

|  |
| --- |
| Business Template  **Retail of Electric vehicles** |
|  |

Contents

[1 Business Description 3](#_Toc188868992)

[1.1 Business background 3](#_Toc188868993)

[1.2 Problems because of poor data management 3](#_Toc188868994)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc188868995)

[1.4 DATASETS DESCRIPTION 3](#_Toc188868996)

[1.5 GRAIN / DIM / FACT 5](#_Toc188868997)

[2 Business Layer 3NF 9](#_Toc188868998)

[3 Business Layer Dimensional Model 9](#_Toc188868999)

[4 Logical Scheme 9](#_Toc188869000)

[5 Data Flow 9](#_Toc188869001)

[6 Fact Table Partitioning Strategy 9](#_Toc188869002)

# 

# Business Description

## Business background

Electric vehicles (EVs) are becoming an essential part of modern transportation as society transitions toward sustainable solutions. This category includes Battery Electric Vehicles (BEVs), which rely solely on electric power, and Plug-in Hybrid Electric Vehicles (PHEVs), which combine electric power with a traditional internal combustion engine. Numerous businesses and organizations focus on providing EVs and related services to meet the growing demand for eco-friendly and efficient mobility options. This field is highly competitive, requiring a strategic approach to succeed, as businesses must consider various factors influencing consumer preferences, such as affordability, charging infrastructure, and technological advancements. Understanding market trends and consumer behavior is key, often achieved by collecting and analyzing data on BEV and PHEV adoption, usage patterns, and performance. These insights can help businesses refine their offerings, enhance customer satisfaction, and remain competitive in the evolving transportation landscape.

## Problems because of poor data management

Ineffective data management in the electric vehicle (EV) sector can significantly impede business growth and decision-making. Without proper tools to gather and analyze data on Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), businesses may lack critical insights into consumer behavior, market demands, and performance trends. This lack of information makes it difficult to devise effective strategies, adapt to changes in the industry, and stay competitive. In a field driven by innovation and data, poor data management can lead to inefficiencies, lost opportunities, and a failure to meet customer expectations.

## Benefits from implementing a Data Warehouse

Implementing a data warehouse can address the challenges outlined above and provide valuable insights into the electric vehicle (EV) market. A data warehouse can help answer critical questions such as:

* What are the most popular EV models based on sales data?
* Which vehicle makes and models are the most commonly registered in specific areas?
* What is the average electric range of registered Battery Electric Vehicles (BEVs) across different regions?
* How has the registration of Plug-in Hybrid Electric Vehicles (PHEVs) evolved over time in specific areas?
* Identify trends in EV adoption rates across cities.
* Analyze the impact of vehicle model year on registration trends.

By leveraging a data warehouse, businesses can gain a competitive edge through informed decision-making and strategic planning.

## DATASETS DESCRIPTION

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

Product Information:

VIN (1-10): The 1st 10 characters of each vehicle's Vehicle Identification Number (VIN), unique identifier.

Model\_Year: The year of production.

Make: The brand or manufacturer of the car.

Model: The category of the car.

Electric Vehicle Type: This distinguishes the vehicle as all electric or a plug-in hybrid.

Clean Alternative Fuel Vehicle (CAFV) Eligibility: This categorizes vehicle as Clean Alternative Fuel Vehicles (CAFVs) based on the fuel requirement and electric-only range requirement in House Bill 2042 as passed in the 2019 legislative session.

Electric Range: Describes how far a vehicle can travel purely on its electric charge.

Sales Information:

TransactionID: The unique identifier of transaction.

Date: The date of the car sale.

Selling\_Price: The actual selling price of the car.

Customer Information:

Customer\_Passport: The unique identifier of customer.

Name: Name of customer

Gender: The gender of customer

Location Information:

State: This is the geographic region of the country associated with the record. (all WA)

City: The city in which the registered owner resides.

Postal Code: The 5 digit zip code in which the registered owner resides.

Employee (who sold a car) Information:

Employee\_id: The unique identifier of employee.

Employee\_name: Name of customer

Employee\_phone: The phone of employee

Sales Channel Information:

Channel\_ID: The unique identifier of channel.

Channel\_Type: Type of Channel (e.g., Online, Showroom, Event, Third\_party)

Channel\_Description: Additional notes about the channel

The second dataset contains the following information:

Product Information:

VIN: The Vehicle Identification Number (VIN), unique identifier.

Year: The year of production

Producer: The brand or manufacturer of the car.

Model: The category of the car.

Electric Vehicle Type: This distinguishes the vehicle as all electric or a plug-in hybrid.

Type CAFV: This categorizes vehicle as Clean Alternative Fuel Vehicles (CAFVs) based on the fuel requirement and electric-only range requirement in House Bill 2042 as passed in the 2019 legislative session.

Electric Range: Describes how far a vehicle can travel purely on its electric charge.

Transmission: type of transmission (automatic/manual)

Color: Color of the car.

Sales Information:

Transaction\_id: Unique identifier of transaction

Price ($): The actual selling price of the car.

Payment\_Method: the method of payment the price

Date: The date of the car sale.

Customer Information:

Customer\_id: The unique identifier of customer.

Customer\_name: Name of customer

Customer\_age: The age of customer

Location Information:

P\_Code: The digit zip code in which the registered owner resides, unique identifier

State: This is the geographic region of the country associated with the record. (all FL)

City: The city in which the registered owner resides.

Employee (who sold a car) Information:

Empl\_id: The unique identifier of employee.

Empl\_name: Name of customer

Empl\_email: The email of employee

The datasets difference based on geographical data, different systems for different states, in this case WA (Washington) and FL (Florida).

The datasets provide a comprehensive overview of the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) sales, allowing for analysis and exploration of trends, sales performance, customer preferences.

## GRAIN / DIM / FACT

The grain of this data is at the level of individual transactions for electric vehicles. Each row represents a single transaction with specific details, such as the vehicle's VIN (unique indentifier), selling price, date, customer details, and the employee handling the transaction. This grain allows analysis at a transactional level, enabling insights such as sales trends, customer behavior, and employee performance.

* **Fact Table:** Contains measurable, numeric data for analysis (e.g., sales data).
* **Dimension Tables:** Contain descriptive attributes to provide context (e.g., customer, vehicle, employee).

**Step 1: Select the Business Process**

The business process represented in the dataset is the vehicle sales process. This process involves selling vehicles to customers, recording vehicle-specific details (e.g., VIN, producer, model, electric range), capturing customer and employee information (e.g., name, age, email), and logging transaction data (e.g., date, price, payment method and transaction ID). This process generates key performance metrics such as the number of vehicles sold, total sales revenue, and customer trends.

**Step 2: Declare the Grain**

The grain of the dataset is one row per vehicle sales transaction. This level of detail ensures that every row corresponds to a specific sale event for a particular vehicle.

Each row captures a unique combination of:

* Vehicle details: (e.g., VIN, producer, model, year, electric type)
* Customer details: (e.g., customer’s passport, name, age, gender)
* Transaction details: (e.g., transaction ID, selling price, date, payment method)
* Employee details: (e.g., employee ID, name, and email)

**Step 3: Identify the Dimensions**

Dimensions describe the contextual details of the fact table (sales transactions). The key dimensions for this dataset include: Vehicle Dimension, Customer Dimension, Location Dimension, Employee Dimension, Sales Channel Dimension.

**Step 4: Identify the Facts**

Facts represent measurable metrics in the dataset that align with the declared grain. For this dataset, the facts include selling price (numeric, additive)

**FACT\_TRANSACTION**

Description: Contains the numeric data related to each transaction.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Transaction\_Id | Unique identifier for the transaction | Serial |
| VIN\_FK | Unique identifier of vehicle | Text |
| Price | Price at which the vehicle was sold | Float |
| Transaction\_Date | Date of the transaction | Date |
| Payment\_Method | Type of Payment | Text |
| Customer\_id\_FK | Unique identifier of customer | Int |
| Employee\_id\_FK | Unique identifier of employee | Text |
| Channel\_id\_FK | Unique identifier of channel | Text |
| Postal\_code\_FK | Unique location identifier | Int |

Example with filled data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Transaction\_id** | **VIN\_FK** | **Custo**  **mer\_id\_FK** | **Price** | | **Transaction\_Date** | **Payment\_Method** | **Employee\_id\_FK** | **Channel\_id\_FK** | **Postal\_code\_FK** |
| TX000001 | 5YJ3E1EA5L | 00000001 | | 21500 | 1/2/2022 |  | F25 | CH0007802 | 98125 |
| 8691788 | FL1000001 | 00000002 | | 66500 | 2-Jan-2022 | Debit Card | F02 |  | 1098125 |

**DIM\_VEHICLE**

**Description**: Contains details about the vehicles.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| VIN | Vehicle identifier | Text |
| Vechile\_Producer | Manufacturer of the vehicle | Text |
| Vechile\_Model | Model of the vehicle | Text |
| Vechile\_Year | Year of manufacture | Int |
| Vechile\_Electric\_Type | Type of electric vehicle | Text |
| Vechile\_Electric\_Range | Vehicle's electric range in miles (how far vehicle can travel on purely electric) | Int |
| Vechile\_Type\_CAFV | Clean Alternative Fuel Vehicle type | Text |
| Vechile\_Color | Color of the vehicle | Text |
| Vechile\_Transmission | Transmission type (manual/automatic) | Text |

Example with filled data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VIN | Producer | Model | Year | | Electric\_Type | Electric\_Range | Type\_CAFV | Color | Transmission |
| FL1000001 | TESLA | MODEL 3 | 2020 | Battery Electric Vehicle (BEV) | | 266 | Clean Alternative | White | Automatic |
| 5YJ3E1EA5L | NISSAN | LEAF | 2021 | Plug-in Hybrid Electric Vehicle (PHEV) | | 395 | Clean Alternative Fuel Vehicle Eligible |  |  |

**DIM\_CUSTOMER**

**Description**: Contains details about the customer.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Customer\_Id | Unique customer identifier | Serial |
| Customer\_Passport | Unique customer identifier | Text |
| Customer\_Name | Name of the customer | Text |
| Customer\_Gender | Gender of the customer | Text |
| Customer\_Age | Age of the customer | Int |

Example with filled data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer\_ID | Customer\_Passport | Customer\_Name | Customer\_Gender | Customer\_Age |
| 00000001 |  | Geraldine |  | 30 |
| 00000002 | AC00128 | Naomi | Female |  |

**DIM\_LOCATION**

**Description**: Contains details about the location.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Postal\_code | Unique location identifier | Int |
| State | Name of state (WA or FL) | Text |
| City | Name of the city | Text |

Example with filled data

|  |  |  |
| --- | --- | --- |
| Postal\_code | State | City |
| 98125 | WA | Seattle |
| 1098125 | FL | Castle Rock |

**DIM\_EMPLOYEE**

**Description:** Contains details about the employees managing the transactions.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Employee\_Id | Unique employee identifier | Serial |
| Employee\_Name | Name of the employee | Text |
| Employee\_Email | Email of the employee | Text |
| Employee\_Phone | Phone number of the employee | Text |

Example with filled data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Employee\_Id | Employee\_Name | Employee\_Phone | | Employee\_Email |
| E01 | Michelle Harring | 1447868801 |  | |
| E02 | Lori Bell |  | lori@gmail.com | |

**DIM\_SALES\_CHANNEL**

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Channel\_Id | Unique channel identifier | Serial |
| Channel\_Type | Type of Channel (e.g., Online, Showroom, Event, Third\_party) | Text |
| Channel\_Desc | Additional notes | Text |

Example with filled data

|  |  |  |
| --- | --- | --- |
| Channel\_Id | Channel\_Type | Channel\_Desc |
| CH0058376 | Online | Sales made through Website. |
| CH0058377 | Third-party | Sales via third-party resellers like Vroom. |

# Business Layer 3NF

1. **Ensure data integrity**

* Break down entities into logical groupings: Producers, models, vehicles, locations, cities, states, customers, employees, channels, transactions are represented by distinct tables.
* Ensure attributes are atomic and eliminate redundancy.
* Use surrogate keys for relationships instead of natural keys for scalability and performance. On 3NF, we will have two fields in each table - \_src\_id and \_id (for example, employee\_id and employee\_src\_id), where \_src\_id will be a natural key and \_id - a surrogate (for SCD2 table – \_id + start\_dt).

1. **Add SCD**

For entities like CE\_EMPLOYEES\_SCD, implement Type 2 Slowly Changing Dimensions (SCD) to track historical data: Fields like START\_DT, END\_DT, and IS\_ACTIVE maintain the history of employee records.

1. **Add Metadata Columns**

* Include SOURCE TRIPLET fields like SOURCE\_System, SOURCE\_Entity, and SOURCE\_ID for traceability.
* Add TA\_INSERT\_DT and TA\_UPDATE\_DT timestamps to support auditing.

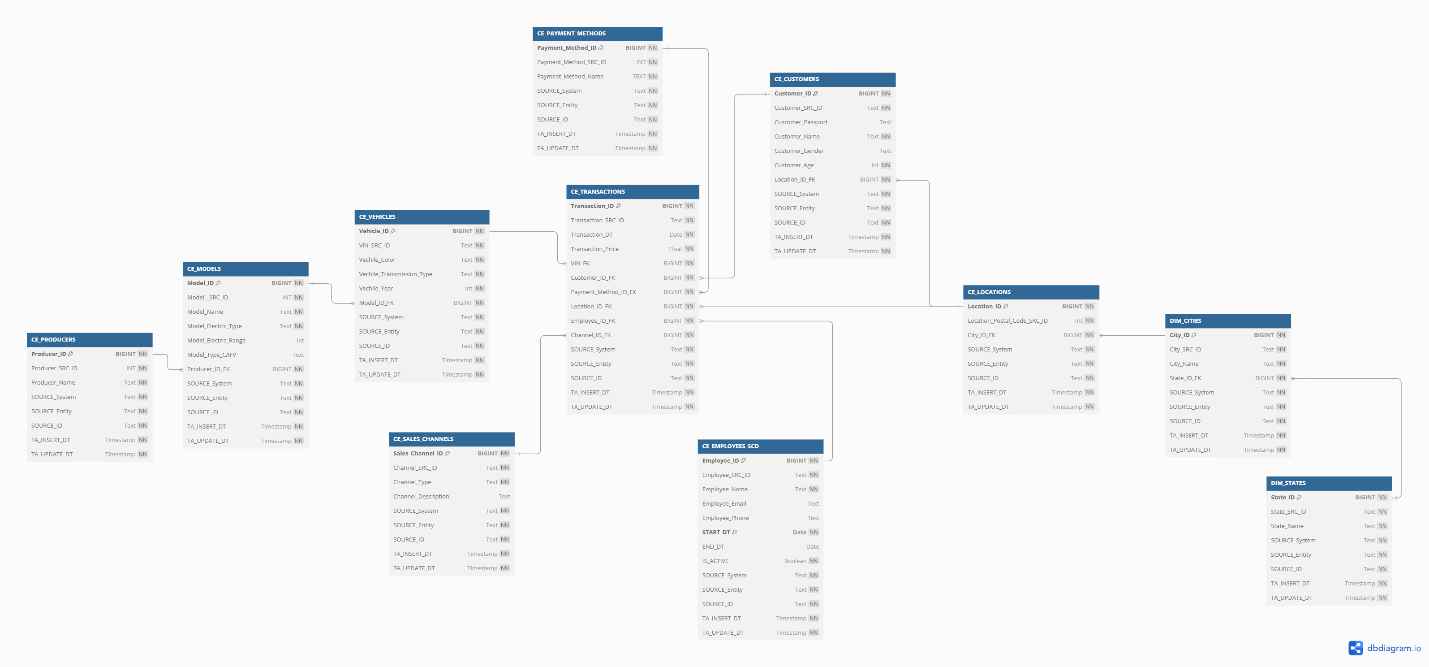
1. **Data Types and Constraints**

* Assign appropriate data types based on expected usage:
* BIGINT: For surrogate keys.
* Text: For descriptive fields (e.g., names, descriptions).
* Date and Timestamp: For time data.
* Int: For age
* Float: For price
* Constraints: NOT NULL for mandatory fields, PK for surrogate keys, FK for establish connection between tables, UNIQUE for name of payment method.

1. **Define Relationships**

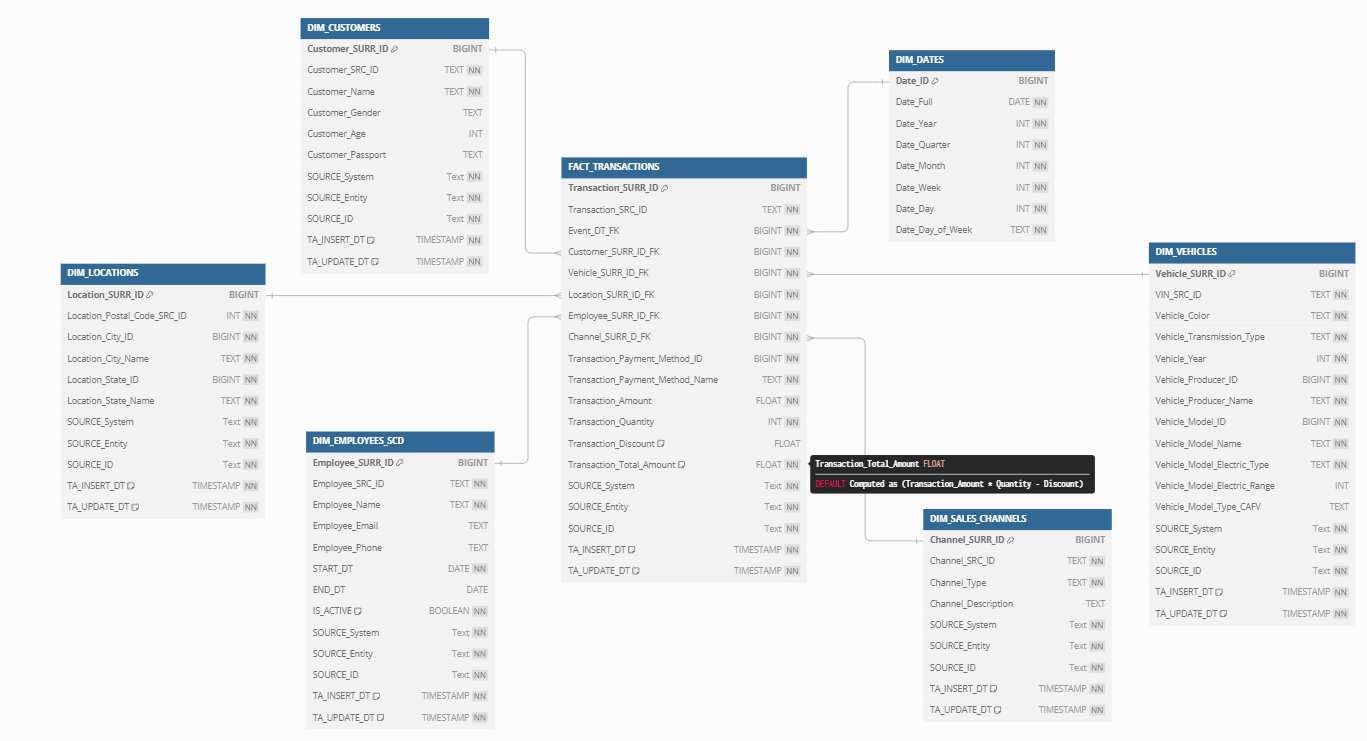
1-to-Many Relationships: A **transaction** is associated with exactly **one vehicle**, **one customer**, **one employee**, **one sales channel**, and **one location**.

* One vehicle may appear in multiple transactions.
* One customer can make many transactions.
* One employee can record many transactions.
* One sales channel can facilitate many transactions.
* One location can host many transactions.
* A producer can produce many models.
* A model can be associated with many vehicles.
* An one payment method can be applied to transaction.
* A state can have many cities.
* A city can have many locations.



# Business Layer Dimensional Model

The fact table FACT\_TRANSACTIONS captures the sales transactions with metrics such as transaction amount, quantity, and discount. It references multiple dimension tables like DIM\_CUSTOMERS, DIM\_VEHICLES, DIM\_LOCATIONS, DIM\_EMPLOYEES, DIM\_SALES\_CHANNELS, and DIM\_DATES to provide contextual information for each transaction. They are connected to the fact table using foreign keys.



In the star schema outlined above, the metrics refer to the numerical data points captured in the fact table (FACT\_TRANSACTIONS). These metrics are the key performance indicators (KPIs) that represent the transactional data in the context of sales transactions. Here's a detailed description of each metric:

Metrics in the FACT\_TRANSACTIONS Table

**Transaction\_Amount**

* Description: Represents the base price of a vehicle at the time of sale. This is the amount before any discounts are applied and is usually the agreed-upon price for the vehicle.
* Data Type: FLOAT
* Use Case: Used to calculate total revenue for the company, excluding discounts. It's a fundamental metric for tracking sales performance.

**Transaction\_Quantity**

* Description: The number of units (vehicles) sold in the transaction. In most cases, this would be 1 (one vehicle), but in cases like bulk sales or fleet sales, this could be greater.
* Data Type: INT
* Use Case: This metric helps in understanding how many vehicles were sold in a given transaction, useful for reporting on total units sold.

**Transaction\_Discount**

* Description: The total discount applied to the transaction. This discount can be in the form of a percentage or a fixed amount off the original price.
* Data Type: FLOAT
* Use Case: Used to calculate how much revenue is lost due to discounts. This is crucial for assessing the impact of discounts on profitability.

**Transaction\_Total\_Amount**

* Description: This is the total value of the transaction after applying the discount to the base transaction amount. The formula is: Transaction\_Total\_Amount = (Transaction\_Amount \* Transaction\_Quantity) - Transaction\_Discount
* Data Type: FLOAT
* Use Case: This is a key metric for understanding the total revenue generated from each transaction. It's the amount that will be invoiced to the customer, and is a fundamental metric for financial analysis and reporting.

# Logical Scheme

A logical model for the data warehouse (DWH) load process include:

Data Sources → ETL Process → Staging Area → Data Transformation & Cleansing → 3NF Relational Layer → Dimensional Modeling → Dimension Layer → Aggregated & Indexed Data → Data Marts → End-User Tools (BI, Reports, Dashboards)

1. DATA SOURCES

Raw data originates from various external sources (databases, files, APIs).

Stored in a structured or semi-structured format before processing.

1. ETL (Extract, Transform, Load)

Extracts data from source systems.

Loads it into a staging area for further processing.

1. Data Staging

A temporary storage area for raw data before transformation.

Helps with data cleansing, deduplication, and validation.

1. Data Transformation & Cleansing

Data is cleaned, formatted, and structured.

Ensures accuracy, consistency, and completeness.

1. 3NF Relational Layer (Normalized Data)

Stores data in a 3rd Normal Form (3NF) schema.

Maintains referential integrity and eliminates redundancy.

1. Dimensional Modeling

Transforms normalized data into a format optimized for business intelligence (BI) reporting.

Uses Fact & Dimension Tables (Star Schema, Snowflake Schema).

1. Aggregated & Indexed Data

Data is pre-aggregated and indexed for faster query performance.

1. Data Marts

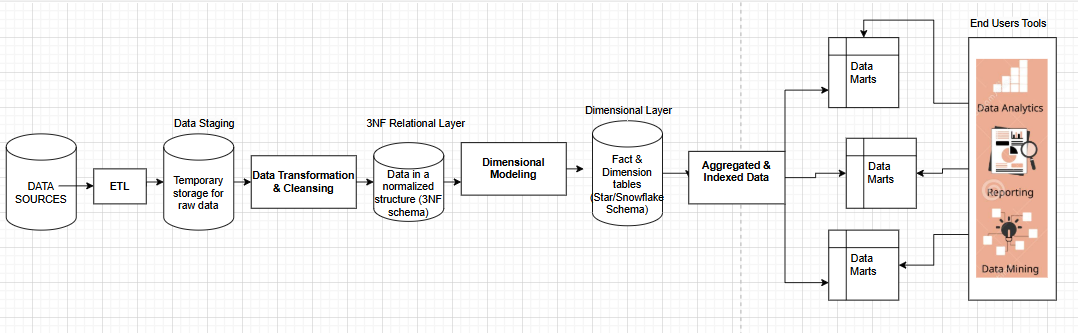
Smaller, department-specific data storage optimized for reporting.

Multiple data marts serve different business functions.

1. End-User Tools

Data is consumed by Business Intelligence & Analytics tools such as:

* Data Analytics – Insights & trend analysis
* Reporting – Dashboards, KPI monitoring
* Data Mining – Advanced pattern recognition & forecasting



# Data Flow

This diagram represents how data flows from raw data sources to a structured dimensional model for analytics and reporting.

1. DATA SOURCES

Input files:

* Electric\_Vehicle\_Population\_Data\_1.csv
* Electric\_Vehicle\_Population\_Data\_2.csv

Data is extracted and loaded into the staging layer.

1. STAGING LAYER

Temporary storage for raw data before transformation. Contains:

* External Tables (ext\_sales\_fl, ext\_sales\_wa).
* Source Tables (src\_sales\_fl, src\_sales\_wa).

1. BL\_3NF (Normalized Layer)

Stores data in a 3rd Normal Form (3NF) relational model for consistency. Includes tables such as:

* CE\_PRODUCERS, CE\_MODELS, CE\_VEHICLES
* CE\_STATES, CE\_CITIES, CE\_LOCATIONS
* CE\_CUSTOMERS, CE\_PAYMENT\_METHODS, CE\_EMPLOYEES\_SCD
* CE\_SALES\_CHANNELS, CE\_TRANSACTIONS

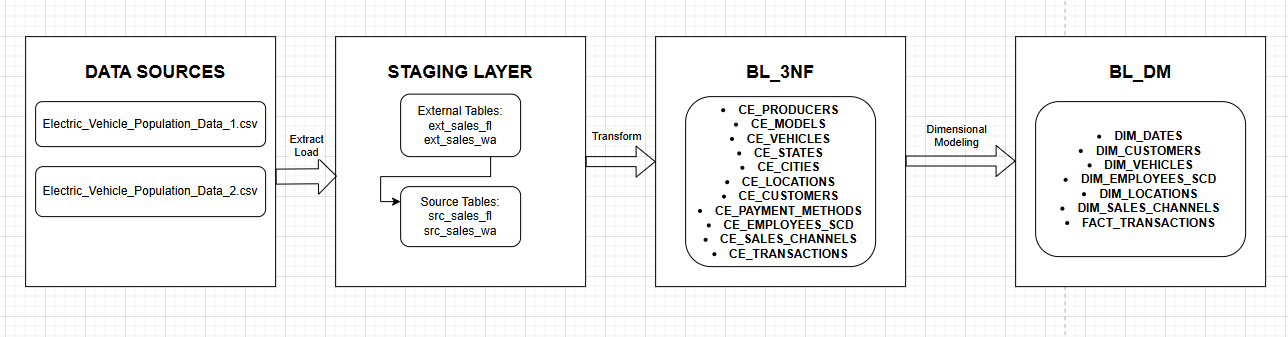
Data is transformed before moving to the next layer.

1. BL\_DM (Dimensional Model Layer)

Stores data in a denormalized structure for analytical queries (fact/dimension tables). Tables include:

* DIM\_DATES, DIM\_CUSTOMERS, DIM\_VEHICLES
* DIM\_EMPLOYEES\_SCD, DIM\_LOCATIONS, DIM\_SALES\_CHANNELS
* FACT\_TRANSACTIONS

This layer is optimized for Business Intelligence (BI) & reporting.



# Fact Table Partitioning Strategy