# Data Warehouse SQL & Analytics Project

**An Overview of Data Insights**

**Vanshika kumar**

Aspiring Data Analyst

# Data Warehouse Project Overview

## 🚀 Welcome to the Data Warehouse and Analytics Project

**This project showcases a comprehensive data warehousing and analytics solution, demonstrating the full data journey from data ingestion and transformation to modeling and actionable business insights.**
**It is designed as a portfolio project that highlights industry best practices in data engineering, data modeling, and analytics. The goal is to simulate a real-world enterprise data environment and demonstrate how raw data can be transformed into meaningful insights that drive decision-making.**

## 📊 Project Overview

**This project involves the following key components:**

### 🏗️ Data Architecture:

- **Designed a modern data warehouse using the Medallion Architecture (Bronze, Silver, and Gold layers) to ensure scalability, data quality, and efficient analytics workflows.**

### ⚙️ ETL Pipelines:

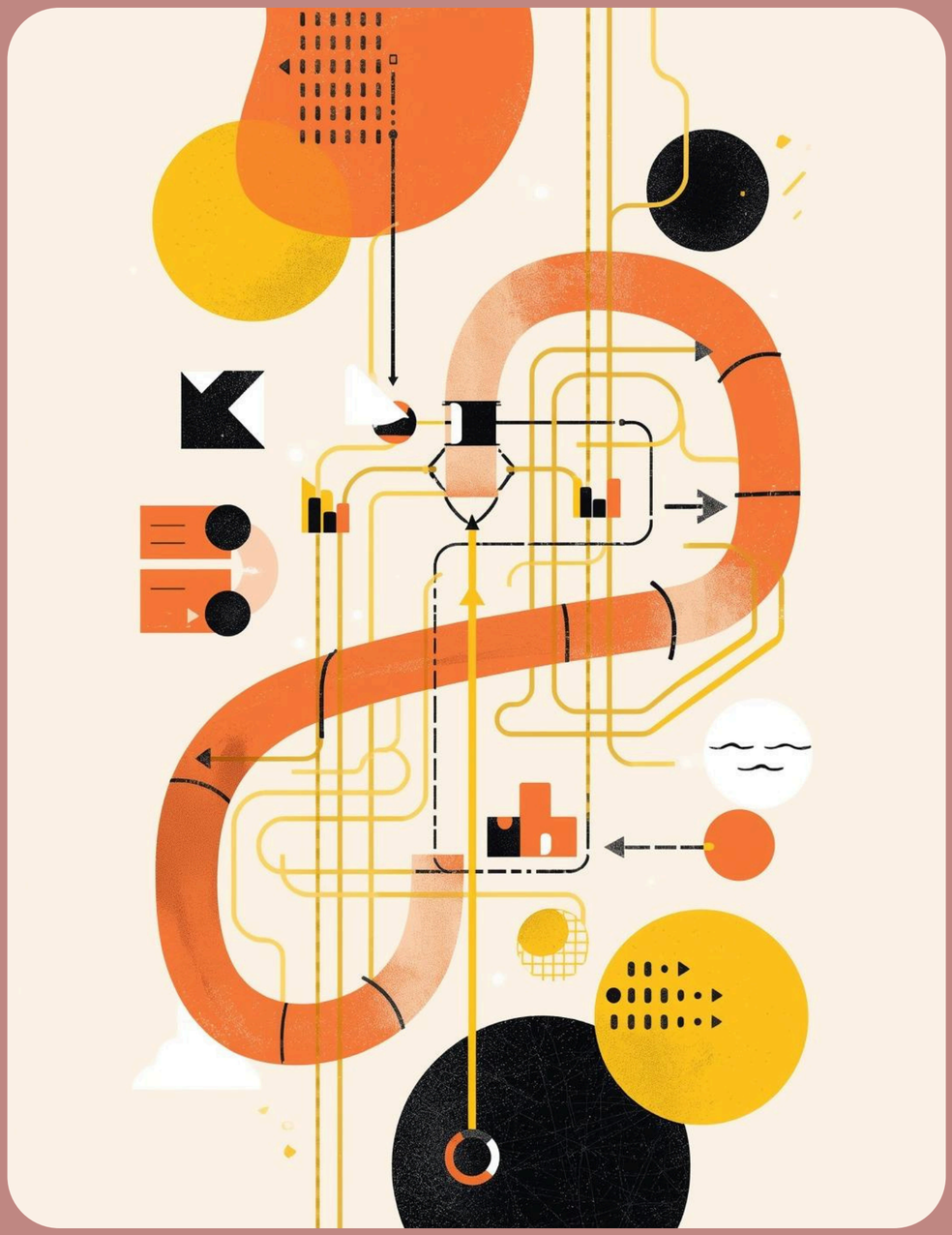
- **Developed robust Extract, Transform, Load (ETL) pipelines to ingest data from multiple sources, clean and transform it, and load it into the warehouse.**

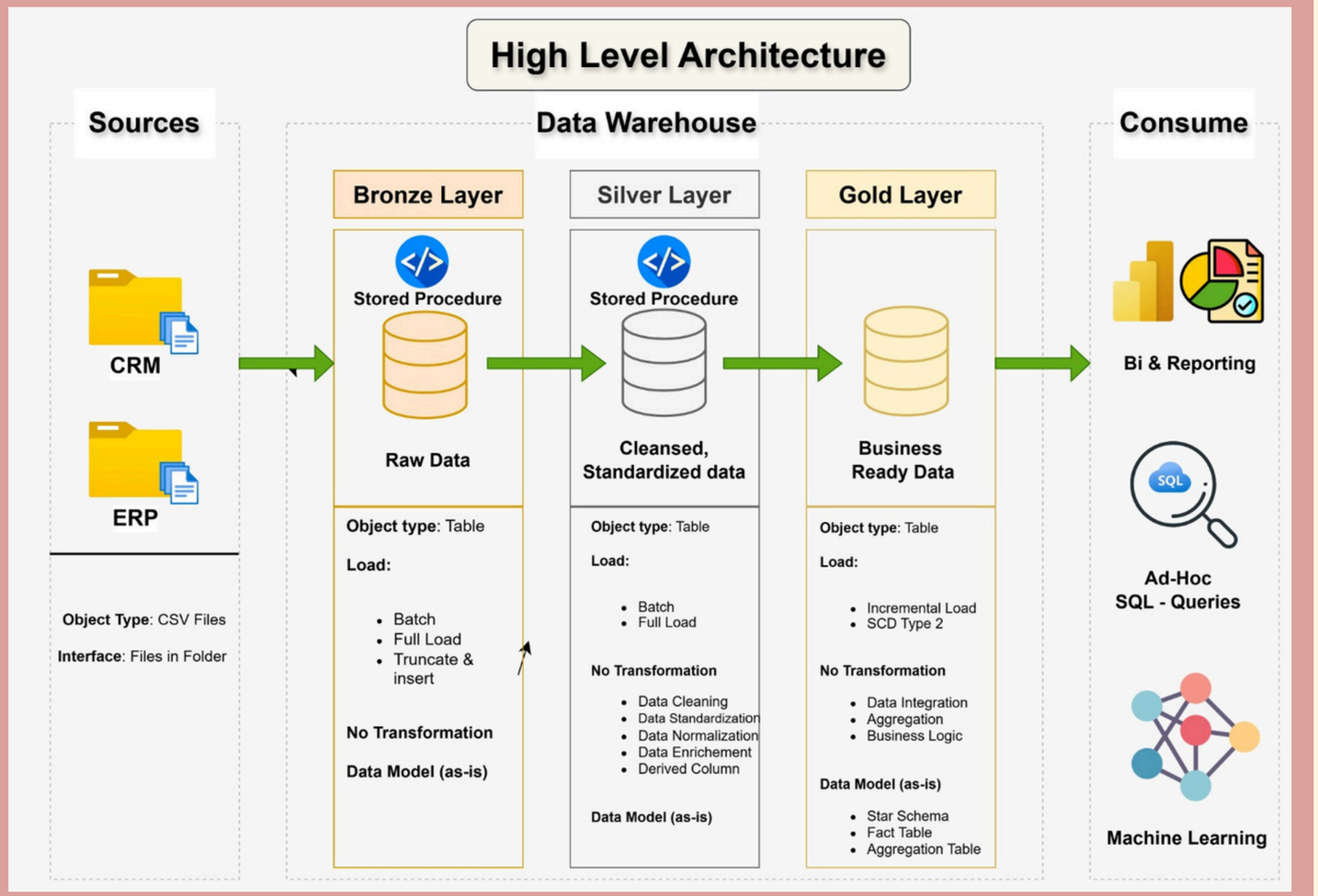
### 🧩 Data Modeling:

- **Implemented star schema and snowflake schema models by building fact and dimension tables optimized for analytical queries and reporting.**

### 📈 Analytics & Reporting:

- **Created SQL-based reports and interactive dashboards to generate actionable business insights and KPIs.**






***High Level Architecture***

# Data Architecture

The Data Architecture for this project follows the Medallion Architecture **Bronze, Silver and Gold layer**

## Bronze Layer

The Bronze layer stores raw data exactly as it is received from the source systems, without any transformations.

In this case, data is ingested from CSV files into the SQL Server database and preserved in its original form for traceability and auditing.

## Silver Layer

The Silver layer performs data cleansing, standardization, and normalization.

This step ensures data quality by handling missing values, correcting inconsistencies, applying business rules, and converting raw data into a structured and usable format.

The output from this layer is clean, standardized data that is ready for modeling and analysis.

## Gold Layer

The Gold layer contains business-ready, curated datasets optimized for analytics and reporting.

Here, the data is modeled into star schemas (facts and dimensions) to support dashboards, self-service BI, and advanced analytics.

This provides a trusted, consistent view of organizational data for decision-making.

# ETL Pipelines: Extract, Transform, Load

## Data Extraction

The **initial phase** involves gathering data from various sources, including databases, APIs, and files, ensuring it is accurate and relevant for further processing and analysis.

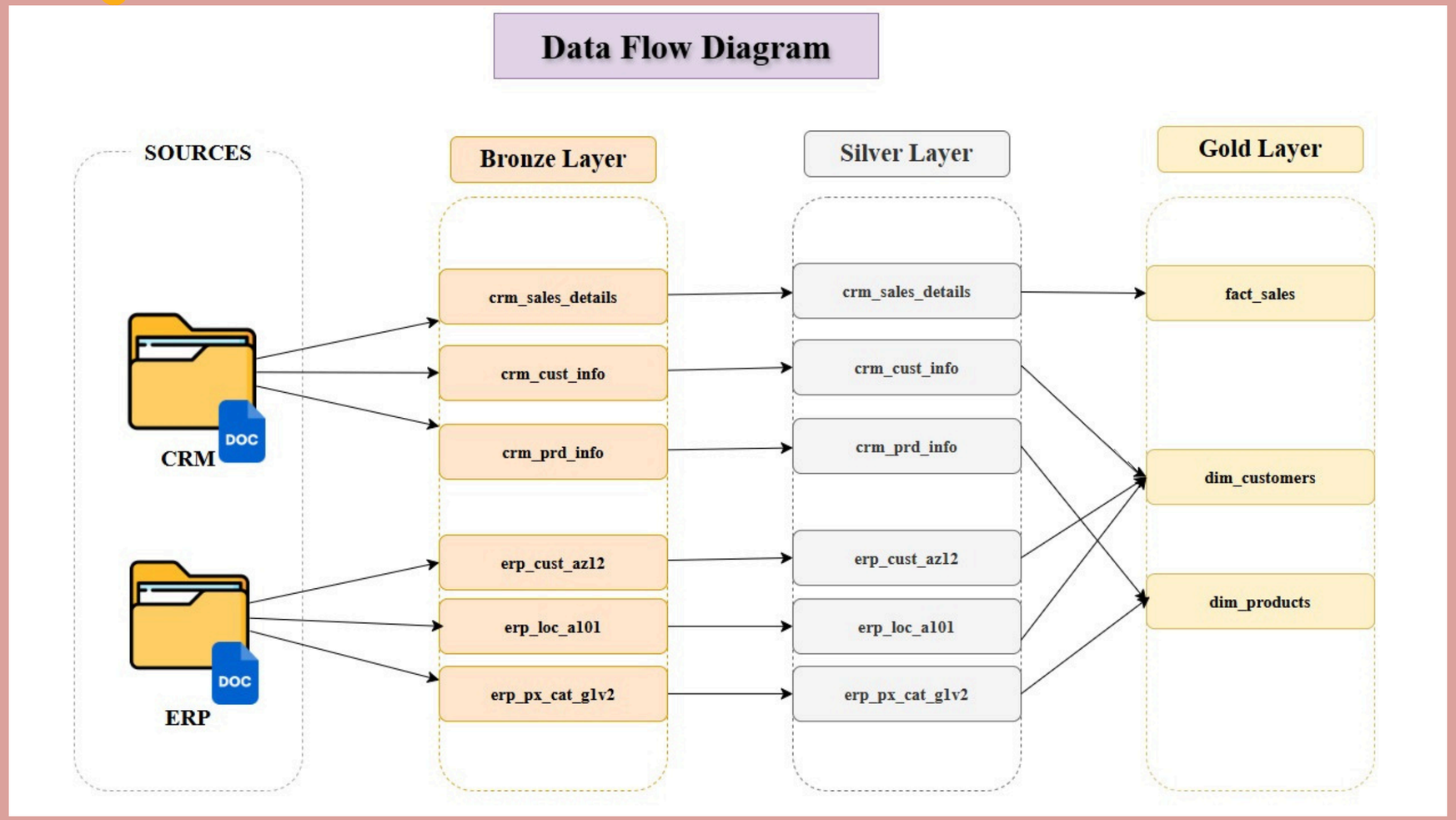
## Data Transformation

This step entails **cleaning and reshaping** the extracted data into a structured format, applying business rules, and enriching the data to enhance its quality and usability for analytics.

## Data Loading

Finally, the transformed data is **loaded into the data warehouse**, making it accessible for business intelligence and reporting, ensuring optimal performance and data integrity throughout the process.

# Data Flow




## ***Data Flow Explanation***

This diagram shows how data moves from CRM and ERP sources to the final analytical layer.

**Sources**: CRM and ERP provide raw data like sales, customers, products, and locations.

**Bronze Layer**: Stores raw data exactly as received from sources.

**Silver Layer**: Performs data cleaning and transformation, making it structured and consistent.

**Gold Layer**: Contains final, business-ready datasets for reporting and analytics, such as sales facts and customer/product dimensions.

# Integration Model

## Integration Model Explanation

This diagram shows how CRM and ERP data are integrated to create a unified view of customers, products, and sales.

**CRM System:**

Holds sales, product, and customer data.

crm_sales_details contains transactional sales records linked to products and customers.

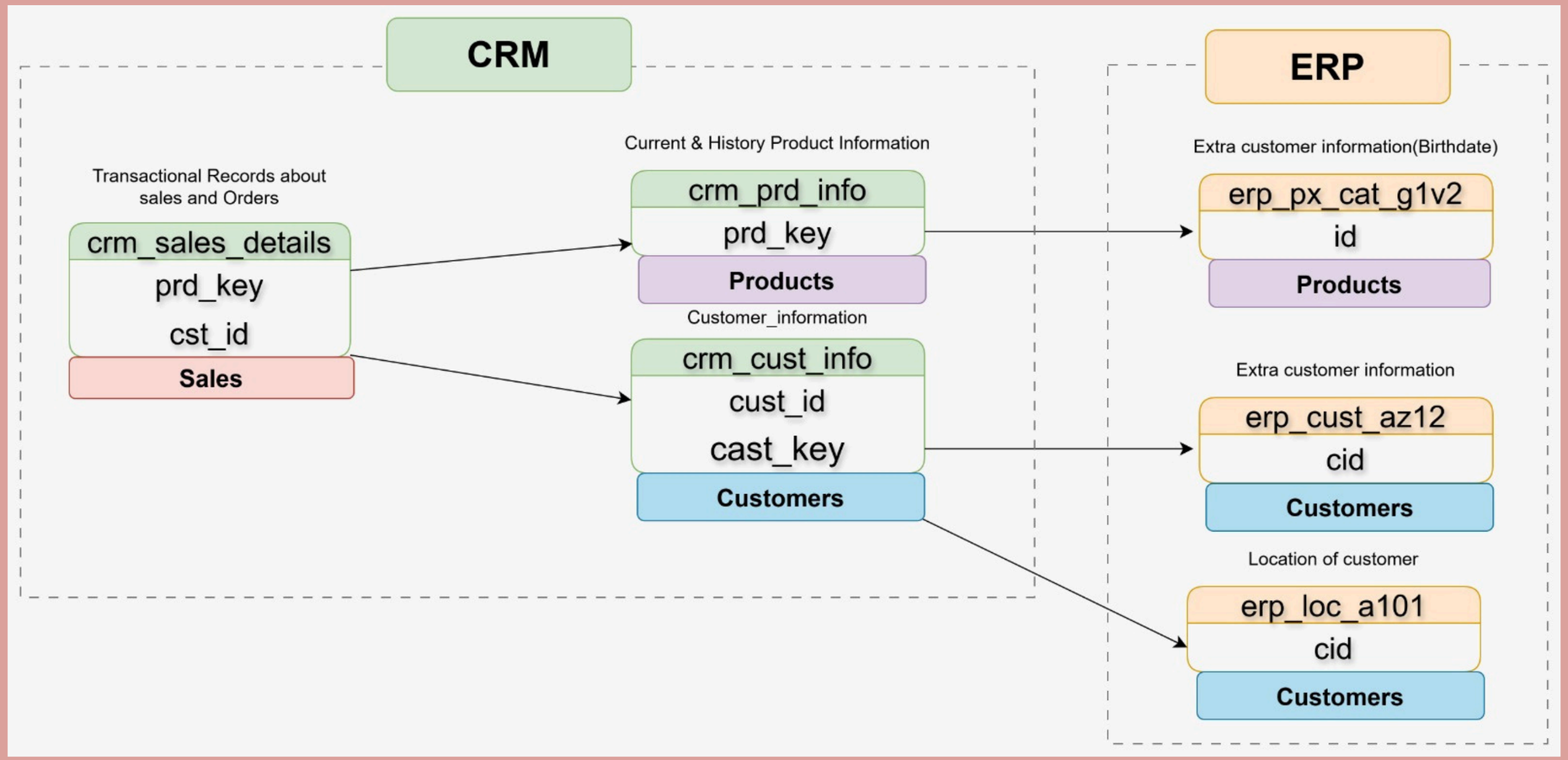
crm_prd_info stores product details, while crm_cust_info stores customer details.

**ERP System:**

Provides additional customer and product information, such as birthdates, locations, and extra product details through tables like erp_cust_az12, erp_loc_a101, and erp_px_cat_g1v2.

**Integration:**

Data from both CRM and ERP systems are merged using matching keys (like prd_key, cust_id, and cid) to create a complete and enriched dataset for analysis.



# Data Modeling: Schemas and Structures

This diagram represents the Star Schema model, which organizes data for efficient reporting and analysis.

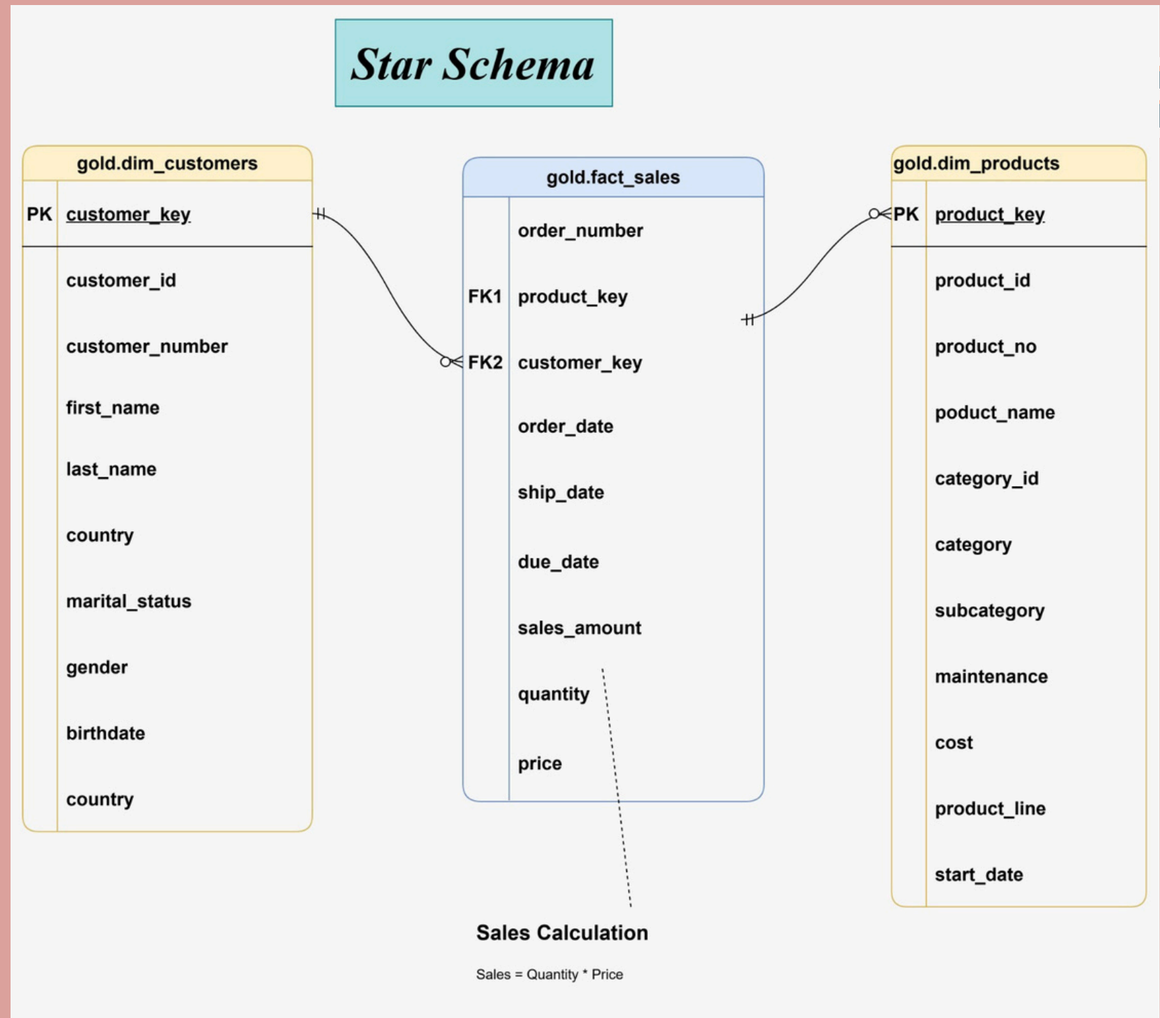
**Fact Table (fact_sales):**
Stores the transactional data like order details, dates, sales amount, quantity, and price.
It connects to dimension tables through foreign keys.
(Sales are calculated as Quantity × Price.)

**Dimension Tables:**

dim_customers: Contains detailed customer information such as name, gender, country, and birthdate.

dim_products: Holds product-related details like product name, category, cost, and start date.

The fact table links to both customer and product dimensions using keys, forming a star-like structure — hence called a Star Schema.



# Analytics: Tools and Techniques

## Data Visualization

Utilizing **charts and graphs** to present complex data sets visually, enabling stakeholders to quickly grasp patterns and insights, enhancing decision-making for business strategies.

## Statistical Analysis

Analyzing **trends and patterns** in historical data to forecast future outcomes, using statistical methods to derive actionable insights that inform strategic planning and operational efficiency.

## Machine Learning

Integrating **predictive analytics** tools to enhance data-driven decisions, leveraging algorithms to analyze data trends and improve forecasting accuracy, driving competitive advantage through informed insights.

# Conclusion

This project demonstrates a complete data warehousing process — from data extraction to final reporting.
Data from CRM and ERP systems is first integrated and stored in the Bronze Layer as raw data, then cleansed and standardized in the Silver Layer, and finally transformed into business-ready datasets in the Gold Layer.
The integration model ensures seamless merging of customer, product, and sales data, while the Star Schema organizes it efficiently for analytics.
Ultimately, this architecture enables accurate reporting, BI insights, and machine learning applications, supporting data-driven business decisions.

# Thank You

**Connecting Through Data Insights**



**Email**

vanshikakr2004@gmail.com

**Linkedin**

www.linkedin.com/in/vanshika-kumar-9268aa291

**Github**

vanshikakr2004@gmail.com