I



|  |
| --- |
| Business Template  **ADIDAS US SALES** |
|  |

**Contents**

[1 Business Description 3](#_1fob9te)

[1.1 Business background 3](#_3znysh7)

[1.2 Problems because of poor data management 3](#_2et92p0)

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

[1.4 DATASETS DESCRIPTION 3](#_3dy6vkm)

[1.5 GRAIN / DIM / FACT 5](#_1t3h5sf)

[2 Business Layer 3NF 5](#_4d34og8)

[3 Business Layer Dimensional Model 5](#_2s8eyo1)

[4 Logical Scheme 5](#_17dp8vu)

[5 Data Flow 5](#_3rdcrjn)

[6 Fact Table Partitioning Strategy 5](#_26in1rg)

# 

# Business Description

## Business background

Adidas, a renowned global leader in the sportswear industry, has a significant presence in the United States market. With a diverse range of products catering to athletes and enthusiasts alike, Adidas has established a strong brand identity synonymous with innovation, quality, and performance.

In the competitive landscape of the sportswear industry, staying ahead requires comprehensive insights into consumer preferences, market trends, and sales performance. Effective data management is crucial for Adidas US Sales to optimize operations, enhance decision-making processes, and drive sustainable growth.

## Problems because of poor data management

However, despite its prominence, Adidas US Sales faces challenges stemming from poor data management practices. Inadequate data collection, disparate data sources, and inconsistent data quality hinder the organization's ability to extract meaningful insights. These problems lead to:

1. Inaccurate forecasting: Without access to reliable historical data and real-time insights, Adidas struggles to accurately forecast demand, resulting in inventory imbalances and missed sales opportunities.
2. Inefficient marketing strategies: Limited visibility into consumer behavior and market trends impedes the development of targeted marketing campaigns, reducing the effectiveness of promotional efforts and brand engagement.
3. Suboptimal inventory management: Poor data integration across supply chain processes leads to inefficiencies in inventory management.
4. Hindered decision-making: Decision-makers lack access to timely and relevant information, resulting in delayed responses to market changes and competitive pressures.

## Benefits from implementing a Data Warehouse

To address these challenges and unlock the full potential of its data assets, Adidas US Sales aims to implement a comprehensive data warehouse solution. By centralizing data from various sources and establishing robust data governance practices, the organization anticipates several benefits:

1. Enhanced business intelligence: A data warehouse enables Adidas to consolidate and analyze vast volumes of data efficiently, providing actionable insights into sales performance, consumer behavior, and market trends.

2. Improved forecasting accuracy: Access to comprehensive historical data and advanced analytics tools empowers Adidas to develop more accurate demand forecasts and optimize inventory levels.

3. Targeted marketing campaigns: By leveraging customer segmentation and predictive analytics, Adidas can tailor marketing strategies to specific audience segments, increasing campaign relevance and driving higher conversion rates.

4. Empowered decision-making: Real-time access to reliable data enables agile decision-making, allowing Adidas to respond swiftly to market dynamics, capitalize on emerging opportunities, and mitigate risks effectively.

In conclusion, the implementation of a data warehouse represents a strategic investment for Adidas US Sales, positioning the organization for sustained growth and competitive advantage in the dynamic sportswear market. By harnessing the power of data, Adidas can unlock new insights, optimize operations, and deliver exceptional experiences to its customers.

## DATASETS DESCRIPTION

*We are working with 2 datasets: one for in-store sales and one for online sales. Each dataset shows individual sales transactions, including product info, customer info, and how/where the sale happened.*

### **In-Store Sales Dataset**

This dataset captures sales made at physical retail locations. Each record represents a sales transaction of a product sold to a customer at a specific store and location.

**Fields include:**

* **Customer Info:** Customer ID, First Name, Last Name
* **Product Info:** Product, Category, Price per Unit
* **Retailer Info:** Retailer, Region, State, City
* **Sales Info:** Units Sold, Total Sales, Operating Profit, Margin
* **Transaction Info:** Invoice Date, Sales Method (always “In-store”), Payment Method

### **Online Sales Dataset**

This dataset captures sales performed through online platforms (e.g., Adidas.com). Each row represents a sales transaction completed online.

**Fields include:**

* **Customer Info:** Customer ID, First Name, Last Name
* **Product Info:** Product, Category, Price per Unit
* **Retailer Info:** Online Retailer (e.g., Adidas.com)
* **Sales Info:** Units Sold, Total Sales, Operating Profit, Margin
* **Transaction Info:** Invoice Date, Sales Method (always “Online”), Payment Method

### Key Entities (Future dimension)

### **In-Store Dataset: Dimensions**

|  |  |
| --- | --- |
| **Dimension** | **Columns** |
| **Customer** | Customer ID, First Name, Last Name |
| **Product** | Product, Product ID, Price per Unit |
| **Retailer** | Retailer, Retailer ID |
| **Location** | Region, Region ID, State, State ID, City, City ID |
| **Product Category** | Product Category ID, Category |
| **Junk** | Payment Method, Sales Method |
| **Date** | Invoice date (will be broken into Date dimension: day, month, year, etc.) |

### **Online Dataset: Dimensions**

|  |  |
| --- | --- |
| **Dimension** | **Columns** |
| **Customer** | Customer ID, First Name, Last Name |
| **Product** | Product, Product ID, Price per Unit |
| **Retailer** | Retailer (Adidas.com), Retailer ID |
| **Product Category** | Product Category ID, Category |
| **Junk** | Payment Method, Sales Method |
| **Date** | Invoice date(Time) |

### Differences Between the Two Datasets

The main differences between the in-store and online sales datasets are:

* **Sales**: Online sales only involve internet-based platforms like Adidas.com; in-store involves physical retailers.
* **Location data**: In-store dataset includes geographic details (city, state, region), while online dataset lacks that.
* **Retailer type**: Online sales are mostly via Adidas.com; in-store includes multiple retail brands.

## GRAIN / DIM / FACT

### Dimensional Design Process

#### Business process

We are analyzing **Adidas sales transactions, so t**his includes in-store and online sales data.

#### Grain

In both datasets, **each row represents one sales transaction**. So:

**Grain:** One sale of a specific product to a specific customer on a specific date and time, by a specific retailer, using a specific payment method.

#### Dimensions

The dimension tables we can extract from both datasets:

|  |  |  |
| --- | --- | --- |
| **Dimension** | **Description** | **Example Columns** |
| **Customer** | Who bought the product | customer\_id, first\_name, last\_name |
| **Product** | What product was sold | product\_id, name, price |
| **Retailer** | Where it was sold (store/website) | retailer\_id, name |
| **Location** | Where in the world | region, state, city + their IDs |
| **Product Category** | Product Category ID, Category | Men, women |
| **Junk** | Payment Method, Sales Method | Cash or Card, Online or Instore |
| **Date** | When it was sold | invoice\_date (break into day/month/year) |
| **Time** | When it was sold in exact time | time\_id, hour, minute, time\_of\_day |

#### Facts

|  |  |
| --- | --- |
| **Fact Table** | **Fact Fields** |
| **FCT\_SALES** | units\_sold, total\_sales, operating\_profit, operating\_margin |

* Units Sold → how many items were sold
* Operating Profit → how much Adidas earned after subtracting costs
* Operating Margin → what % of the sales is profit

### Table descriptions:

#### FCT\_SALES

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| sale\_id | Unique identifier for each sale | INT / BIGINT |
| date\_id | Reference to date dimension | INT |
| customer\_id | Reference to customer dimension | INT |
| product\_id | Reference to product dimension | INT |
| retailer\_id | Reference to retailer dimension | INT |
| location\_id | Reference to location dimension (in-store only) | INT |
| payment\_method\_id | Reference to payment method dimension | INT |
| sales\_method\_id | Reference to sales method dimension | INT |
| units\_sold | Number of items sold in this transaction | INT |
| total\_sales | Total revenue from this transaction | DECIMAL(10,2) |
| operating\_profit | Profit after subtracting costs | DECIMAL(10,2) |
| operating\_margin | Profit margin as a percentage | DECIMAL(5,2) |
| avg\_unit\_price | total\_sales / units\_sold | DECIMAL(10,2) |
| profit\_per\_unit | operating\_profit / units\_sold | DECIMAL(10,2) |
| discount\_applied | CASE WHEN price\_per\_unit > avg\_unit\_price THEN 'Y' ELSE 'N' | CHAR(1) |
| is\_high\_margin | CASE WHEN operating\_margin > 30 THEN 'Y' ELSE 'N' | CHAR(1) |

### **Dimension Table: DIM\_CUSTOMER**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| customer\_id | Unique customer ID | INT |
| first\_name | Customer's first name | TEXT |
| last\_name | Customer's last name | TEXT |
| full\_name | Customer’s full name | TEXT |
| customer\_type | New or Returning | TEXT |

### **Dimension Table: DIM\_PRODUCT**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| product\_id | Unique product ID | INT |
| product\_name | Name of the product | TEXT |
| category | Category name | TEXT |
| category\_id | Category ID | INT |
| price\_per\_unit | Price for one item | DECIMAL(10,2) |

### **Dimension Table: DIM\_RETAILER**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| retailer\_id | Unique retailer ID | INT |
| retailer | Retailer name | TEXT |

### **Dimension Table: DIM\_LOCATION (In-store only)**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| location\_id | Surrogate key (generated) | INT |
| region | Region name | TEXT |
| region\_id | Region ID | INT |
| state | State name | TEXT |
| state\_id | State ID | INT |
| city | City name | TEXT |
| city\_id | City ID | INT |

### **Dimension Table: DIM\_PRODUCT\_CATEGORY**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| product\_category\_id | Unique category ID | INT |
| category | e.g., Men or Women | TEXT |

### **Dimension Table: DIM\_JUNK**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| junk\_id | Unique junk ID | INT |
| payment\_method | e.g., Cash or Card | TEXT |
| sales\_method | "Online", "In-store" or “Outlet” | TEXT |

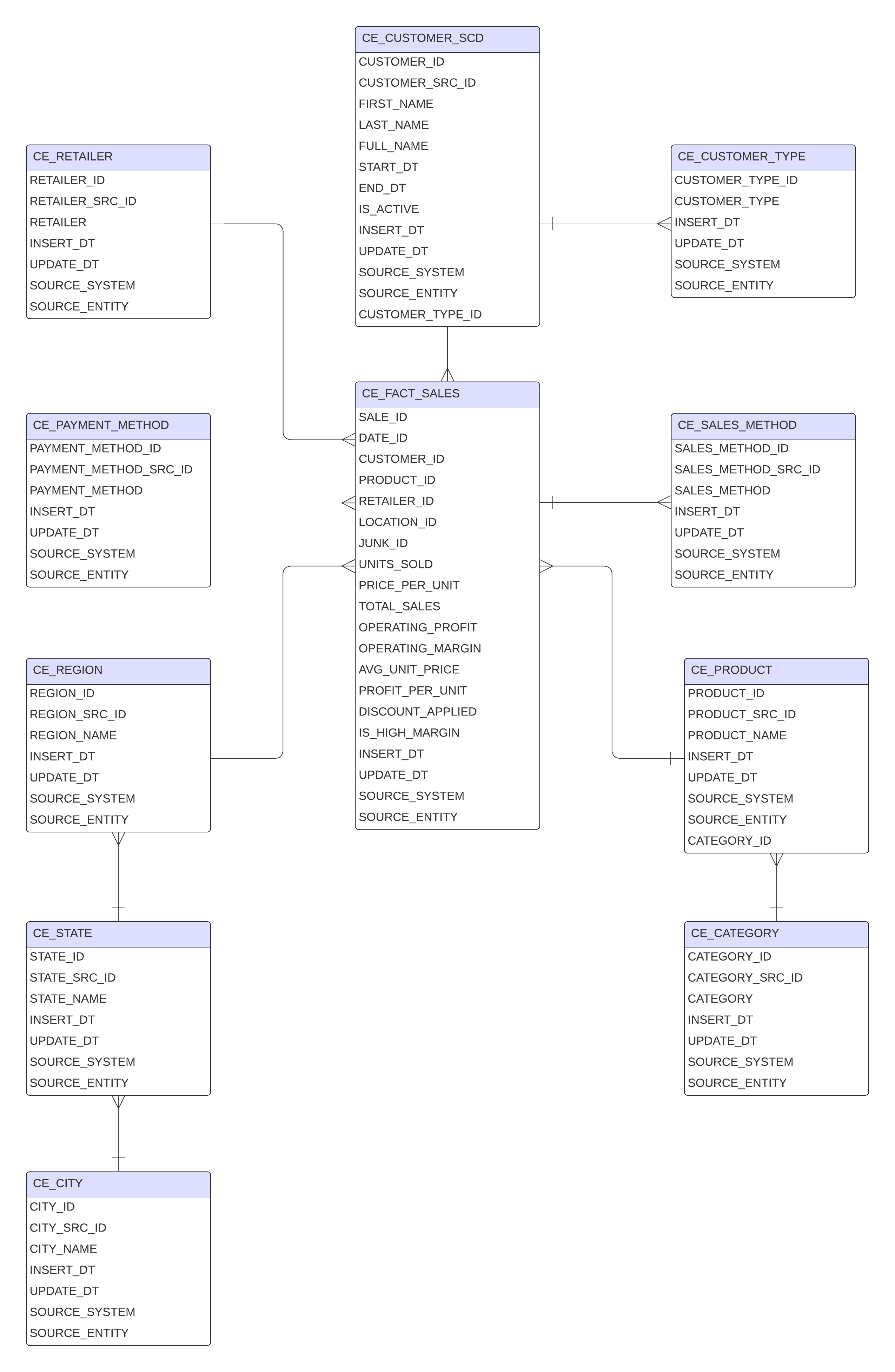
### **Dimension Table: DIM\_DATE**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| date\_id | Surrogate key (YYYYMMDD) | INT |
| day | Day of the month | INT |
| month | Month number | INT |
| year | Year | INT |
| weekday | Name of the weekday | TEXT |
| time | Optional: HH:MM | TEXT |
| week\_number | Number of week | INT |
| is\_weekend | Y or N | CHAR(1) |

### **Dimension Table: DIM\_TIME**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| time\_id | Surrogate key (HH:MM) | INT |
| hour | Hour | INT |
| minute | Minute | INT |
| time\_of\_day | Time of day. For example: between 7 and 10, between 11 and 14, between 15 and 18, between 19 and 22).etc | TEXT |

# Business Layer 3NF



### **Objective**

The objective of this task is to design a normalized data model (3rd Normal Form) for Adidas US sales data. The model must integrate data from multiple datasets, unify entities where applicable, apply Slowly Changing Dimension Type 2 (SCD2) to one selected entity, and include appropriate metadata columns and source tracking (source triplet) to support auditability and historical tracking.

### **Design Approach**

#### **1. Entity Consolidation**

Multiple datasets with overlapping entities (e.g., Customer, Product) were unified into single logical tables. This ensures data consistency and avoids redundancy. For instance, if both sources contain customer information, the resulting model includes a single Customer\_SCD2 table that merges these inputs.

#### **2. SCD Type 2 Implementation**

The Customer entity was selected for SCD Type 2 tracking due to the high likelihood of changes in customer attributes (e.g., name, contact information). The implementation uses the following columns:

* START\_DT, END\_DT: define the validity period of each version.
* IS\_ACTIVE: identifies the currently valid record.
* INSERT\_DT, UPDATE\_DT: track the loading and update timestamps.
* Table name updated to Customer\_SCD2 according to EPAM naming standards.

This structure allows the business to maintain a full history of customer data changes, supporting accurate historical analysis.

#### **3. Source Triplet Columns**

Each table includes the following metadata fields:

* SRC\_SYS: source system name.
* SRC\_TBL: original table name.
* SRC\_COL: original column name.

These columns support traceability, ETL debugging, and audit requirements.

#### **4. Normalization to 3NF**

Each entity table adheres to Third Normal Form (3NF), ensuring:

* Elimination of partial and transitive dependencies.
* Each non-key column depends only on the primary key.
* Entities are structured with surrogate keys and natural keys where applicable.

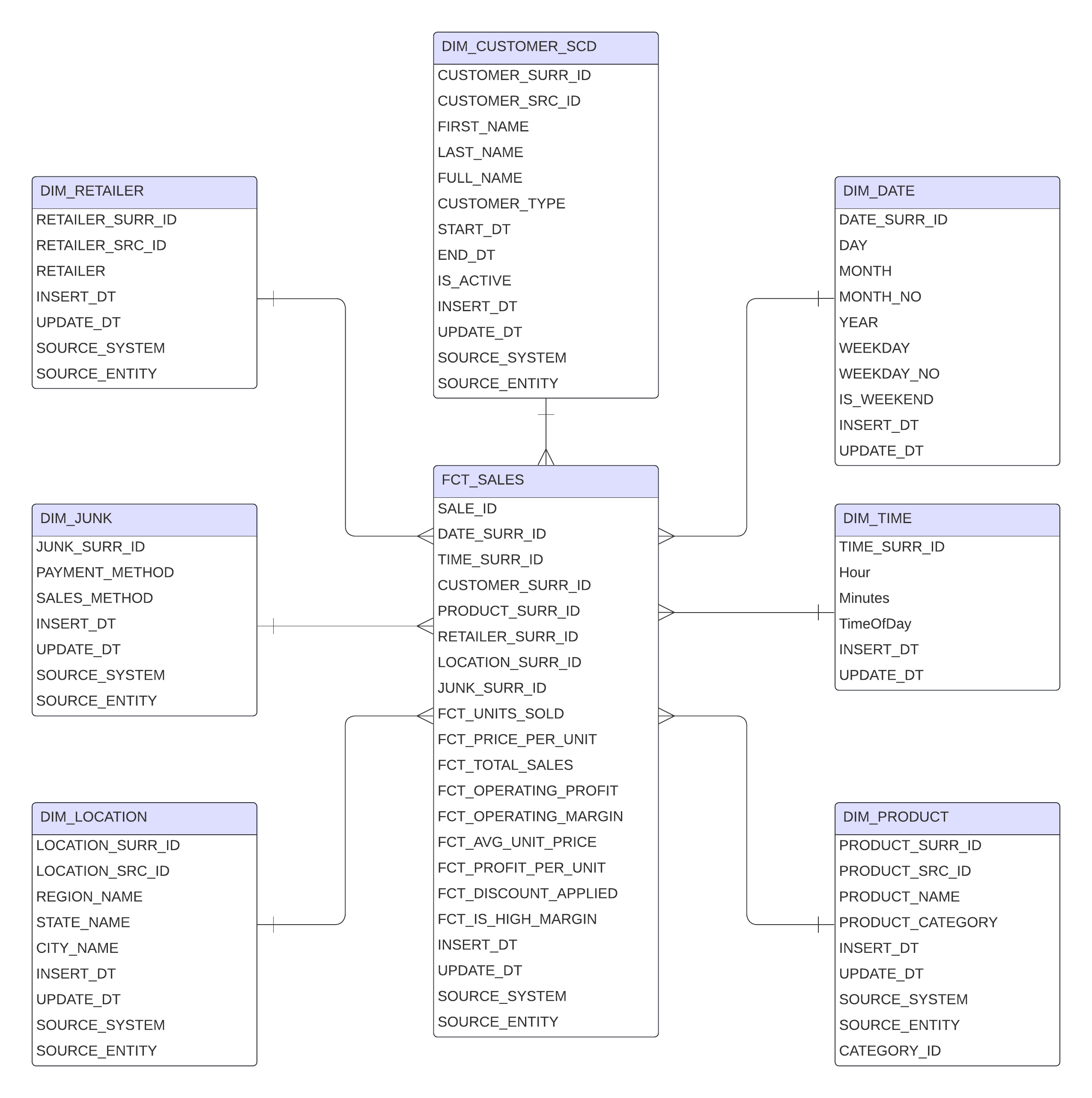
#### **5. Calculated Fields**

The fact table includes additional derived metrics:

* AVG\_UNIT\_PRICE = TOTAL\_SALES / UNITS\_SOLD
* PROFIT\_PER\_UNIT = OPERATING\_PROFIT / UNITS\_SOLD
* IS\_HIGH\_MARGIN = CASE WHEN OPERATING\_MARGIN > 30 THEN 'Y' ELSE 'N'
* DISCOUNT\_APPLIED where applicable based on price comparisons

These fields support more advanced KPI analysis.

# Business Layer Dimensional Model



## **3.1 Star schema**

**Dimensions**

* **DIM\_CUSTOMER\_SCD** (SCD2):  
   CUSTOMER\_SURR\_ID (PK), CUSTOMER\_SRC\_ID, FIRST\_NAME, LAST\_NAME, FULL\_NAME, CUSTOMER\_TYPE, START\_DT, END\_DT, IS\_ACTIVE, INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.
* **DIM\_PRODUCT** (category merged):  
   PRODUCT\_SURR\_ID (PK), PRODUCT\_SRC\_ID, PRODUCT\_NAME, PRODUCT\_CATEGORY, INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.
* **DIM\_RETAILER**:  
   RETAILER\_SURR\_ID (PK), RETAILER\_SRC\_ID, RETAILER, INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.
* **DIM\_LOCATION** (flattened for DM):  
   LOCATION\_SURR\_ID (PK), REGION\_NAME, STATE\_NAME, CITY\_NAME, INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.  
   *Include a default row for online: e.g., CITY\_NAME='N/A', STATE\_NAME='N/A', REGION\_NAME='ONLINE'.*
* **DIM\_JUNK** (sales + payment):  
   JUNK\_SURR\_ID (PK), SALES\_METHOD, PAYMENT\_METHOD, INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.
* **DIM\_DATE** (generated once):  
   DATE\_SURR\_ID (PK, YYYYMMDD), DATE\_DT, DAY, MONTH, MONTH\_NO, YEAR, WEEKDAY, WEEKDAY\_NO, IS\_WEEKEND, INSERT\_DT, UPDATE\_DT.

**Fact**

* **FCT\_SALES\_DD** (daily transaction grain):  
   Keys: DATE\_SURR\_ID, CUSTOMER\_SURR\_ID, PRODUCT\_SURR\_ID, RETAILER\_SURR\_ID, LOCATION\_SURR\_ID (nullable), JUNK\_SURR\_ID.  
   Degenerate: SALE\_ID.  
   Measures:  
   UNITS\_SOLD, PRICE\_PER\_UNIT, TOTAL\_SALES, OPERATING\_PROFIT, OPERATING\_MARGIN (0–1),  
   **Derived**: AVG\_UNIT\_PRICE, PROFIT\_PER\_UNIT, DISCOUNT\_APPLIED (Y/N), IS\_HIGH\_MARGIN (Y/N),  
   Extra attr: **TIME\_OF\_DAY** (‘Morning’, ‘Afternoon’, ‘Evening’, ‘Night’),  
   Tech: INSERT\_DT, UPDATE\_DT, SOURCE\_SYSTEM, SOURCE\_ENTITY.

## **3.2 Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Definition** | **Formula / Rule** |
| **UNITS\_SOLD** | Quantity of items sold | source field |
| **PRICE\_PER\_UNIT** | Actual unit price paid | source field |
| **TOTAL\_SALES** | Line revenue | source field *(or UNITS\_SOLD × PRICE\_PER\_UNIT)* |
| **OPERATING\_PROFIT** | Profit after costs | source field |
| **OPERATING\_MARGIN** | Profit ratio (0–1) | OPERATING\_PROFIT / NULLIF(TOTAL\_SALES,0) |
| **AVG\_UNIT\_PRICE** | Average realized price | TOTAL\_SALES / NULLIF(UNITS\_SOLD,0) |
| **PROFIT\_PER\_UNIT** | Profit per item | OPERATING\_PROFIT / NULLIF(UNITS\_SOLD,0) |
| **DISCOUNT\_APPLIED** | Discount flag | Y if PRICE\_PER\_UNIT > AVG\_UNIT\_PRICE else N |
| **IS\_HIGH\_MARGIN** | High-margin flag | Y if OPERATING\_MARGIN > 0.30 else N |
| **TIME\_OF\_DAY** | Time bucket | ETL-derived from timestamp: Morning(6–11), Afternoon(12–17), Evening(18–22), Night(23–5) |

# Logical Scheme

# Data Flow