****

ETL & Data Warehouse Project

# 

Mira Ben Halima

Abir Lakhal

Pranav Prakash Nair

Mamadou Ousmane Keita

Om Prakash

The Dataset: [Company Financials Dataset](https://www.kaggle.com/code/rajatraj0502/company-financials-dataset/input)

**Table Of Contents**

[**Introduction 3**](#_jrtm0mp72t2m)

[**1. Selected Dataset and Rationale 3**](#_rwti6ry5kjx)

[**2. Star Schema Design 4**](#_qc32krd87sdd)

[**3. ETL Process and Key Transformations 5**](#_o7wtjovvu4wt)

[a. Extraction: 5](#_1yk7rz4twk6m)

[b. Transformation: 5](#_numpc6gmfgji)

[c. Loading: 5](#_jg89x89iwmre)

[**4. Insights and Conclusions 5**](#_25tg7g379oed)

[**Conclusion 6**](#_81pzww83xd8)

### Introduction

This project aims to demonstrate the creation of a data warehouse system using a real-world dataset containing international sales transactions. By transforming the original flat file into a structured star schema, we enable efficient querying, business intelligence analysis, and insightful reporting. The process involves designing appropriate dimension and fact tables, implementing a robust ETL (Extract, Transform, Load) pipeline using SQL Server Integration Services (SSIS), and analyzing the data to uncover key trends and metrics.

### 1. Selected Dataset and Rationale

For this project, the selected dataset contains sales transactions across different markets. Each record captures the segment, country, product, discount band, units sold, pricing details, gross sales, discounts, net sales, cost of goods sold (COGS), and profit, along with date information.

**Rationale:**

* The dataset is comprehensive and structured, ideal for demonstrating star schema design.
* It includes multiple dimensions (Segment, Country, Product, Date) that are critical for multidimensional analysis.
* It provides key metrics (Units Sold, Sales, Profit) useful for business intelligence and reporting.
* The dataset's variety allows for exploring relationships between different business aspects.

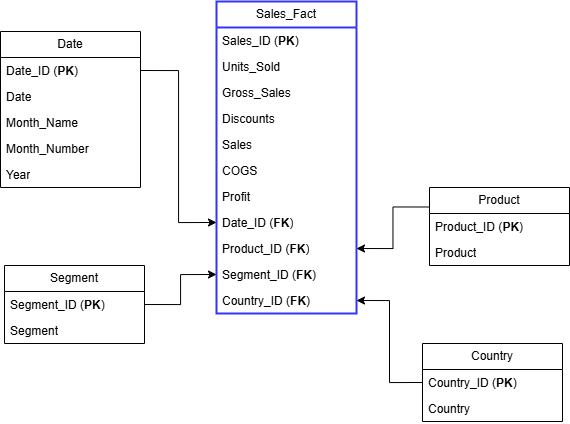
**Note:** The original dataset did not contain a unique identifier for each row. To support the creation of the fact table and enable referencing, an ID column was added using a SQL query with an auto-increment feature in the staging database (before loading into the final data warehouse).

**Data Dictionary:**

| **Column Name** | **Description** |
| --- | --- |
| Sales\_ID | Unique identifier for each sales transaction |
| Segment | Market segment (e.g., Government, Midmarket) |
| Country | Country of sale |
| Product | Product sold |
| Discount Band | Discount applied |
| Units Sold | Quantity sold |
| Manufacturing Price | Cost to manufacture a single unit |
| Sale Price | Selling price per unit |
| Gross Sales | Total sales before discount |
| Discounts | Amount discounted |
| Sales | Final sales value after discount |
| COGS | Cost of Goods Sold |
| Profit | Profit from the transaction |
| Date | Date of transaction |
| Month Number | Numerical month |
| Month Name | Name of the month |
| Year | Year of the transaction |

### 2. Star Schema Design

The star schema consists of one fact table and four dimension tables:



**Fact Table: Fact\_Sales**

* Sales\_ID (PK)
* Segment\_ID (FK)
* Country\_ID (FK)
* Product\_ID (FK)
* Date\_ID (FK)
* Units\_Sold
* Manufacturing\_Price
* Sale\_Price
* Gross\_Sales
* Discounts
* Sales
* COGS
* Profit

**Dimension Tables:**

* Dim\_Segment (Segment\_ID, Segment)
* Dim\_Country (Country\_ID, Country)
* Dim\_Product (Product\_ID, Product)
* Dim\_Date (Date\_ID, Date,Month\_Number, Month\_Name, Year)

### 3. ETL Process and Key Transformations

#### Data Extraction:

* **Source:** Data was extracted from a CSV source containing sales records.
* **Tools:** SSIS (SQL Server Integration Services) was used for this project. SSIS is a component of Microsoft SQL Server that can be used to perform a broad range of data migration and ETL tasks.
* **Process:**SSIS was used to connect to the CSV file.The data was read into the SSIS data flow.Data types were initially defined, noting that some columns required later conversion.

#### Data Transformation:

* **Surrogate Keys:** Textual values (Segment, Country, Product, Date) were transformed into surrogate keys using Lookup transformations. This improves query performance and reduces storage.
* **Data Cleaning:** Numeric columns were cleaned by removing extra symbols like "$" and ",". Values with "S-" in the Discounts column were handled by assuming they represent zero and replacing them with "0". This assumption was made due to the lack of context in the original data
* **Null/Empty Value Handling:** Null or empty values were handled with default replacements or removed depending on relevance to ensure data integrity.

#### Data Loading:

**SSIS Loading:**

* OLE DB Destination was used to connect to the target data warehouse database.
* "Table or View - Fast load" was employed for efficient data insertion.
* Error handling was implemented within the SSIS packages to manage potential issues during the loading process

### 4. Data Insights and Conclusions

1. SQL Queries:

* **What factors influence key performance metrics?**

This query analyzes how segment, country, and product impact sales and profit.

SELECT

ds.Segment,

dc.Country,

dp.Product,

SUM(fs.Sales) AS TotalSales,

SUM(fs.Profit) AS TotalProfit

FROM Fact\_Sales fs

JOIN Dim\_Segment ds ON fs.Segment\_ID = ds.Segment\_ID

JOIN Dim\_Country dc ON fs.Country\_ID = dc.Country\_ID

JOIN Dim\_Product dp ON fs.Product\_ID = dp.Product\_ID

GROUP BY ds.Segment, dc.Country, dp.Product

ORDER BY TotalSales DESC, TotalProfit DESC;

* **How do different attributes impact business outcomes?**

**How discount bands affect sales?**

SELECT

db."Discount Band",

SUM(fs.Sales) AS TotalSales,

AVG(fs.Profit / fs.Sales) AS AverageProfitMargin

FROM Fact\_Sales fs

JOIN Dim\_Segment ds ON fs.Segment\_ID = ds.Segment\_ID

JOIN Dim\_Country dc ON fs.Country\_ID = dc.Country\_ID

JOIN Dim\_Product dp ON fs.Product\_ID = dp.Product\_ID

JOIN Dim\_Date dd ON fs.Date\_ID = dd.Date\_ID

CROSS JOIN (SELECT DISTINCT "Discount Band" FROM convertcsv) db ON 1=1

WHERE fs.Sales IS NOT NULL

GROUP BY db."Discount Band"

ORDER BY TotalSales DESC;

**How sales vary by month?**

SELECT

dd.Year,

dd.Month\_Name,

SUM(fs.Sales) AS MonthlySales

FROM Fact\_Sales fs

JOIN Dim\_Date dd ON fs.Date\_ID = dd.Date\_ID

GROUP BY dd.Year, dd.Month\_Name

ORDER BY dd.Year, STR\_TO\_DATE(dd.Month\_Name, '%M'); -- Order by actual month order

1. Insights:

After analyzing the data warehouse, the following insights were observed:

* The Government and Midmarket segments have the highest sales volume, indicating they are key customer segments.
* Specific countries (Canada, Germany) and products ( VTT) contribute more significantly to overall profit, suggesting areas of strength.
* Seasonal trends are identifiable, with sales peaking during certain months (e.g., October, November), which could be linked to specific promotions or market cycles.
* Discount bands correlate directly with profit margins; higher discount bands may lead to increased sales volume but potentially lower profit margins.

### Conclusion

This project demonstrates how to transform raw transactional data into a structured, query-optimized star schema for decision-making. The ETL process enables clean, consistent, and enriched data that supports business intelligence and analytics.