

|  |
| --- |
| Business Template  transportation sales data |

Contents

[1 Business Description 3](#_Toc155614186)

[1.1 Business background 3](#_Toc155614187)

[1.2 Problems because of poor data management 3](#_Toc155614188)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc155614189)

[1.4 DATASETS DESCRIPTION 3](#_Toc155614190)

# 

# Business Description

## Business background

Planes, trains, and cars are essential elements of modern transportation. Numerous companies and manufacturers offer various types of vehicles suitable for different needs, routes, and user preferences. This industry is highly competitive, so to be successful in this field, one must approach it responsibly and consider numerous factors that influence people's choice of transportation. First of all, this can be achieved by collecting product sales information and analyzing it using specialized tools.

## Problems because of poor data management

Poor data management can severely hinder the success of a business due to the lack of sufficient information to guide decision-making. Without effective tools to gather, analyze, and interpret data, companies are unable to develop informed business strategies. This deficiency makes it difficult to stay competitive in any industry. In the transportation sector, for instance, failing to manage sales and customer data properly can result in missed opportunities, inefficient operations, and an inability to meet customer demands. Effective data management is essential for making strategic decisions, improving customer satisfaction, and maintaining a competitive edge.

## Benefits from implementing a Data Warehouse

Implementing a data warehouse can help address the problems described above. A data warehouse allows businesses to compile, store, and analyze vast amounts of data efficiently. This implementation can provide answers to key questions such as:

* Which transportation modes (planes, trains, cars) have the highest sales volumes?
* Which vehicle types have the widest distribution of prices?
* Is there a typical sales distribution across different transportation modes or within specific vehicle types?

Further processing of data through a data warehouse would also allow businesses to:

* Correlate specific vehicle features with changes in sales or prices.
* Identify differences in preferences between various customer segments.
* Optimize inventory management by understanding seasonal sales trends.
* Enhance customer satisfaction by identifying and addressing pain points.
* Develop targeted marketing strategies based on comprehensive data insights.
* And many other critical analyses that can drive business success.

## DATASETS DESCRIPTION

The first dataset contains the following information about sales on the American market.

#### Order Information:

* **ORDERNUMBER:** A unique identifier for each order.
* **QUANTITYORDERED:** The number of units of a product ordered.
* **PRICEEACH:** The price per unit of the product.
* **ORDERLINENUMBER:** The number of order lines associated with the order.
* **SALES:** The total sales amount for the order (calculated as QUANTITYORDERED \* PRICEEACH).
* **ORDERDATE:** The date when the order was placed.
* **DEALSIZE\_ID:** The size of the deal (e.g., Small, Medium, Large).
* **STATUS:** The current status of the order (e.g., Shipped, Resolved, Cancelled, In Process, On Hold).

#### Time Information:

* **QTR\_ID:** The quarter of the year in which the order was placed (1 to 4).
* **DAY\_ID:** The day of the month when the order was placed.
* **MONTH\_ID:** The month when the order was placed.
* **YEAR\_ID:** The year when the order was placed.

#### Product Information:

* **PRODUCTLINE\_ID:** A unique identifier for the product line.
* **PRODUCTLINE:** The category or type of product (e.g., Motorcycles, Classic Cars, Trucks and Buses, Trains, Vintage Cars, Planes).
* **MSRP:** The manufacturer's suggested retail price of the product.
* **PRODUCTCODE:** The specific code of the product.

#### Customer Information:

* **CUSTOMER\_ID:** A unique identifier for each customer.
* **CUSTOMERNAME:** The name of the customer.
* **CONTACTFIRSTNAME:** The first name of the customer.
* **CONTACTLASTNAME:** The last name of the customer.
* **PHONE:** The phone number of the customer.

#### Address Information:

* **ADDRESS\_ID:** A unique identifier for the address.
* **ADDRESSLINE1:** The street address of the customer.
* **CITY:** The city where the customer resides.
* **STATE:** The state where the customer resides.
* **POSTALCODE:** The postal code of the customer's address.
* **COUNTRY:** The country where the customer resides.

#### Payment method:

* **PAYMENT\_ID:** unique identifier for the payment.
* **PAYMENT\_METHOD:** With what customer paid.

#### Dealsize:

* **DEALSIZE\_ID:** unique identifier for the deal size .
* **DEALSIZE:** what deal size customer bought

For the second data there is following structure

#### Order Information:

* **ORDER\_ID:** A unique identifier for each order.
* **QUANTITY:** The number of units of a product ordered.
* **PRICE\_FOR\_EACH:** The price per unit of the product.
* **NUMBER\_OF\_ORDERLINE:** The number of order lines associated with the order.
* **SALES\_AMOUNT:** The total sales amount for the order (calculated as QUANTITY \* PRICE\_FOR\_EACH).
* **DATE\_OF\_ORDER:** The date when the order was placed.
* **DEAL\_SIZE:** The size of the deal (e.g., Small, Medium, Large).
* **CURRENT\_STATUS:** The current status of the order (e.g., Shipped, Resolved, Cancelled, In Process, On Hold).

#### Time Information:

* **QUAERTER:** The quarter of the year in which the order was placed (1 to 4).
* **DAY:** The day of the month when the order was placed.
* **MONTH:** The month when the order was placed.
* **YEAR:** The year when the order was placed.

#### Product Information:

* **PRODUCTLINE\_ID:** A unique identifier for the product line.
* **PRODUCT\_LINE:** The category or type of product (e.g., Motorcycles, Classic Cars, Trucks and Buses, Trains, Vintage Cars, Planes).
* **MS\_RP:** The manufacturer's suggested retail price of the product.
* **PRODUCT\_MODEL:** The specific model of the product.

#### Customer Information:

* **CUSTOMERS\_ID:** A unique identifier for each customer.
* **CUST\_NAME:** The first name of the customer.
* **CUST\_LASTNAME:** The last name of the customer.
* **PHONE\_NUMBER:** The phone number of the customer.

#### Address Information:

* **ADDRESS\_ID\_NUM:** A unique identifier for the address.
* **ADDRESS\_LINE:** The street address of the customer.
* **CITY\_NAME:** The city where the customer resides.
* **POSTCODE:** The postal code of the customer's address.
* **COUNTRY\_NAME:** The country where the customer resides.

#### Payment method:

* **PAYMENT\_ID:** unique identifier for the payment.
* **PAYMENT\_METHOD:** With what customer paid.

#### Dealsize:

* **DEALSIZE\_ID:** unique identifier for the deal size .
* **DEALSIZE:** what deal size customer bought

.

Main difference between two datasets are that first one contains information about USA sales, it has STATE column additionally.

Second data contains is rest of the countries, without USA.

The data is sales transactions involving various transportation products, customers, and geographic locations. The business process involves capturing the details of each sale, including quantities, prices, customer details, and the status of the order and addresses.

## gRAin

Grain as understand, the single row in the fact table, will be (after I denormalize it):

 ORDERNUMBER

 QUANTITYORDERED

 SALES

 STATUS

 PRODUCT\_ID

 CUSTOMER\_ID

 ADDRESS\_ID

* TIME\_ID
* PAYEMENT\_ID
* DEALSIZE\_ID

Dimensions will be:

TIME TABLE (DIMENSION):

* TIME\_ID (WHICH WILL BE ADDED LATER)
* QRT\_ID
* YEAR
* MONTH
* DAY

PRODUCT TABLE (DIMENSION):

|  |  |  |
| --- | --- | --- |
|  |  |  |

* PRODUCTLINE\_ID
* PRODUCTLINE
* MSRP
* PRODUCTCODE

CUSTOMER TABLE (DIMENSION):

 CUSTOMER\_ID

 CUSTOMERNAME

 CONTACTFIRSTNAME

 CONTACTLASTNAME

 PHONE

 ADDRESSLINE1

 CITY

 STATE

 POSTALCODE

 COUNTRY

PAYMENT\_METHODS TABLE (DIMENSION)

* PAYMENT\_ID
* PAYMENT\_METHOD

DEALSIZE TABLE (DIMENSION)

* DEALSIZE\_ID
* DEALSIZE

With chosen grain it will be possible to have detailed analysis of sales transactions at the most granular level. It enables to understand the specifics of what products are being sold, in what quantities, at what prices, and to which customers.

Information can be used for inventory management, sales forecasting, marketing strategies, and customer relationship management.

Additionally, With this grain, the data can be aggregated to higher levels to support various types of analysis. Also, Capturing data at the line item level ensures that no detail is lost, which is crucial for accurate reporting and decision-making.

ORDERNUMBER - uniquely identifies the order and ties the transaction back to a specific purchase event. It helps in tracking the entire order process from initiation to completion.

QUANTITYORDERED - represents the number of units of the product ordered. This is crucial for inventory management, sales forecasting, and understanding customer purchasing patterns.

SALES - captures the total sales amount for the line item. It is essential for revenue analysis and financial reporting.

STATUS - indicates the current state of the order (Shipped, Pending). This is important for order fulfillment tracking and customer service.

PRODUCT\_ID - links the order line item to a specific product. It is used to analyze product performance, manage product lifecycles, and optimize inventory.

CUSTOMER\_ID - ties the order to a specific customer, allowing for customer segmentation, personalization, and targeted marketing efforts.

ADDRESS\_ID - relates to the delivery address for the order. It aids in logistics planning, delivery performance analysis, and geographic sales analysis.

TIME\_ID - connects the order to the time dimension, capturing when the order was placed. It enables time-based analysis, such as identifying sales trends over time, seasonality, and peak buying periods.

PAYMENT\_ ID - links the order to the payment method used. This is critical for understanding customer payment preferences to analyze payment method performance, and managing financial processes.

DEAL\_ID - links to the deal size customer purchased. It is important to know the preferences of the customer, which deal size they prefer, to make business decisions accordingly

ORDER TABLE

| **Column name** | **Description** | **Data Type** |
| --- | --- | --- |
| ORDERNUMBER | Unique identifier for the order | Int |
| QUANTITYORDERED | Number of units ordered | Int |
| SALES | Total sales amount | Decimal |
| ORDERDATE | The date when the order was placed | Date |
| DEALSIZE | Size of the deal | Text |
| STATUS | Status of the order | Text |
| PRODUCT\_ID | Foreign key to the product dimension | Int |
| CUSTOMER\_ID | Foreign key to the customer dimension | Int |
|  |  |  |
| TIME\_ID | Foreign key to the time dimension | Int |

| **ORDERNUMBER** | **QUANTITYORDERED** | **SALES** | **ORDERDATE** | **DEALSIZE** | **STATUS** | **PRODUCT\_ID** | **CUSTOMER\_ID** | **ADDRESS\_ID** | **TIME\_ID** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 30 | 2871 | 2/24/2003 | Small | Shipped | 1 | 1 | 1 | 1 |
| 2 | 45 | 3746.7 | 8/25/2003 | Medium | Shipped | 1 | 2 | 2 | 2 |

TIME TABLE

| **Column name** | **Description** | **Data Type** |
| --- | --- | --- |
| TIME\_ID | Unique identifier for the time entry | Int |
| QRT\_ID | Quarter of the year | Int |
| YEAR | Year of the date | Int |
| MONTH | Month of the date | Int |
| DAY | Day of the date | Int |

| **TIME\_ID** | **QRT\_ID** | **DAY\_ID** | **MONTH\_ID** | **YEAR\_ID** |
| --- | --- | --- | --- | --- |
| 1 | 30 | 95.7 | 2871 | 2/24/2003 |
| 2 | 45 | 83.26 | 3746.7 | 8/25/2003 |

PRODUCT TABLE

| **Column name** | **Description** | **Data Type** |
| --- | --- | --- |
| PRODUCTLINE\_ID | Unique identifier for the product line | Int |
| PRODUCTLINE | Name of the product line | Text |
| PRICEEACH | Price per unit | Decimal |
| MSRP | Manufacturer's suggested retail price | Decimal |
| PRODUCTCODE | Specific code of the product | Text |

| **PRODUCTLINE\_ID** | **PRODUCTLINE** | **MSRP** | **PRODUCTCODE** | **PRICEEACH** |
| --- | --- | --- | --- | --- |
| 1 | Motorcycles | 95 | S10\_1678 | 10 |
| 2 | car | 87 | S42\_166 | 20 |

CUSTOMER\_TABLE

| **CUSTOMER\_ID** | **CUSTOMERNAME** | **CONTACTFIRSTNAME** | **CONTACTLASTNAME** | **PHONE** | **ADDRE** | **PRODUCTCODE** | **ADDRESSLINE1** | **CITY** | **STATE** | **COUNTRY** | **POSTALcODE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Land of Toys Inc. | Kwai | Yu | 2125557818 | 95 | S10\_1678 | Street 5 | NYC | NY | USA | 1010 |
| 2 | Toys4GrownUps.com | Julie | Young | 6265557265 | 87 | S42\_166 | Bara street | Tokyo | NA | Japan | 2020 |

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| CONTACTFIRSTNAME | First name of the contact person | Text |
| CONTACTLASTNAME | Last name of the contact person | Text |
| PHONE | Contact phone number | Text |
| ADDRESSLINE1 | Street address | Text |
| CITY | City | Text |
| STATE | State | Text |
| POSTALCODE | Postal code | Int |
| COUNTRY | Country | Text |

Payment Method

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| PAYMENT\_ID | Unique identifier of the payment | int |
| PAYMENT\_METHOD | What method customer used for payment | Text |

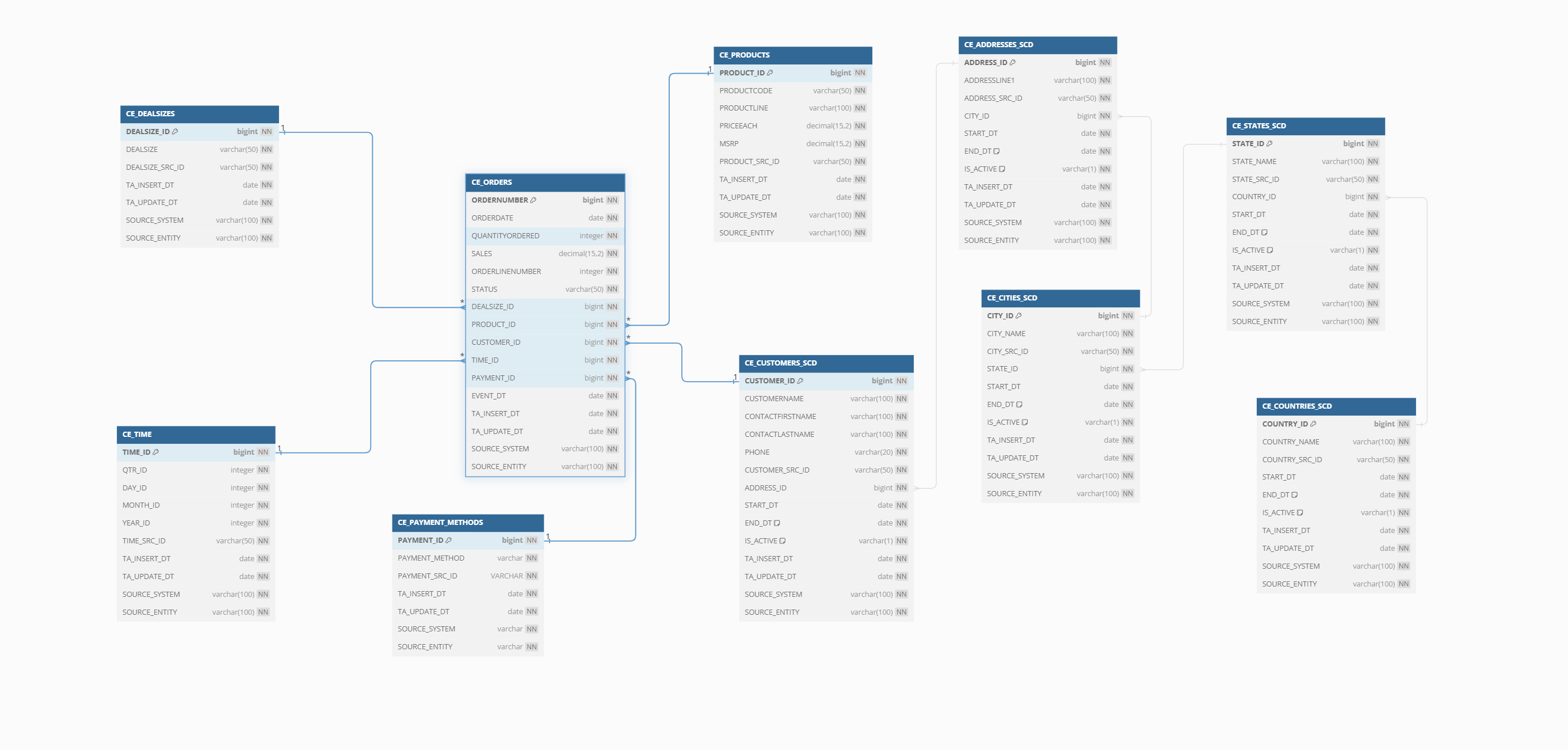
| PAYMENT\_ID | PAYMENT\_METHOD |
| --- | --- |
| 1 | CARD |
| 2 | CASH |

DEAL SIZE

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| DEALSIZE\_ID | Unique identifier of the dealsize | int |
| DEALSIZE | Description of the dealsize | Text |

| dealsize\_ID | DEALSIZE |
| --- | --- |
| 1 | MEDIUM |
| 2 | SMALL |

# Business Layer 3NF



In transforming the original data model into a 3NF schema, I aimed to ensure data integrity, eliminate redundancy, and maintain a clear hierarchy:

**CE\_ORDERS table** is the Fact Table that stores transactional data. It includes attributes like ORDERNUMBER (Primary Key), ORDERDATE, QUANTITYORDERED, SALES, ORDERLINENUMBER, STATUS, DEALSIZE\_ID (Foreign Key to CE\_DEALSIZES), PRODUCT\_ID (Foreign Key to CE\_PRODUCTS), CUSTOMER\_ID (Foreign Key to CE\_CUSTOMERS\_SCD), TIME\_ID (Foreign Key to CE\_TIME), EVENT\_DT (date of the event), TA\_INSERT\_DT (row creation date), and TA\_UPDATE\_DT (last row modification date). This table is the core of the schema, capturing detailed transaction data.

**CE\_PRODUCTS table** is an SCD Type 1 Dimension Table. It includes PRODUCT\_ID (Primary Key), PRODUCTCODE, PRODUCTLINE, PRICEEACH, MSRP, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table holds the current state of each product, with changes overwriting previous values.

**CE\_CUSTOMERS\_SCD table** is an SCD Type 2 Dimension Table. It includes CUSTOMER\_ID (Primary Key), CUSTOMERNAME, CONTACTFIRSTNAME, CONTACTLASTNAME, PHONE, ADDRESS\_ID (Foreign Key to CE\_ADDRESSES), SOURCE\_SYSTEM, SOURCE\_ENTITY, START\_DT (Start Date of the active period), END\_DT (End Date of the active period), IS\_ACTIVE (status indicator), TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table tracks historical changes in customer information, allowing for a historical view over time.

**DIM\_ADDRESSES table** is a Dimension Table. It includes ADDRESS\_ID (Primary Key), ADDRESSLINE1, CITY\_ID (Foreign Key to CE\_CITY), CITY\_NAME, STATE\_ID (Foreign Key to CE\_STATE), STATE\_NAME, COUNTRY\_ID (Foreign Key to CE\_COUNTRY), COUNTRY\_NAME, POSTALCODE, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table keeps the hierarchical structure of location data, linking cities, states, and countries through foreign keys.

**CE\_CITY table** is a Dimension Table with CITY\_ID (Primary Key), CITY\_NAME, STATE\_ID (Foreign Key to CE\_STATE), SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. It captures city information and links it to states.

**CE\_STATE table** is another Dimension Table, with STATE\_ID (Primary Key), STATE\_NAME, COUNTRY\_ID (Foreign Key to CE\_COUNTRY), SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table captures state information and links it to countries.

**CE\_COUNTRY table** is a Dimension Table with COUNTRY\_ID (Primary Key), COUNTRY\_NAME, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. It captures country information.

**CE\_TIME table** is a dimension Table that includes TIME\_ID (Primary Key), QTR\_ID, DAY\_ID, MONTH\_ID, YEAR\_ID, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table holds time-related data, essential for time-based analysis.

**CE\_DEALSIZES table** is another dimension Table, with DEALSIZE\_ID (Primary Key), DEALSIZE, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table captures the sizes of deals for orders.

CE\_PAYMENT\_METHODS table is dimension Table, with PAYMENT\_ID (Primary Key), PAYMENT\_METHOD, SOURCE\_SYSTEM, SOURCE\_ENTITY, TA\_INSERT\_DT, and TA\_UPDATE\_DT. This table captures the methods of payment for orders.

Additionally, technical attributes TA\_INSERT\_DT and TA\_UPDATE\_DT are added to all tables to track row creation and modification dates. The attributes SOURCE\_SYSTEM and SOURCE\_ENTITY and SMTH\_SOURCE\_ID identify the data's origin, crucial for tracking the data flow from source systems to the data warehouse.

The hierarchical organization is maintained in the DIM\_ADDRESSES table, which connects to the CE\_CITY, CE\_STATE, and CE\_COUNTRY tables through foreign keys. This setup ensures data integrity across different geographical levels.