Data Warehouse Report

Pierpaolo Spadafora - 263722

July 18, 2025

1 Introduction

I chose the **Online Sales Dataset**[1] which is a sample e-commerce database on Kaggle. To support richer analytics the data warehouse includes economic indicators such as *Gross Domestic Product* (GDP) [2] and *GDP per capita*[3] provided from the World Bank.
The aim is to provide multidimensional analytical perspectives relating sales performance to geographical context and economic trends, at a good level of quality, consistency and detail.

2 Dataset Analysis and Preparation

2.1 Primary Dataset Overview

The Online Sales Dataset consists of **49,782 records** across **17 columns**, covering the period from *January 2020* to *September 2025*. Key attributes include:

- **Transaction details**: `InvoiceNo`, `InvoiceDate`, `CustomerID`, `Discount`,
- **Product information**: `StockCode`, `Description`, `Category`, `UnitPrice`, `Quantity`,
- **Geographic and operational data**: `Country`, `SalesChannel`, `ShipmentProvider`, `OrderPriority`, `ReturnStatus`
- **Financial and logistical metadata**: `ShippingCost`, `PaymentMethod`, `WarehouseLocation`

2.2 Economic Indicators Integration

The World Bank economic datasets contains GDP and GDP per capita indicators in current US dollars for 266 countries from 1960 to 2024. The original dataset is structured as:
`"Country Name","Country Code","Indicator Name","Indicator Code","1960","1961",...,"2024"`

[1] `https://www.kaggle.com/datasets/yusufdelikkaya/online-sales-dataset`
[2] `https://data.worldbank.org/indicator/NY.GDP.MKTP.KD`
[3] `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`

2.3 Data Quality Assessment

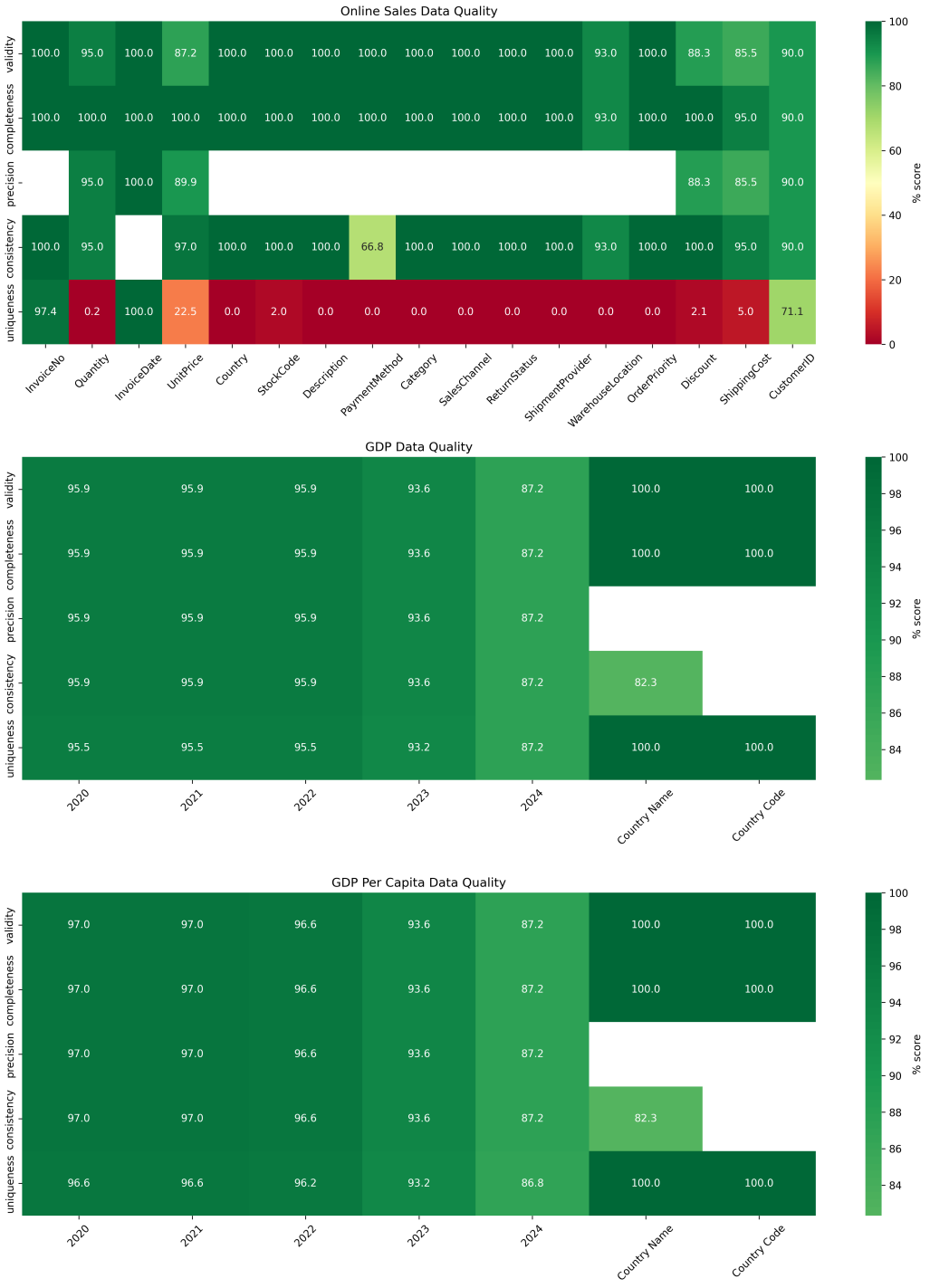

Figure 1: Initial data quality assessment heatmap for the Online Sales Dataset.

Online Sales Dataset Quality Issues:

- **Validity violations**: Fields such as `Quantity` and `UnitPrice` contained invalid negative values; `PaymentMethod` showed typographical errors; `ShipmentProvider`, `Discount`, and `ShippingCost` suffered from inconsistent formats.
- **Completeness issues**: Missing values were present in various fields.

- **Precision anomalies**: Numerical fields exhibited incorrect decimal formatting and inconsistent precision.
- **Consistency Problems**: The allegedly synthetic nature of the dataset resulted in random and inconsistent values.

World Bank Economic Data Quality Issues: The GDP datasets exhibited data gaps primarily in small nation-states and recent years where official statistics had not yet been published.

3 Data Integration with World Bank Economic Indicators

3.1 World Bank Economic Data Processing

The `2_Pulizia_CSV_GDP.py` script was used to rescale World Bank GDP datasets in a consistent manner. The time period was filtered for values between 2020-2024, matching the range of the Online Sales Dataset.
Data quality improvement handling missing values in one of these two ways:

- Linear interpolation was applied where there was minimal missing data to impute missing values
- Those with over two years of consecutive missing data were removed in an attempt to maintain continuity

3.2 Data Cleaning and Preparation Process

The raw datasets required extensive cleaning before integration.

1. **Online Sales Dataset Cleaning**
   Regular expressions were used to filter the correct values. Negative values in numerical fields were deleted or changed in absolute value depending on the type of field. The `Discount` column was normalized between 0 and 1.
   Text standardization corrected common typos and ensured consistent formatting for InvoiceNo and `StockCode` values. A new column `EstimatedUnitCost` was added, calculated as a random variation of the `UnitPrice` to simulate a plausible price of the items and calculate more meaningful KPI's with ease. A significant enhancement was creating realistic multi-item orders. The original dataset contained mostly single-item transactions, but with repeating InvoiceNumbers due to its allegedly random-generation nature, so related items were grouped into single multi-article orders
2. **Economic Data Integration**
   Integrating the World Bank economic indicators required a careful alignment between the two datasets. The destination countries of the customers in the Online Sales Dataset (OSD) were matched with the countries in the GDP CSV file, using their respective ISO alpha-3 codes to simplify representation with a standardized one. Wherever economic data were missing they were handled as previously mentioned.

Final Dataset Quality The cleaning process transformed the original 49,782 records into a refined dataset of 43,434 records significantly improving data quality.

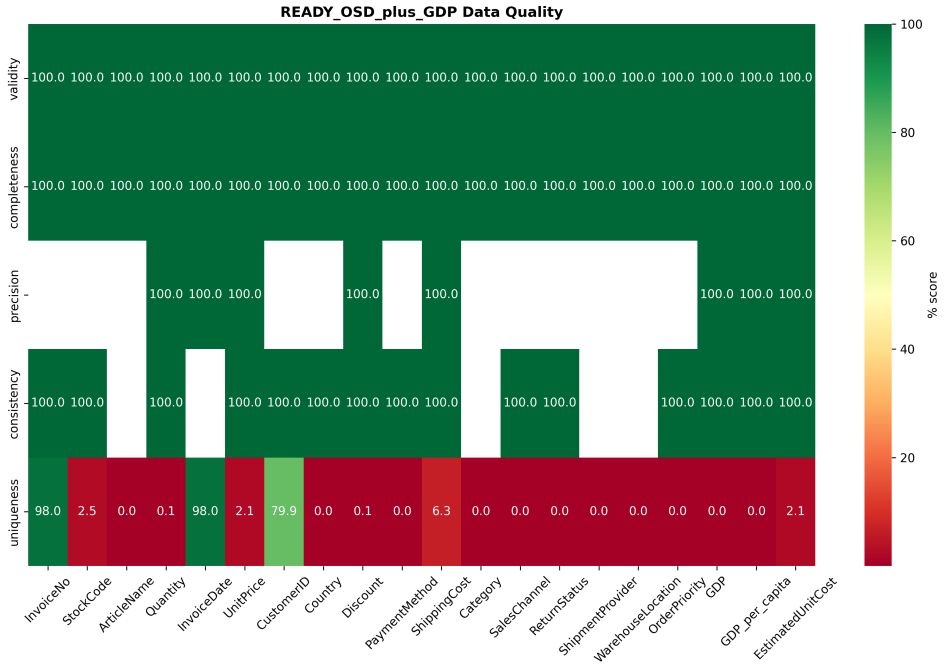

Figure 2: Final data quality assessment heatmap for the cleaned Online Sales Dataset + GDP.

4 Conceptual and Logical Schema

4.1 Initial E/R Schema

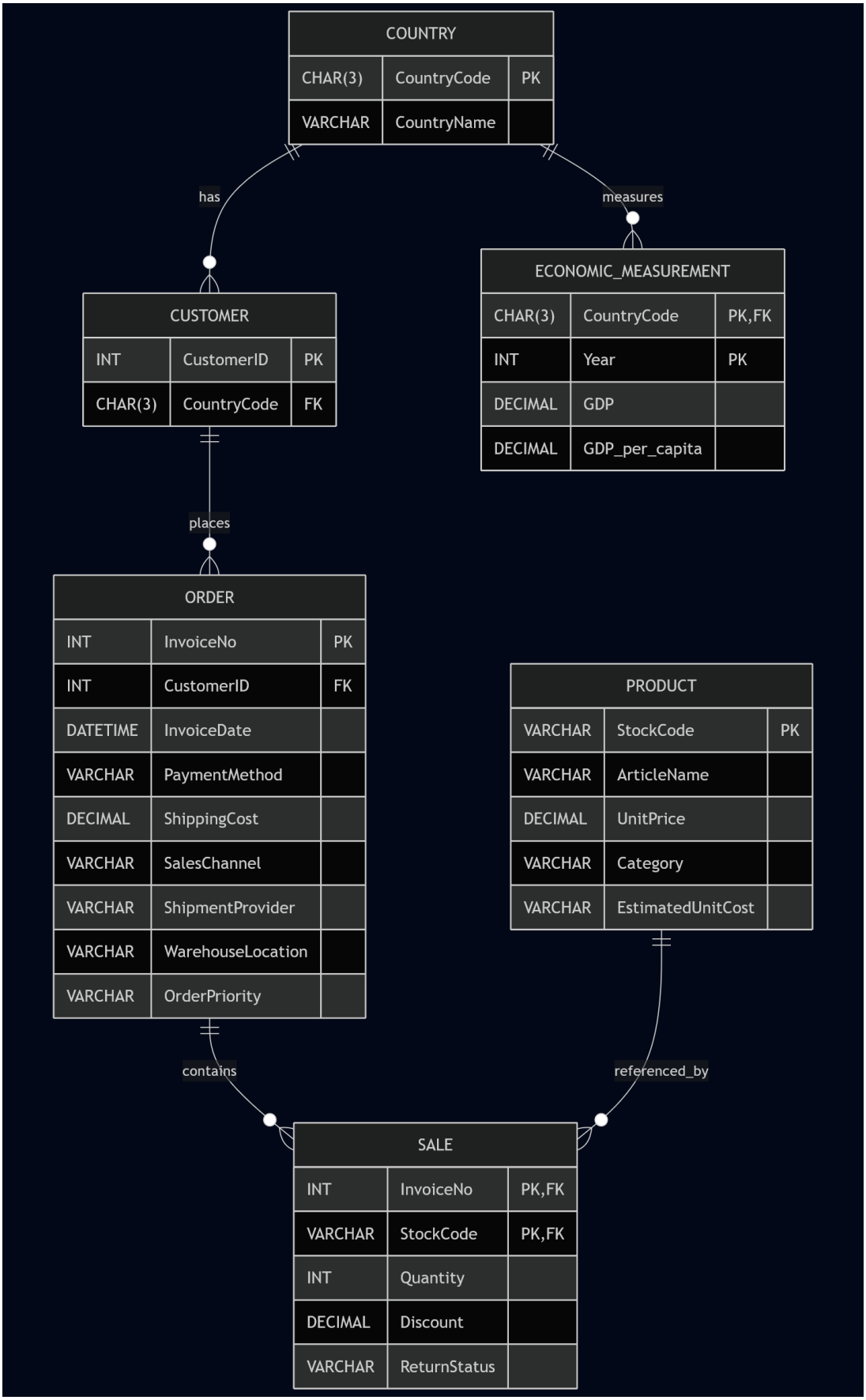

Figure 3: Initial Entity-Relationship diagram of the *Online Sales Dataset*.

The central entities in the first E/R model are: **Sale**, **Order**, **Product**, **Customer**, **Country**, and **Economic_Measurement**. `Order` is central in the schema: each Order is associated with one or more `Sale` entries, each linked to a `Product`. The Geo-Economical situation is addressed by the `Customer-Country` and the `Economic_Measurement` entity which provides GDP context for each country.

4.2 Attribute Tree

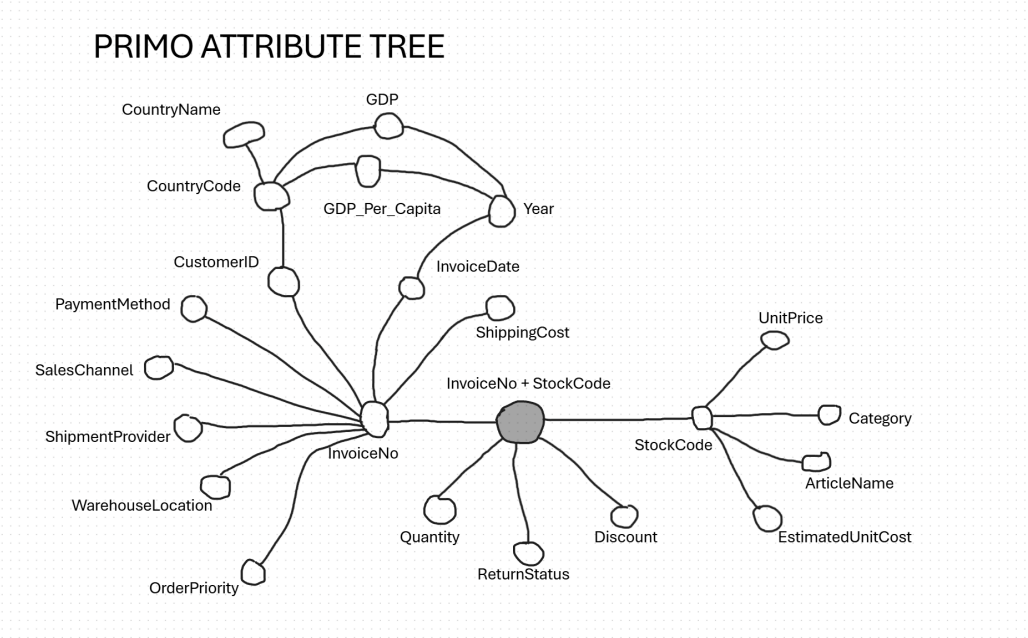

Figure 4: Attribute tree — full version.

Step 1 - initial attribute tree: all the raw dataset attributes were mapped giving a complete but rough idea of the data available.

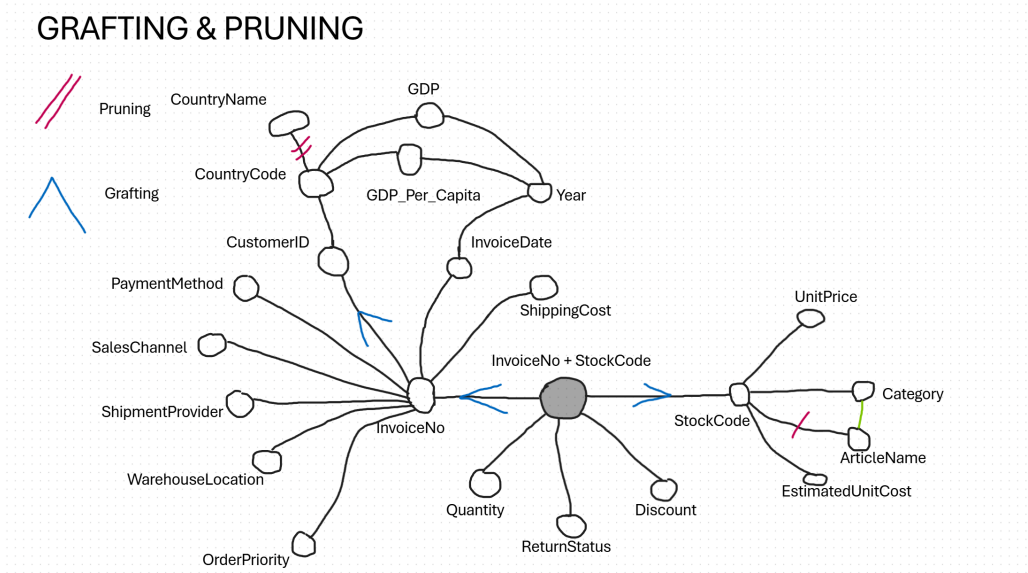

Figure 5: Pruning and grafting of the attribute tree.

Step 2 - pruning and grafting: Attributes such as `InvoiceNo`, `StockCode`, and `CustomerID`

were grafted into the tree root as they didn't add particular analytical value. The `CountryName` attribute was pruned as `CountryCode` is sufficient to uniquely identify countries. Additionally, `ArticleName` was repositioned as part of the hierarchy under `Category`, rather than being directly linked to `StockCode`.

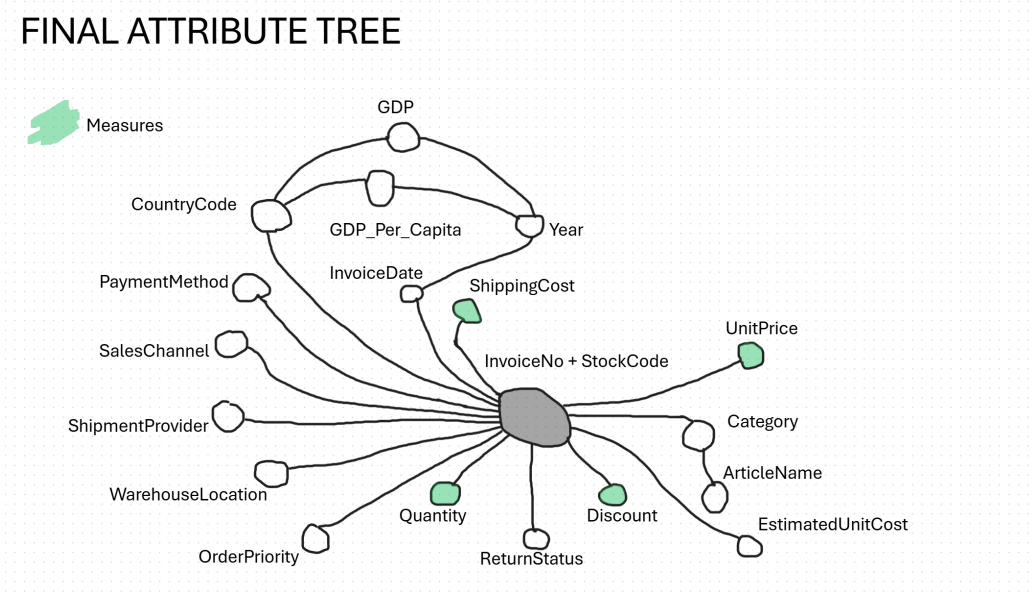

Figure 6: Final attribute tree.

Step 3 - final version: the final attribute tree contains only the attributes relevant for analytical purposes.

4.3 Dimensional Fact Model

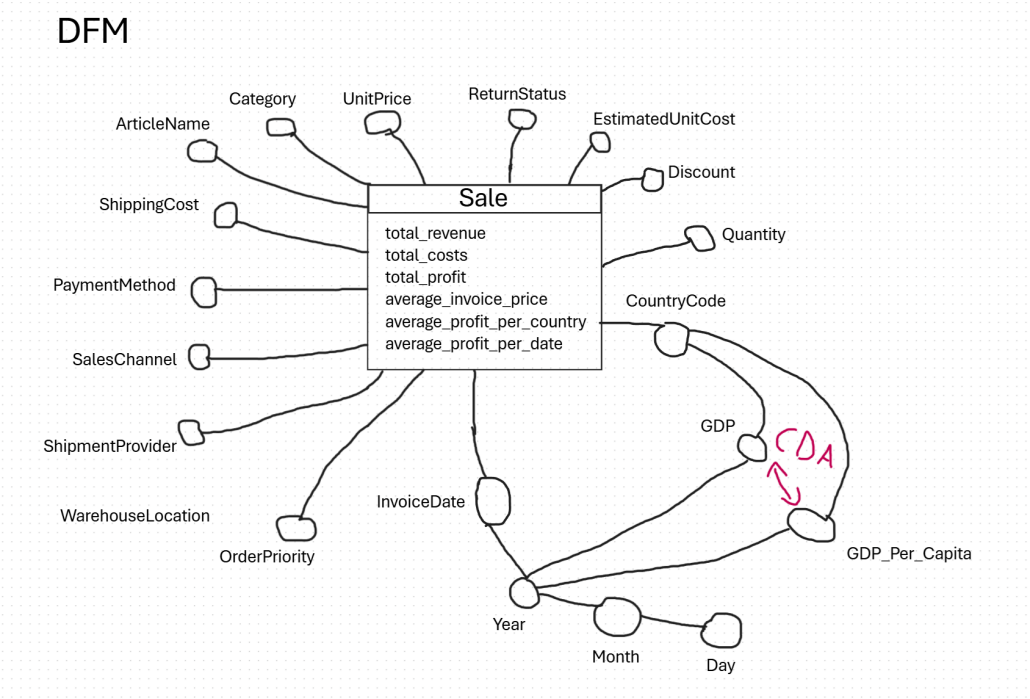

Figure 7: Dimensional Fact Model.

Given the final attribute tree, the dimensional fact model was created. The `Sales` table is the central fact table, containing measures such as `TotalRevenue`, `TotalCost`, and `TotalProfit` etc.

4.4 Star Schema

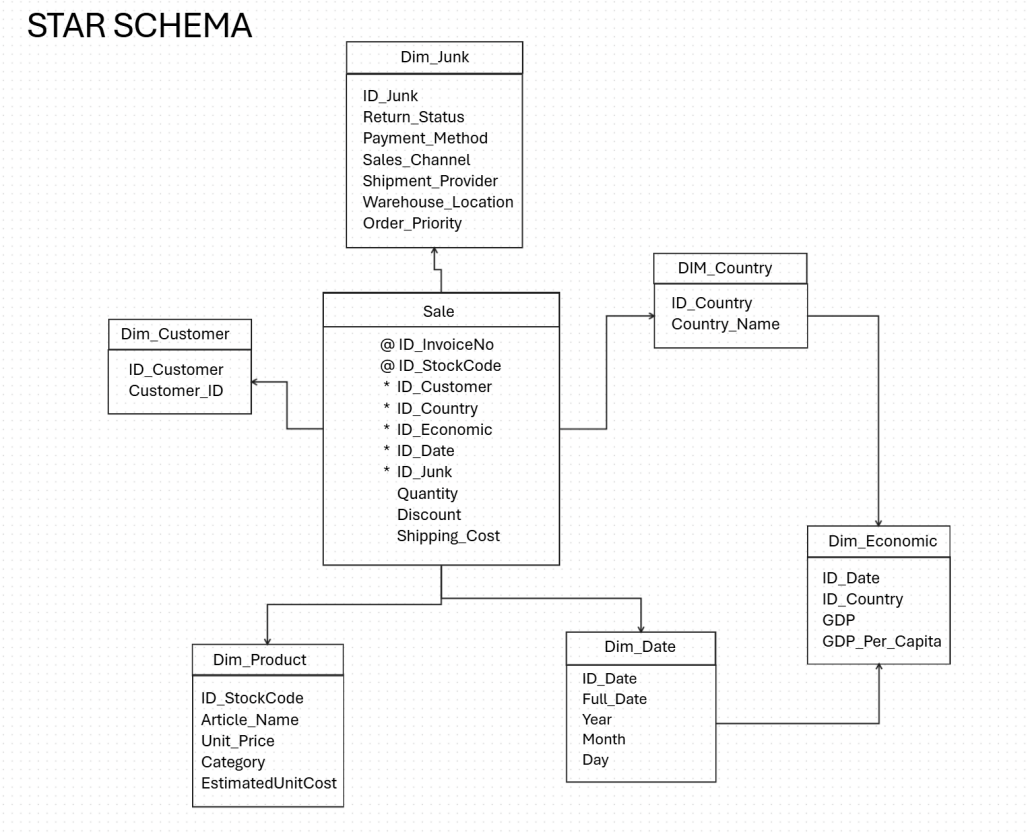

Figure 8: Final Star Schema of the Data Warehouse.

To avoid having many degenerate dimension we created a Junk one, `Dim_Junk`, to group attributes such as (e.g., `Return_Status`, `Sales_Channel`, `Payment_Method`) into a compact structure. The `Dim_Economic` dimension, linked via `ID_Country` and `ID_Date`, enables geo-economic analyses by correlating KPIs with GDP and GDP per capita over time.

References

[1] `https://www.kaggle.com/datasets/yusufdelikkaya/online-sales-dataset`

[2] `https://data.worldbank.org/indicator/NY.GDP.MKTP.KD`

[3] `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`

[4] `https://github.com/PierpaoloSpadafora/Data_Warehouse`