number | The unique identifier of the order. | Text |
| Ship date | The date of shipment. | Date |
| Order priority | The priority of the order delivery (Critical, High, Medium, Low). | Text |
| Ship Mode | The type of shipment (Same day, First class, Second class, Standard class). | Text |

Example with filled data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order ID | Order Number | Ship date | Order priority | Ship Mode |
| 1 | IN-2023-266375 | 07-01-2023 | High | Same Day |

Dim Date

Shows the date of the order

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Date ID | The primary key of the date table | Bigint |
| Order Date | The date of the order. | Date |

Example with filled data

|  |  |
| --- | --- |
| Date ID | Order Date |
| 1 | 07-01-2023 |

Dim Market

Shows the market distribution unit

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Market ID | The primary key of this table | Bigint |
| Market | The market distribution unit (US, Canada, EU, EMEA, APAC, Africa, LATAM) | Text |

Example with filled data

|  |  |
| --- | --- |
| Market ID | Market |
| 1 | APAC |

Fact table

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| FK\_Employee ID | Foreign key to the employee dimension table | Bigint |
| FK\_Customer ID | Foreign key to the customer dimension table | Bigint |
| FK\_Shipping address | Foreign key to the address dimension table | Bigint |
| FK\_Product ID | Foreign key to the product dimension table | Bigint |
| FK\_Order ID | Foreign key to the employee dimension table | Bigint |
| FK\_Market ID | Foreign key to the market dimension table | Bigint |
| FK\_Order Date | Foreign key to the date dimension | Bigint |
| Sales | The actual selling amount in US $. This is an additive measurement. | Float |
| Quantity | The number of units sold. This is an additive measurement. | Int |
| Discount | The discount rate of the order. This is a non-additive measurement. | Float |
| Profit | The profit amount realized by Global superstore on the order. This is an additive measurement. | Float |
| Shipping cost | Cost associated with the delivery. This is an additive measurement. | Float |

# Business Layer 3NF

In the BL\_3NF layer I defined 14 dimension tables. Three of the dimension tables are slowly changing dimensions 2 (SCD2) because if any of the data in these dimension changes except the Source\_id, the new row with the same id generated from a sequence is inserted having the START\_DT the date of change, the END\_DT 9999.12.31 and IS\_ACTIVE is set to true, in the old row with the same id the END\_DT is updated to the date of the change and IS\_ACTIVE is set to false. The three SCD2 are: Dim\_Products\_SCD2, Dim\_Customers\_SCD2 and Dim\_Employees\_SCD2. In the slowly changing dimensions type 2 the primary key is a composite key that consists of bigint number generated from a sequence and the START\_DT.

The rest of the dimension tables except Dim\_Date are SCD1, that means that any change in a row (except source triplets) will overwrite the previous data. Dim Date is an SCD0 so the dimension doesn’t change.

In every dimension table there is the source triplet:

* Source\_system - ‘Globalstore\_sales’
* Source\_entity - ‘Globalstore\_consumer’, ‘Globalstore\_corporate\_home\_office’
* Source\_id- for example ‘JR-16210’ (Customer NR), ‘QB109’ (Employee NR), ‘TEC-AC-10003033’ (Product Number), ‘APAC’ (Market), ‘Furniture’ (Category), ‘Chairs’ (Sub-category), ‘CA-2023-3869607’ (Order Number), etc.

Each dimension has a surrogate key that gets values from sequences.

|  |  |
| --- | --- |
| Dim\_Products\_SCD2 | |
| Column name | Data Type |
| PK\_Product ID | Bigint |
| FK\_Sub-category ID | bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Product NR | Text |
| Product name | Text |
| Category | Text |
| START\_DT | date |
| END\_DT | date |
| IS\_ACTIVE | boolean |

|  |  |
| --- | --- |
| Dim\_Customers\_SCD2 | |
| Column name | Data Type |
| PK\_Customer ID | Bigint |
| FK\_Sector ID | Bigint |
| FK\_Segment ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Customer NR | Text |
| Tax Number | Text |
| Contact person | Text |
| Customer Name | Text |
| Email | Text |
| Gender | Varchar(1) |
| Age Group | Text |
| START\_DT | date |
| END\_DT | date |
| IS\_ACTIVE | boolean |

|  |  |
| --- | --- |
| Dim\_Employees\_SCD2 | |
| Column name | Data Type |
| PK\_Employee ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Employee NR | Text |
| employee FirstName | Text |
| employee LastName | Text |
| START\_DT | date |
| END\_DT | date |
| IS\_ACTIVE | boolean |

|  |  |
| --- | --- |
| Dim\_Sectors | |
| Column name | Data Type |
| PK\_Sector ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Sector | Text |

|  |  |
| --- | --- |
| Dim\_Segments | |
| Column name | Data Type |
| PK\_Segment ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Segment | Text |

|  |  |
| --- | --- |
| Dim\_Date | |
| Column name | Data Type |
| PK\_Date ID | Bigint |
| Order Date | Date |

|  |  |
| --- | --- |
| Dim\_Markets | |
| Column name | Data Type |
| PK\_Market ID | Bigint |
| Market | Text |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |

|  |  |
| --- | --- |
| Dim\_Orders | |
| Column name | Data Type |
| PK\_Order ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Order Number | Text |
| Ship date | Date |
| Order priority | Text |
| Ship Mode | Text |

|  |  |
| --- | --- |
| Dim\_Addresses | |
| Column name | Data Type |
| PK\_Address ID | Bigint |
| FK\_Country | Bigint |
| Shipping address | Text |
| City | Text |
| State | Text |
| Postal Code | Int |
| Region | Text |

|  |  |
| --- | --- |
| Dim\_Countries | |
| Column name | Data Type |
| PK\_Country ID | Bigint |
| FK\_Region ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Country | Text |
| State | Text |

|  |  |
| --- | --- |
| Dim\_Regions | |
| Column name | Data Type |
| PK\_Region ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Region | Text |

|  |  |
| --- | --- |
| Dim\_Sub-categories | |
| Column name | Data Type |
| PK\_Sub-category ID | Bigint |
| FK\_Category ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Sub-category | Text |

|  |  |
| --- | --- |
| Dim\_Categories | |
| Column name | Data Type |
| PK\_Category ID | Bigint |
| Source\_system | varchar |
| Source\_entity | varchar |
| Source\_id | varchar |
| Category | Text |

The fact table has foreign keys to the following dimension tables: Dim\_Employees\_SCD2, Dim\_Customers\_SCD2, Dim\_Products\_SCD2, Dim\_Addresses, Dim\_Markets, Dim\_Orders, Dim\_Date. The measurements in the fact table are: Sales, Quantity, Discount, Profit.

|  |  |
| --- | --- |
| FCT\_ORDERS\_DD | |
| Column name | Data Type |
| FK\_Employee ID | Bigint |
| FK\_Customer ID | Bigint |
| FK\_Shipping address | Bigint |
| FK\_Product ID | Bigint |
| FK\_Order ID | Bigint |
| FK\_Market ID | Bigint |
| FK\_Order Date | Bigint |
| Sales | Float |
| Quantity | Int |
| Discount | Float |
| Profit | Float |
| Shipping cost | Float |

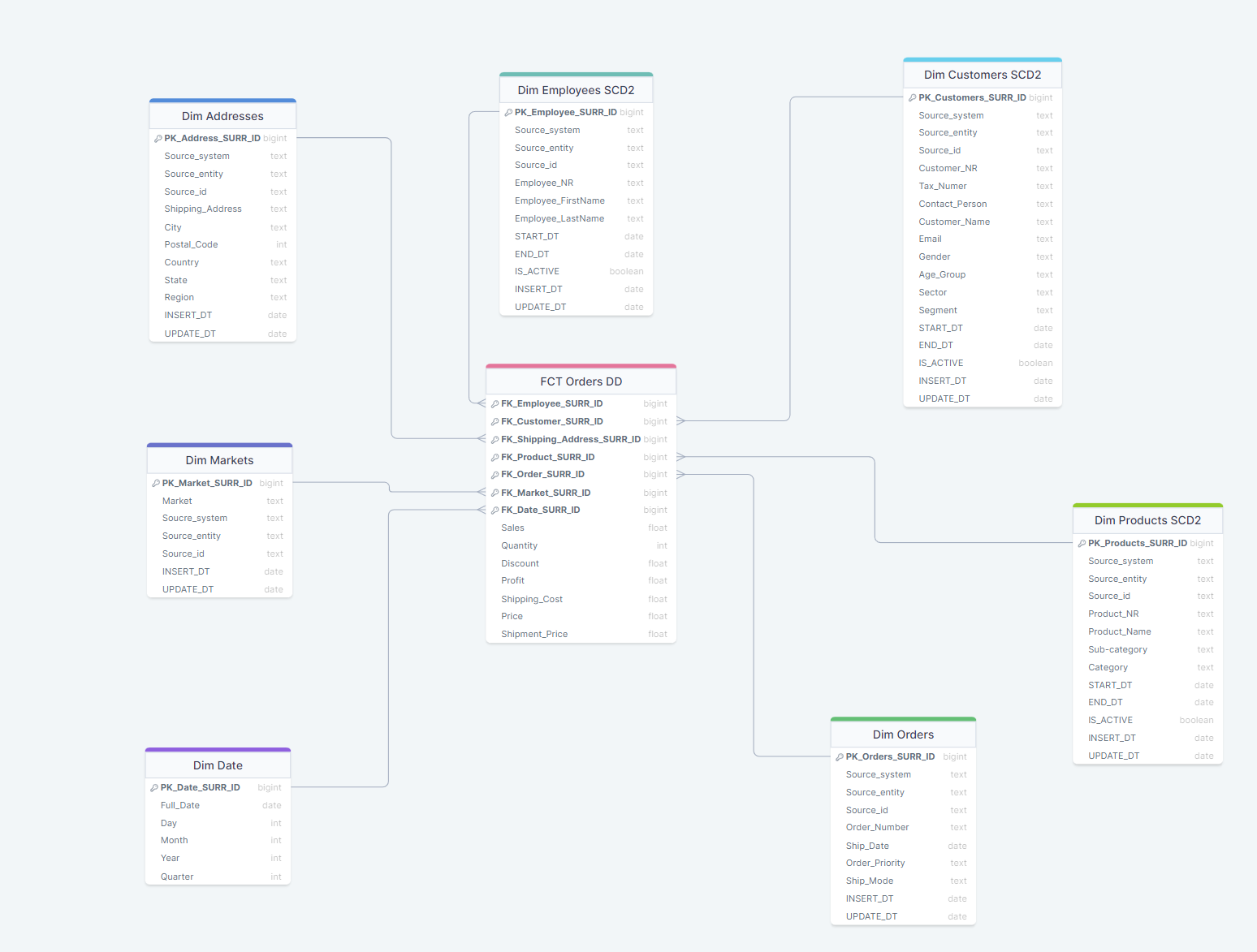
# Business Layer Dimensional Model

I created the BL\_DM from the BL\_3NF schema. The BL\_DM Schema is a star schema. I denormalized three of the dimension from BL\_3NF. These three dimensions have hierarchical attributes. These three dimensions are Dim\_Employees\_SCD2, Dim\_Customers\_SCD2 and Dim\_Products\_SCD2, these are also slowly changing dimensions type 2. The Primary key in these dimensions is a single bigint number generated from a sequence for each, for example PK\_Employee\_SURR\_ID (meaning it is a surrogate key). In the BL\_DM the primary key is not a composite key. The source\_id in the dimension tables originate from the number in the composite primary key of the dimensions in BL\_3NF, of course, converted to varchar2. The main difference between the dimension in BL\_DM and BL\_3NF is that in BL\_DM there is no composite key. All the other attributes of the dimensions in the BL\_DM are inherited from the BL\_3NF. The SCD2 dimensions have also the START\_DT, END\_DT and IS\_ACTIVE attributes.

All the dimensions have the attribute INSERT\_DT and UPDATE\_DT – the dates when the record was inserted or updated, respectively.

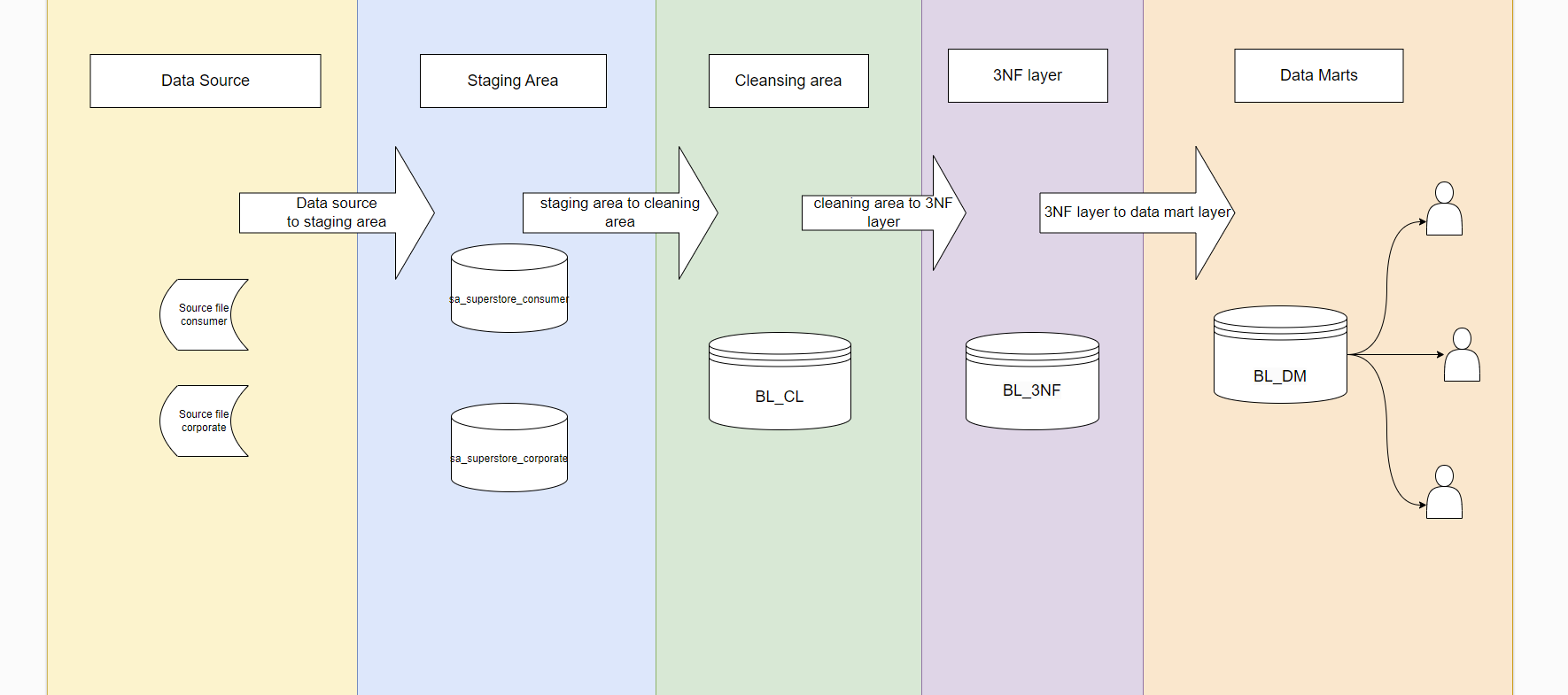
Since the Dim Date dimension is a unique and special dimension, it is not be populated from the normalized layer like other dimensions but I created a script for it to populate with all the dates from 2022-01-01 to 2024-12-31.

The fact table has foreign keys to the dimensions: Dim Employee SCD2, Dim Customer SCD2, Dim Products SCD2, Dim Orders, Dim Addresses, Dim Markets and Dim Date. The measurements in the fact table are sales quantity, discount, profit, shipping cost- these are from the BL\_3NF, and two new measurements calculated in this layer: price calculated as sales/quantity and shipment price calculated as shipping cost/quantity. The primary key of the fact table is a composite key made up of the foreign keys to the dimension tables.



# Logical Scheme

Below there is the logical schema of the data load process.



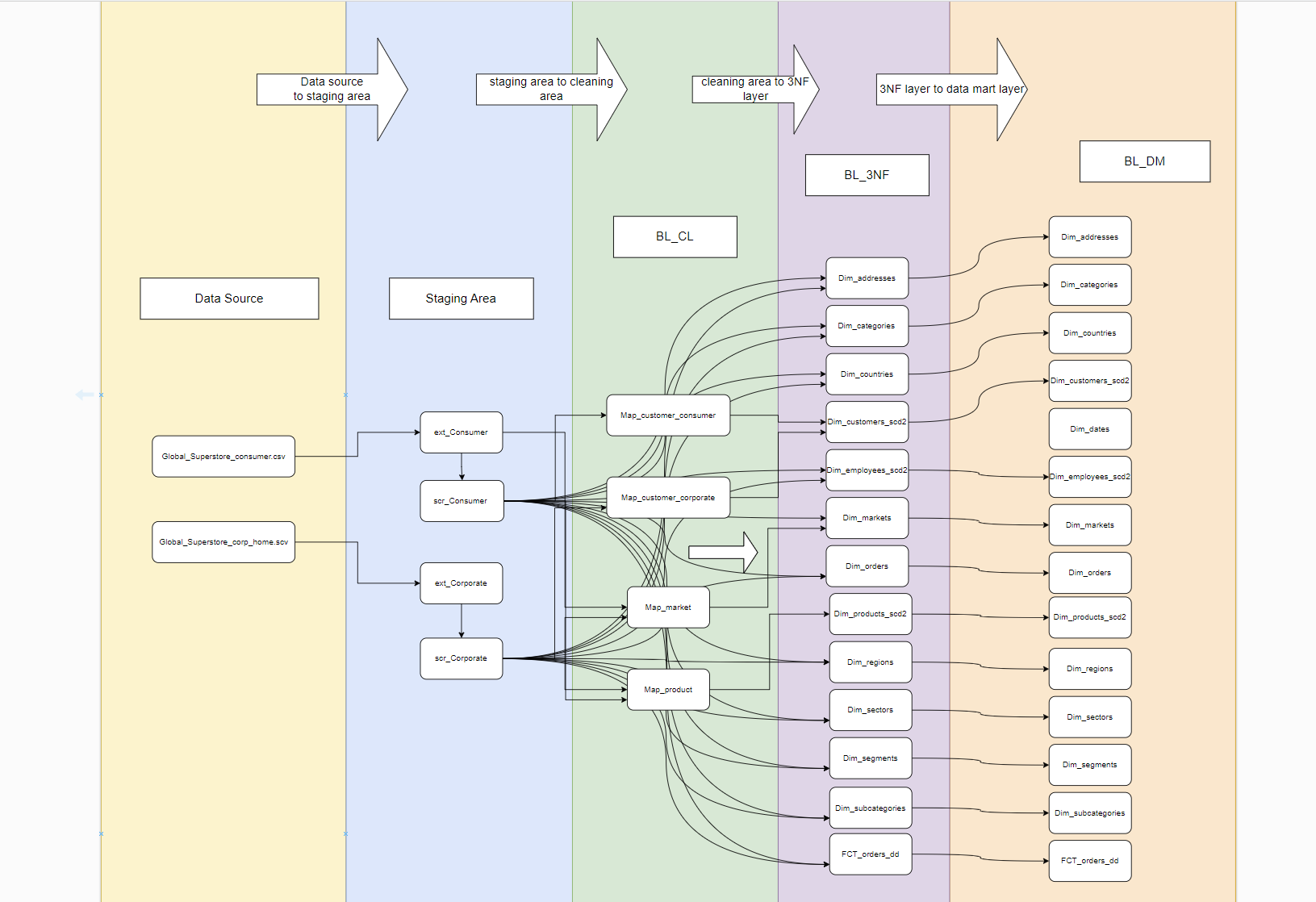
# Data Flow

The diagram below reflects the data flow from the source to the data consumers. From the source files the data is loaded in the staging area where there are two separate schemas for the two sources: sa\_superstore\_consumer and sa\_superstore\_corporate. In each of these schemas there is a foreign table defined for accessing the external table and a table defined for loading the data unchanged from the external tables.

There is a cleaning area, the BL\_CL schema, used for deduplication and data cleaning. The customer, product and market data is cleaned and loaded into map\_customer\_consumer, map\_customer\_corporate, map\_product and map\_market respectively.

In the 3NF layer, BL\_3NF schema, there are the dimension tables and the fact table. 3 of the dimensions are scd2 type: Dim\_products\_scd2, Dim\_customer\_scd2, Dim\_employee\_scd2. the rest of the dimension tables are scd1 type. The customer markets and products are loaded using the map tables from the BL\_CL.

The data mart layer, BL\_DM schema, is a snowflake schema in which the data in the dimensions is loaded from the dimensions in the layer BL\_3NF and the fact is loaded from the fact table in BL\_3NF. There is a special dimension, dim\_dates loaded with the calendar dates and day, month, year and quarter.



# Fact Table Partitioning Strategy