Data Warehouse User Manual

---

1. Introduction

The datawarehouse database is a central repository for enterprise data supporting:

- Analytics
- Business Intelligence
- Reporting
- Machine Learning

Audience: Data Analysts, BI Developers, Data Scientists, ML Engineers, Database Architects.

Data Layers:

- **Bronze:** Raw data ingestion from CSV/external sources.
- **Silver:** Data cleaning, deduplication, type casting, basic transformations.
- **Gold:** Refined, business-ready data for dashboards, reporting, and ML features.

---

2. Database Architecture

Schemas:

- bronze → Raw data tables
- silver → Cleaned and transformed tables
- gold → Fact and dimension views

Fact & Dimension Tables:

- Gold Layer Fact: gold.fact_sales
- Gold Layer Dimensions: gold.dim_customers, gold.dim_products

Data Flow:

1. Bronze Layer → Extract CSV → Load raw tables.
2. Silver Layer → Transform data, deduplicate, cast types, clean invalid data.
3. Gold Layer → Build views for reporting and analytics.

---

3. Bronze Layer – Raw Data

Tables:

Table	Description	Key Columns
crm_cust_info	Raw CRM customer data	cst_id, cst_key
crm_prd_info	Raw product master	prd_id, prd_key
crm_sales_details	Raw sales transactions	sls_ord_num, sls_prd_key, sls_cust_id
erp_cust_az12	ERP customer info	cid, bdate, gen
erp_loc_a101	ERP customer location	cid, cntry
erp_px_cat_g1v2	Product categories	id, cat, subcat, maintenance

ETL Procedure: bronze.load_bronze

Notes: Contains raw, uncleaned data. Analysts should generally not query Bronze directly.

---

4. Silver Layer – Transformed Data

Tables:

Table	Description
crm_cust_info	Deduplicated and cleaned customer master
crm_prd_info	Product master with categories and mappings
crm_sales_details	Cleaned sales transactions with corrected prices and dates
erp_cust_az12	ERP customer info, cleaned
erp_loc_a101	Cleaned customer location data
erp_px_cat_g1v2	Product category mapping

Transformations:

- Deduplication (ROW_NUMBER() used per key)
- Standardizing codes (e.g., M → Mountain, F → Female)
- Correcting invalid dates and numeric fields

- Calculating corrected sales and prices

ETL Procedure: silver.load_silver

---

5. Gold Layer – Business Ready

Views:

View	Description
dim_customers	Customer dimension with enriched ERP and location info
dim_products	Product dimension with category, subcategory, product line
fact_sales	Fact table linking products and customers with sales

Business Rules:

- Only current product versions (prd_end_dt IS NULL)
- Deduplicated customers
- Fact table joins customers and products for easy BI reporting

---

6. Sample Queries

Data Analysts:

```
-- Total sales per customer
SELECT dc.customer_number, SUM(fs.sales) AS total_sales
FROM gold.fact_sales fs
JOIN gold.dim_customers dc
  ON fs.customer_key = dc.customer_key
GROUP BY dc.customer_number;


-- Product revenue by category
SELECT dp.category, SUM(fs.sales) AS revenue
FROM gold.fact_sales fs
JOIN gold.dim_products dp
```

```
 ON fs.product_key = dp.product_key
GROUP BY dp.category;
```

Business Intelligence / Dashboard:

- Use fact_sales with dimensions dim_customers and dim_products
- Star Schema:

```
dim_customers
    |
    |
fact_sales ---- dim_products
```

Data Scientists / ML:

```
SELECT dc.customer_key,
    dp.category,
    COUNT(fs.order_number) AS purchase_count,
    SUM(fs.sales) AS total_spent
FROM gold.fact_sales fs
JOIN gold.dim_customers dc ON fs.customer_key = dc.customer_key
JOIN gold.dim_products dp ON fs.product_key = dp.product_key
GROUP BY dc.customer_key, dp.category;
```

---

7. Best Practices

- Use **Gold** for reporting, **Silver** for historical analysis, **Bronze** for raw data only
- Naming conventions: lowercase, underscores, schema.table_name
- Document all transformations for reproducibility

---

8. Troubleshooting

Issue	Cause	Fix
NULL/invalid dates	Incorrect CSV	Clean in Silver layer

Issue	Cause	Fix
Duplicate customers	Multiple raw entries	Deduplicate using ROW_NUMBER()
Incorrect sales/prices	Invalid input	Recalculate as quantity * ABS(price)
Slow queries	Large fact tables	Index customer_key, product_key

---

9. Tools Used

- Microsoft SQL Server Management Studio
- Draw.io (for diagrams)
- Notion (for maintaining track of tasks)
- GPT-5 (for manual creation)