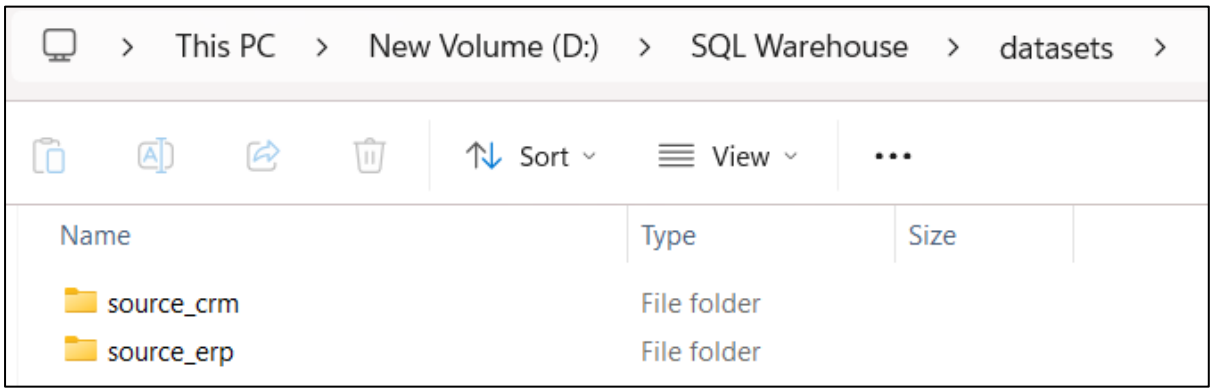
NovaMart Data Warehouse

Project Overview

This project focuses on designing and implementing a complete SQL-based Data Warehouse from scratch using MySQL. The objective was to convert raw operational data from multiple source systems into a clean, integrated, and analytics-ready data model following industry best practices.

The project simulates a real-world data engineering workflow, covering data ingestion, data cleaning, transformation, integration, and dimensional modeling. The final output is a star schema that can be directly consumed by BI tools such as Power BI.

Source Systems and Dataset



The project uses six CSV files originating from two source systems:

CRM System

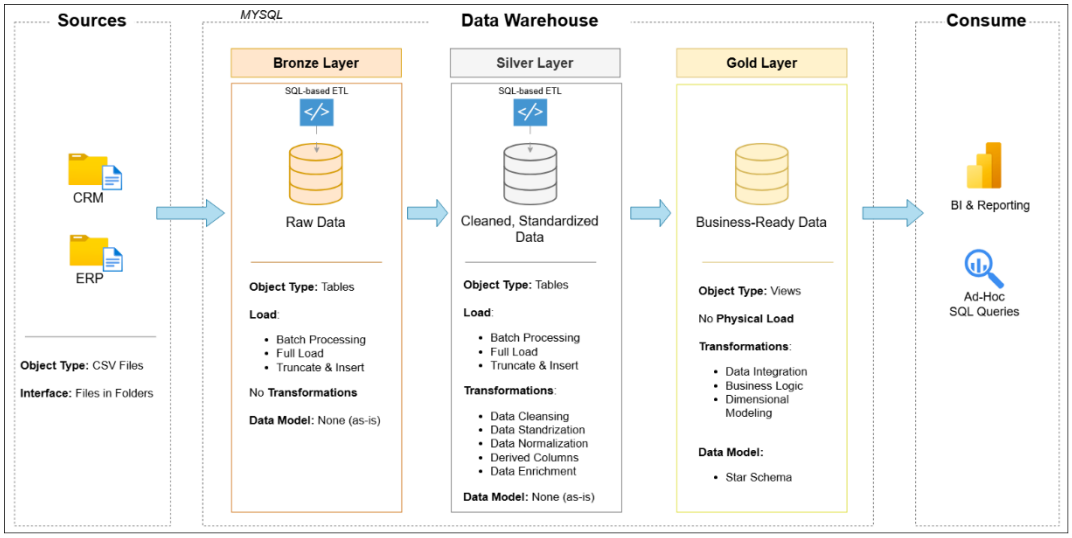
- Customer Information
- Product Information
- Sales Transactions

ERP System

- Customer Demographics
- Location Data
- Product Categories

These datasets contained raw, inconsistent, and partially invalid data, including duplicate records, inconsistent codes, missing values, and invalid dates.

Data Warehouse Architecture



The Data Warehouse is implemented using a layered architecture:

- **Bronze Layer**: Raw data ingestion
- **Silver Layer**: Cleaned and standardized data
- **Gold Layer**: Business-ready star schema

All layers are implemented within a single MySQL database named NovaMart_DW.

Bronze Layer (Raw Ingestion)

cst_id	cst_key	cst_firstname	cst_lastname	cst_marital_status	cst_gndr	cst_create_date
11000	AW00011000	Jon	Yang	M	M	2025-10-06
11001	AW00011001	Eugene	Huang	S	M	2025-10-06
11002	AW00011002	Ruben	Torres	M	M	2025-10-06
11003	AW00011003	Christy	Zhu	S	F	2025-10-06
11004	AW00011004	Elizabeth	Johnson	S	F	2025-10-06
11005	AW00011005	Julio	Ruiz	S	M	2025-10-06
11006	AW00011006	Janet	Alvarez	S	F	2025-10-06
11007	AW00011007	Marco	Mehta	M	M	2025-10-06
11008	AW00011008	Rob	Verhoff	S	F	2025-10-06
11009	AW00011009	Shannon	Carlson	S	M	2025-10-06
11010	AW00011010	Jacquelyn	Suarez	S	F	2025-10-06
11011	AW00011011	Curtis	Lu	M	M	2025-10-06
11012	AW00011012	Lauren	Walker	M	F	2025-10-06

The Bronze layer stores data exactly as received from the source systems.

Key characteristics:

- One table per CSV file
- No transformations or business rules applied
- Preserves original data issues for traceability

Implementation details:

- Tables created using MySQL
- Data loaded using LOAD DATA LOCAL INFILE
- Dates and numeric fields ingested as-is
- Acts as a historical and audit layer

Silver Layer (Data Cleaning and Transformation)

cst_id	cst_key	cst_firstname	cst_lastname	cst_marital_status	cst_gndr	cst_create_date	dwh_create_date
11000	AW00011000	Jon	Yang	Married	Male	2025-10-06	2026-02-09 20:18:20
11001	AW00011001	Eugene	Huang	Single	Male	2025-10-06	2026-02-09 20:18:20
11002	AW00011002	Ruben	Torres	Married	Male	2025-10-06	2026-02-09 20:18:20
11003	AW00011003	Christy	Zhu	Single	Female	2025-10-06	2026-02-09 20:18:20
11004	AW00011004	Elizabeth	Johnson	Single	Female	2025-10-06	2026-02-09 20:18:20
11005	AW00011005	Julio	Ruiz	Single	Male	2025-10-06	2026-02-09 20:18:20
11006	AW00011006	Janet	Alvarez	Single	Female	2025-10-06	2026-02-09 20:18:20
11007	AW00011007	Marco	Mehta	Married	Male	2025-10-06	2026-02-09 20:18:20
11008	AW00011008	Rob	Verhoff	Single	Female	2025-10-06	2026-02-09 20:18:20
11009	AW00011009	Shannon	Carlson	Single	Male	2025-10-06	2026-02-09 20:18:20
11010	AW00011010	Jacquelyn	Suarez	Single	Female	2025-10-06	2026-02-09 20:18:20
11011	AW00011011	Curtis	Lu	Married	Male	2025-10-06	2026-02-09 20:18:20

The Silver layer is responsible for data quality and standardization.

Key transformations performed:

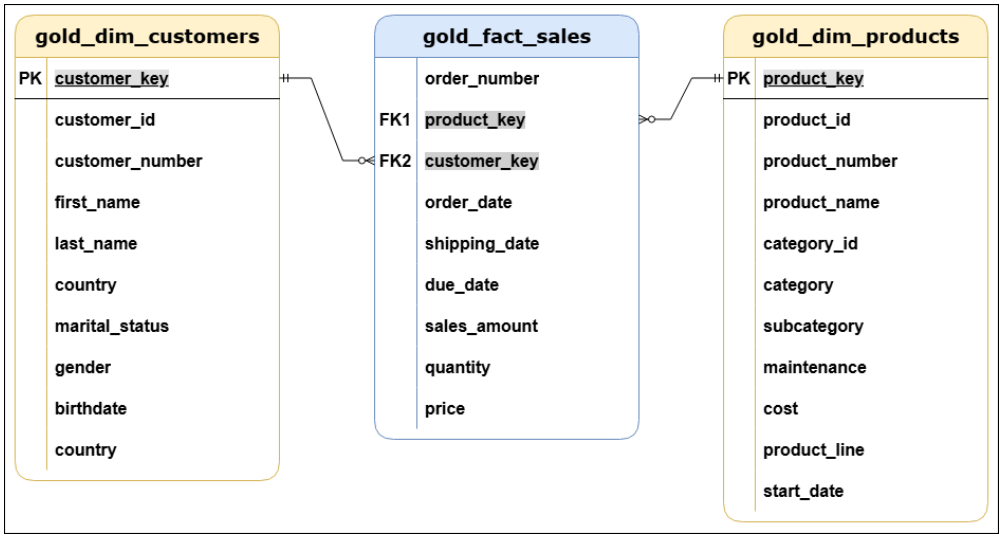
- Removal of duplicate customer records using window functions
- Trimming unwanted spaces from text fields
- Normalization of coded values (gender, marital status, product line, country)
- Handling invalid and missing values using NULL and default values
- Conversion of integer-based dates into proper DATE format
- Fixing invalid dates such as 0000-00-00
- Recalculation of incorrect sales values using quantity and price
- Standardization of customer and product keys for integration

The Silver layer ensures that data is consistent, reliable, and ready for modeling.

Gold Layer (Dimensional Modeling)

The Gold layer represents the final analytical data model. Instead of physical tables, MySQL views are created.

Star Schema Design



The Gold layer follows a star schema consisting of:

Dimensions

- dim_customers: Customer attributes, demographics, and location
- dim_products: Product details, categories, and attributes

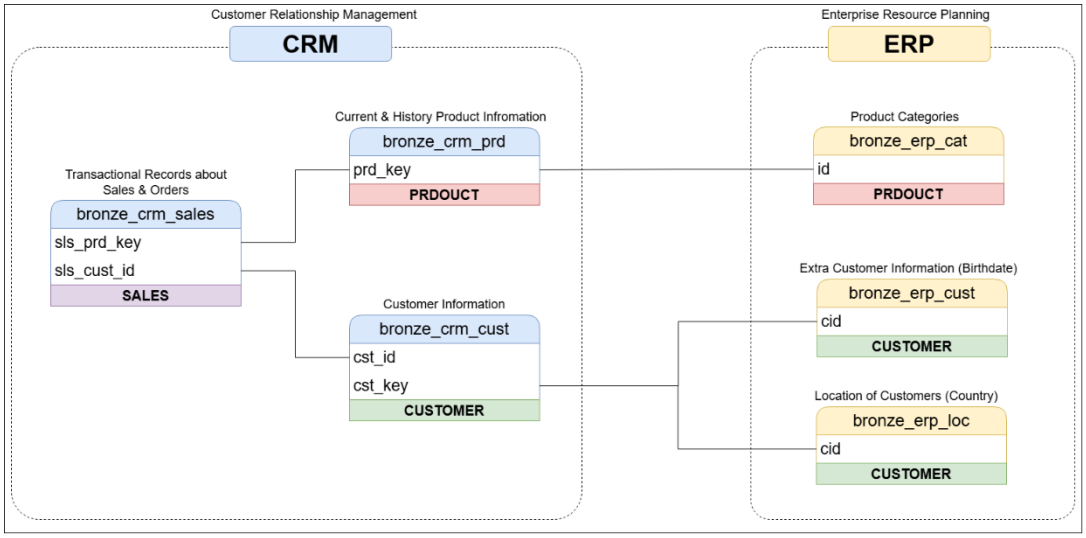
Fact

- fact_sales: Sales transactions linked to customers and products

Key features:

- Surrogate keys generated using ROW_NUMBER()
- Business keys replaced with surrogate keys
- Only current (active) product records included
- Fact table contains measurable metrics such as sales amount, quantity, and price

Data Integration



Data from CRM and ERP systems is integrated in the Silver and Gold layers:

- CRM is treated as the primary source for customer identity
- ERP enriches customer data with demographics and location
- Product category data from ERP is merged with CRM product data
- Referential integrity validated through join checks

Validation and Quality Checks

```
88 •  SELECT *
89    FROM gold_fact_sales f
90    LEFT JOIN gold_dim_customers c
91    ON c.customer_key = f.customer_key
92    LEFT JOIN gold_dim_products p
93    ON p.product_key = f.product_key
94    WHERE p.product_key IS NULL OR c.customer_key IS NULL;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

order_number	product_key	customer_key	order_date	shipping_date	due_date	sales_amount	quantity	price	customer_key	customer_id	customer_number	first_name	last_name	country	mari

To ensure correctness:

- Orphan record checks were performed between fact and dimensions
- Surrogate key uniqueness was validated
- Data consistency across joins was verified
- Sample reconciliation queries were executed between layers

Final Output

The final deliverables of the project are three analytics-ready views:

1. gold_dim_customers

customer_key	customer_id	customer_number	first_name	last_name	country	marital_status	gender	birthdate	create_date
1	11000	AW00011000	Jon	Yang	Australia	Married	Male	1971-10-06	2025-10-06
2	11001	AW00011001	Eugene	Huang	Australia	Single	Male	1976-05-10	2025-10-06
3	11002	AW00011002	Ruben	Torres	Australia	Married	Male	1971-02-09	2025-10-06
4	11003	AW00011003	Christy	Zhu	Australia	Single	Female	1973-08-14	2025-10-06
5	11004	AW00011004	Elizabeth	Johnson	Australia	Single	Female	1979-08-05	2025-10-06
6	11005	AW00011005	Julio	Ruiz	Australia	Single	Male	1976-08-01	2025-10-06
7	11006	AW00011006	Janet	Alvarez	Australia	Single	Female	1976-12-02	2025-10-06
8	11007	AW00011007	Marco	Mehta	Australia	Married	Male	1969-11-06	2025-10-06
9	11008	AW00011008	Rob	Verhoff	Australia	Single	Female	1975-07-04	2025-10-06
10	11009	AW00011009	Shannon	Carlson	Australia	Single	Male	1969-09-29	2025-10-06

2. gold_dim_products

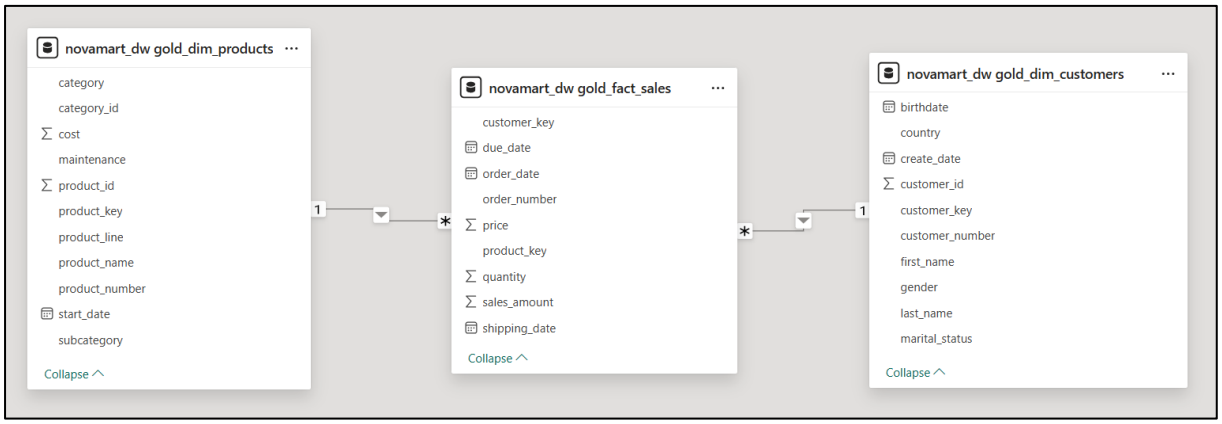
product_key	product_id	product_number	product_name	category_id	category	subcategory	maintenance	cost	product_line	start_date
1	210	FR-R92B-58	HL Road Frame - Black- 58	CO_RF	Components	Road Frames	Yes	0	Road	2003-07-01
2	211	FR-R92R-58	HL Road Frame - Red- 58	CO_RF	Components	Road Frames	Yes	0	Road	2003-07-01
3	348	BK-M82B-38	Mountain-100 Black- 38	BI_MB	Bikes	Mountain Bikes	Yes	1898	Mountain	2011-07-01
4	349	BK-M82B-42	Mountain-100 Black- 42	BI_MB	Bikes	Mountain Bikes	Yes	1898	Mountain	2011-07-01
5	350	BK-M82B-44	Mountain-100 Black- 44	BI_MB	Bikes	Mountain Bikes	Yes	1898	Mountain	2011-07-01
6	351	BK-M82B-48	Mountain-100 Black- 48	BI_MB	Bikes	Mountain Bikes	Yes	1898	Mountain	2011-07-01
7	344	BK-M82S-38	Mountain-100 Silver- 38	BI_MB	Bikes	Mountain Bikes	Yes	1912	Mountain	2011-07-01
8	345	BK-M82S-42	Mountain-100 Silver- 42	BI_MB	Bikes	Mountain Bikes	Yes	1912	Mountain	2011-07-01
9	346	BK-M82S-44	Mountain-100 Silver- 44	BI_MB	Bikes	Mountain Bikes	Yes	1912	Mountain	2011-07-01
10	347	BK-M82S-48	Mountain-100 Silver- 48	BI_MB	Bikes	Mountain Bikes	Yes	1912	Mountain	2011-07-01

3. gold_fact_sales

order_number	product_key	customer_key	order_date	shipping_date	due_date	sales_amount	quantity	price
SO43697	20	10769	2010-12-29	2011-01-05	2011-01-10	3578	1	3578
SO43698	9	17390	2010-12-29	2011-01-05	2011-01-10	3400	1	3400
SO43699	9	14864	2010-12-29	2011-01-05	2011-01-10	3400	1	3400
SO43700	41	3502	2010-12-29	2011-01-05	2011-01-10	699	1	699
SO43701	9	4	2010-12-29	2011-01-05	2011-01-10	3400	1	3400
SO43702	16	16646	2010-12-30	2011-01-06	2011-01-11	3578	1	3578
SO43703	20	5625	2010-12-30	2011-01-06	2011-01-11	3578	1	3578
SO43704	6	6	2010-12-30	2011-01-06	2011-01-11	3375	1	3375
SO43705	7	12	2010-12-30	2011-01-06	2011-01-11	3400	1	3400
SO43706	17	16622	2010-12-31	2011-01-07	2011-01-12	3578	1	3578

These views are the only objects intended for BI consumption.

Power BI Usage



In Power BI:

- Connect directly to the MySQL database
- Import the three Gold views
- Build relationships using surrogate keys
- Perform reporting and analytics on top of the star schema

Bronze and Silver layers are not used in Power BI.

Key Skills Demonstrated

- SQL Data Warehousing
- ETL implementation using MySQL
- Data quality handling
- Multi-source data integration
- Dimensional modeling (Star Schema)
- Business-ready data preparation

Project Outcome

This project delivers a complete end-to-end Data Warehouse pipeline that transforms raw operational data into structured, reliable, and analytics-ready datasets. It closely mirrors how modern data warehouses are built and maintained in real-world organizations.