

|  |
| --- |
| Business Template  **supermarket sales data** |
| 130+ Thousand Supermarket Logo Royalty-Free Images, Stock Photos & Pictures  | Shutterstock |

Contents

[1 Business Description 3](#_Toc201611847)

[1.1 Business background 3](#_Toc201611848)

[1.2 Problems because of poor data management 3](#_Toc201611849)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc201611850)

[1.4 DATASETS DESCRIPTION 3](#_Toc201611851)

[1.4.1 Dataset 1: Online Orders 3](#_Toc201611852)

[1.4.2 Dataset 2: Offline Orders 3](#_Toc201611853)

[1.5 GRAIN / DIM / FACT 4](#_Toc201611854)

[1.5.1 Fact Table 5](#_Toc201611855)

[1.5.2 Dim Tables 6](#_Toc201611856)

[2 Business Layer 3NF 8](#_Toc201611857)

[3 Business Layer Dimensional Model 8](#_Toc201611858)

[4 Logical Scheme 8](#_Toc201611859)

[5 Data Flow 8](#_Toc201611860)

[6 Fact Table Partitioning Strategy 8](#_Toc201611861)

# 

# Business Description

## Business background

The business operates in the **retail industry**, specifically selling consumer products across various city branches. The organization handles both **online and offline orders**, with a broad product portfolio classified into categories and sold through multiple store locations. The company employs sales staff, offers occasional discounts, and operates in a competitive environment where timely and accurate reporting of sales data is crucial to drive strategic decisions.

## Problems because of poor data management

The company has faced several challenges due to fragmented and inconsistent data management practices. The lack of a unified reporting system has resulted in discrepancies between online and offline sales data, complicating the process of performance evaluation and financial forecasting. Manual changing of records not only delays strategic decision-making but also increases the risk of human error. Also, the business has struggled to gain insights into key performance metrics such as cost efficiency, revenue generation, and discount impact. Without integrated customer and product-level analytics, it is difficult to tailor offerings, track performance, and stay ahead of the competition.

## Benefits from implementing a Data Warehouse

Implementing a data warehouse will significantly enhance the company's ability to collect, store, and analyze data from both operational systems. With structured access to consistent and cleansed sales and cost data, decision-makers will be able to track profitability, identify high-performing products and branches, and monitor employee performance across locations. The warehouse will support time-based, product-based, and geography-based reporting and allow for the exploration of patterns in customer purchases. Hierarchical structures within the data, such as those linking product categories to individual items and time-based groupings like year, month, and day, will provide valuable multidimensional views. Ultimately, the business will benefit from improved forecasting, targeted marketing strategies, and optimized inventory management.

## DATASETS DESCRIPTION

### ****Dataset 1: Online Orders****

Includes all orders placed through the company's e-commerce platform. The dataset contains:

* **Product Information:**

Item Code, Item Name, Category Code, Category Name, Wholesale Price, Loss Rate

* **Sales Facts:**

Date, Time, Quantity Sold, Unit Selling Price, Discount, Total Sales, Cost, Gross Income

* **Organizational Data:**

Employee ID, Branch, City

* **Customer Information:**

Simulated using Customer IDs

* **Other:**

Transaction ID, Source System = "Online", Date hierarchy (Year, Month, Day)

### ****Dataset 2: Offline Orders****

Includes all orders placed in company’s physical branches

* **Product Information:**

Item Code, Category Code, Wholesale Price, Loss Rate

* **Sales Facts:**

Date, Time, Quantity Sold, Unit Selling Price, Discount, Total Sales, Cost, Gross Income

* **Organizational Data:**

Employee ID, Branch, City

* **Customer Information:**

Customer ID

* **Other:**

Transaction ID, Source System = "Offline", Date hierarchy (Year, Month, Day)

In this dataset we have several entities, such as

1. **Product**

Which represents items available for sale. Includes product codes, names, categories, pricing, and expected loss rates. Used to analyze what is being sold.

1. **Customer**

Which represents the buyer.

1. **Date**

Which represents the transaction date and supports hierarchical analysis by day, month, and year. Enables time-based trend reporting.

1. **Employee**

Which represents staff responsible for handling sales (online processing or in-store). Helps assess staff performance and assignment.

1. **Branch**

Which represents physical store locations and helps us to analyze geographic trends and branch-level performance.

1. **Channel**

That distinguishes the source of the sale — either **Online** or **Offline** — helping compare e-commerce vs. physical store performance.

1. **Sales Facts**

That is the central numeric data of each transaction: quantity sold, unit price, discounts, total revenue, cost, and gross income. Tied to all dimensions for complete analysis.

The two datasets differ in several important ways. The online dataset contains complete information for all attributes, while the offline version is limited and that shows the lack of advanced data systems. Additionally, transaction IDs are uniquely prefixed to distinguish between the two sources, and the structure of certain dimensions varies to reflect differing system designs. Despite these differences, both datasets share common dimensions, such as date, customer, and geography, which will allow them to be merged in the data warehouse.

## GRAIN / DIM / FACT

The business process selected is 'Sales Transactions'. This process captures all data generated from both online and offline product sales.  
  
The grain of the fact table is defined at the transaction line-item level, meaning one row represents one product sold in a specific transaction. This ensures the atomic level of detail necessary for flexible and accurate aggregation.  
  
The following dimension tables are identified to provide descriptive context: Product, Customer, Employee, Date, Branch (Geography), and Channel (Online/Offline).  
  
The fact table captures measurable metrics such as Quantity Sold, Unit Price, Total Amount, Discount, Cost, and Gross Income. These values are all additive and useful for analysis across various dimensions.

### Fact Table

FCT\_SALES\_DD

This is the central fact table that stores detailed information about each sales transaction. Each row represents a product sold in a specific transaction, along with measurable data such as quantity sold, sales amount, cost, discount, and gross income. It references all dimension tables through foreign keys.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Transaction\_ID | Unique transaction identifier | INT, Serial, PK |
| Date\_ID | FK to DIM\_DATE | INT |
| Product\_ID | FK to DIM\_PRODUCT | INT |
| Customer\_ID | FK to DIM\_CUSTOMER | INT |
| Employee\_ID | FK to DIM\_EMPLOYEE | INT |
| Branch\_ID | FK to DIM\_BRANCH | INT |
| Channel\_ID | FK to DIM\_CHANNEL | INT |
| Quantity\_Sold | Number of items sold | INT |
| Unit\_Price | Selling price per item | FLOAT |
| Total\_Amount | Total revenue for the sale | FLOAT |
| Discount | Discount applied to sale | FLOAT |
| Cost | Cost of the items sold | FLOAT |
| Gross\_Income | Profit = Total - Cost | FLOAT |

Examples:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Transaction\_ID | Date\_ID | Product\_ID | Customer\_ID | Employee\_ID | Branch\_ID | Channel\_ID | Quantity\_Sold | Unit\_Price | Total\_Amount | Discount | Cost | Gross\_Income |
| ONL\_000001 | 2024-05-12 | 101 | 501 | 3002 | 5 | 1 | 3 | 49.99 | 149.97 | 10.00 | 100.00 | 39.97 |

### Dim Tables

DIM\_PRODUCT

This table contains descriptive information about products such as item code, item name, category, wholesale price, and expected loss rate. It allows analysis of sales performance by product or product category.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Product\_ID | Surrogate primary key | INT, PK |
| Item\_Code | Product code from source system | VARCHAR(20) |
| Item\_Name | Name of the product | TEXT |
| Category\_Code | Category code | VARCHAR(20) |
| Category\_Name | Name of the category | TEXT |
| Wholesale\_Price | Base cost of product | FLOAT |
| Loss\_Rate | Expected loss percentage | FLOAT |

Examples:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Product\_Key | Item\_Code | Item\_Name | Category\_Code | Category\_Name | Wholesale\_Price | Loss\_Rate |
| 101 | A456 | Men's Running Shoes | C001 | Footwear | 33.33 | 0.05 |

DIM\_CUSTOMER

This table holds information about customers including ID, name, gender, age, and segment type. Enables customer-based reporting such as segmentation, frequency of purchase, and targeting.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Customer\_ID | Unique ID for each customer | INT, PK |
| Customer\_Name | Full name of customer | TEXT |
| Gender | Customer's gender | VARCHAR(15) |
| Age | Customer's age | INT |
| Segment | Customer segment type | VARCHAR(10) |

Example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer\_ID | Customer\_Name | Gender | Age | Segment |
| 501 | Ana Kordzaia | Female | 29 | Premium |

DIM\_EMPLOYEE

This table represents sales employees or staff members. Includes employee IDs, names, roles, and hire dates. Useful for analyzing employee performance and sales by staff.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Employee\_ID | Unique ID for each employee | INT, PK |
| Employee\_Name | Full name of the employee | VARCHAR(50) |
| Role | Job position | VARCHAR(50) |
| Hire\_Date | Date of employment | DATE |

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Employee\_ID | Employee\_Name | Role | Hire\_Date |
| 3002 | David Jashi | Sales Associate | 2021-07-10 |

DIM\_DATE

This table stores calendar-related attributes for each transaction date including year, month, day, and weekday. Enables time-based reporting like monthly sales, year-over-year growth, and seasonal trends.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Date\_ID | Unique date ID | INT, PK |
| Year | Calendar year | INT |
| Month | Month number | INT |
| Day | Day of the month | INT |
| Weekday | Day of the week | VARCHAR(20) |

Example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date\_ID | Year | Month | Day | Weekday |
| 2024-05-12 | 2024 | 5 | 12 | Sunday |

DIM\_BRANCH

This table provides details about physical store locations. Includes branch ID, name, city, and region. Used to evaluate store-level performance and geographic trends.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Branch\_ID | Unique branch ID | INT, PK |
| Branch\_Name | Branch name | VARCHAR(50) |
| City | City of the branch | VARCHAR(50) |
| Region | Region/State | VARCHAR(20) |

Example:

|  |  |  |  |
| --- | --- | --- | --- |
| Branch\_ID | Branch\_Name | City | Region |
| 5 | Liberty Mall | Chicago | Illinois |

DIM\_CHANNEL

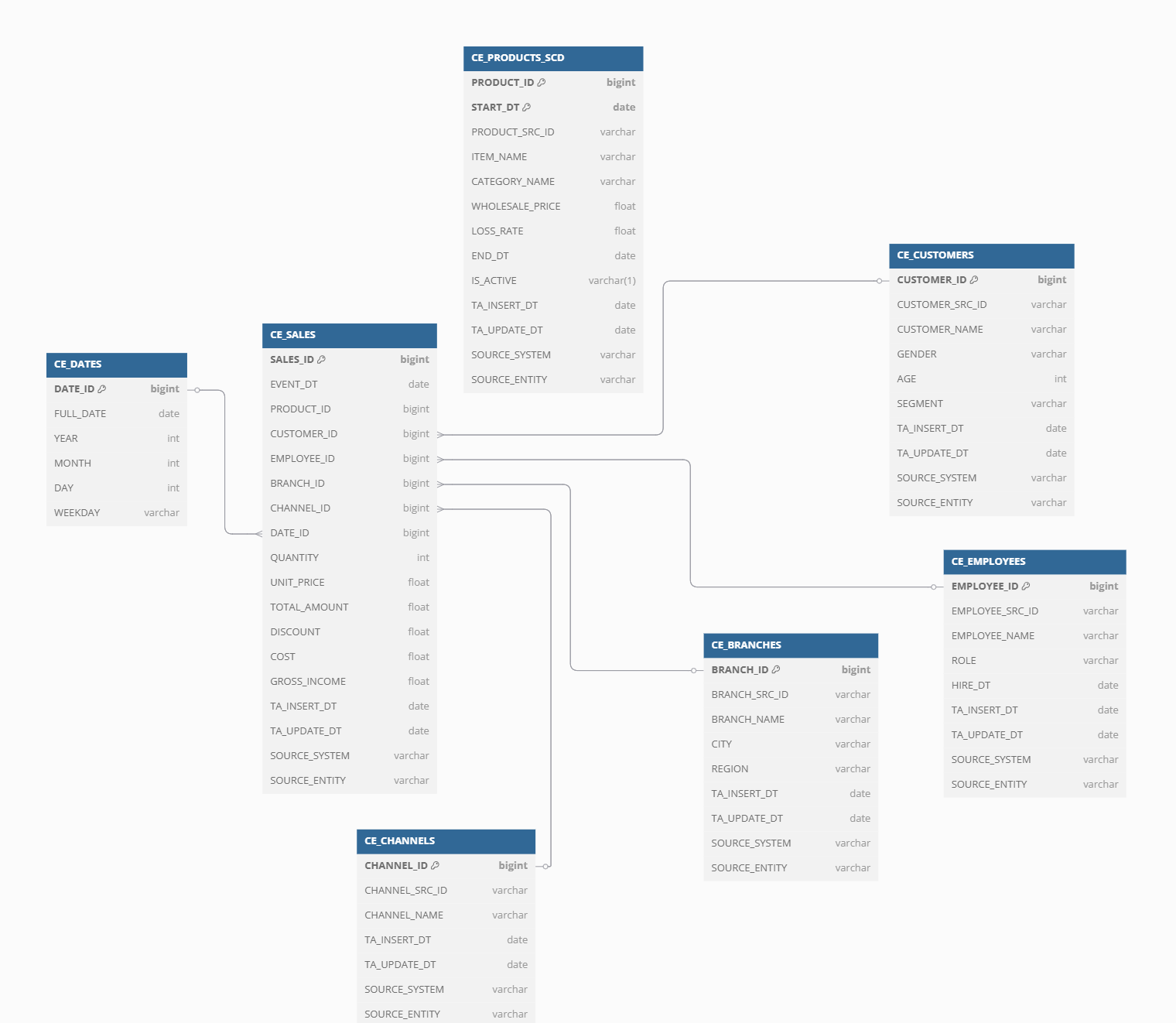
This table distinguishes between sales made online and offline. This dimension enables comparison of performance between different sales platforms.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Data Type |
| Channel\_ID | Unique channel ID | INT, PK |
| Channel\_Name | Sales channel: Online/Offline | VARCHAR(25) |

Example:

|  |  |
| --- | --- |
| Channel\_ID | Channel\_Name |
| 1 | Online |

# Business Layer 3NF



**Overview and steps**

**1. CE\_PRODUCTS\_SCD**

**SCD Type 2 dimension** for products.  
Tracks the history of product changes such as name, category, and price over time.

PRODUCT\_ID is the surrogate key

PRODUCT\_SRC\_ID identifies the original product from the source

START\_DT, END\_DT, and IS\_ACTIVE manage versioning

Multiple records can exist for the same product over time

Used to analyze product evolution or performance historically

**2. CE\_CUSTOMERS**

Holds the unified customer master data across all channels.

CUSTOMER\_ID is the surrogate key used in joins

CUSTOMER\_SRC\_ID stores the business/customer identifier from source

Includes demographic and segmentation data (gender, age, segment)

**3. CE\_EMPLOYEES**

Stores employee information across the organization.

EMPLOYEE\_ID is the system-generated key

EMPLOYEE\_SRC\_ID is the natural identifier from source

Includes personal attributes such as name, role, and hire date

Supports analysis of sales by employee or branch assignments

**4. CE\_BRANCHES**

Represents physical store locations (offline branches).

BRANCH\_ID is the surrogate key

BRANCH\_SRC\_ID comes from the operational system

Provides geographic attributes such as city and region

Enables branch-level performance reporting

**5. CE\_CHANNELS**

Represents the source of the transaction (sales channel).

CHANNEL\_ID is the primary key

CHANNEL\_SRC\_ID indicates source-level identifier

CHANNEL\_NAME describes the mode: e.g., **Online**, **Offline**

Useful for comparing channel performance, adoption, and trends

**6. CE\_DATES**

Standard **date dimension** supporting all time-based analysis.

DATE\_ID is the surrogate key

Stores a single row per date (granularity = 1 day)

Includes calendar breakdown: year, month, day, weekday

Enables slicing metrics by time, including trends and seasonal patterns

**7. CE\_SALES**

Core **fact-like table** that records sales transaction events.

SALES\_ID is the surrogate key (one row per transaction)

Stores all measurable sales metrics: quantity, price, discount, cost, income

Foreign keys reference all relevant dimensions via surrogate keys

Includes CUSTOMER\_ID, EMPLOYEE\_ID, BRANCH\_ID, CHANNEL\_ID, DATE\_ID

**Uses** PRODUCT\_ID **logically only** to avoid FK conflict with SCD2Includes source tracking (SOURCE\_SYSTEM, SOURCE\_ENTITY) and audit timestamps

**Step 1: Analyzed the Source Systems**

I started by reviewing both the online and offline sales datasets. Though they had structural differences (e.g. presence/absence of product names or item categories), I unified shared business entities into single logical tables. This included entities like Customer, Product, Employee, Branch, Channel, and Date.

**Step 2: Normalized into Entities**

Each real-world concept was converted into a normalized table. All non-key attributes were placed into separate entities to remove redundancy and preserve referential integrity. For instance, sales transactions were separated from product details, employee data, and customer profiles.

**Step 3: Applied Naming Conventions**

All business layer tables are prefixed with CE\_ (Core Entity).

Surrogate keys use the pattern <ENTITY>\_ID, and natural keys use <ENTITY>\_SRC\_ID.

All timestamp columns are prefixed with TA\_ for example TA\_INSERT\_DT, TA\_UPDATE\_DT.

SCD2 fields include START\_DT, END\_DT, and IS\_ACTIVE

**Step 4: Defined Surrogate and Natural Keys**

Every table includes:

A **surrogate key** (\*\_ID) that is system-generated and acts as the primary key.

A **natural key** (\*\_SRC\_ID) derived from source data, which forms part of the **source triplet** for lineage.

For example, PRODUCT\_ID is a surrogate key in CE\_PRODUCTS\_SCD, while PRODUCT\_SRC\_ID comes directly from the transactional source.

**Step 5: Added Source Triplet and Technical Metadata**

To ensure traceability and support data governance:

Each table includes the **source triplet**: SOURCE\_SYSTEM, SOURCE\_ENTITY, and \*\_SRC\_ID.

Timestamps for data ingestion and update are consistently named: TA\_INSERT\_DT and TA\_UPDATE\_DT.

For Slowly Changing Dimensions (like Products), we also added: START\_DT, END\_DT, and IS\_ACTIVE.

**Step 6: Handled Slowly Changing Dimension**

The CE\_PRODUCTS\_SCD table was designed as an **SCD Type 2** dimension. This allows tracking of changes in product attributes over time (e.g., price updates, renaming). Its primary key is **composite**: PRODUCT\_ID + START\_DT.

This composite key allows versioned records to coexist and supports historical analysis. IS\_ACTIVE flags the current version.

**Step 7: Defined Logical Relationships via Surrogate Keys**

The fact table CE\_SALES references all related dimensions using surrogate keys:

CUSTOMER\_ID, EMPLOYEE\_ID, BRANCH\_ID, CHANNEL\_ID, DATE\_ID

**Special handling for SCD2:**  
The reference to PRODUCT\_ID in CE\_SALES is maintained **logically only** — no foreign key constraint is applied. This aligns with the guideline that SCD2 tables have composite PKs, and joins to them should rely on business rules, not enforced keys.

# Business Layer Dimensional Model

# Logical Scheme

# Data Flow

# Fact Table Partitioning Strategy