DEPARTMENT OF ECONOMICS AND BUSINESS

DEPARTMENT OF ELECTRICAL ELECTRONIC AND COMPUTER ENGINEERING

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

DATA BASE AND BIG DATA ANALYTICS

DATA WAREHOUSE & STREAMLIT DASHBOARD

Final Report – Homework 3

Author
Francesca Calcagno

Supervisor:
Giovanni Morana

ACADEMIC YEAR 2024/2025

TABLE OF CONTENTS

INTRODUCTION

The aim of this project was to simulate a real-world Business Intelligence process, from raw transactional data to the development of a data warehouse and the implementation of an interactive dashboard using Streamlit. The work was carried out in a consistent and methodical way, ensuring the accuracy of data transformations, the coherence of the star schema, and the analytical relevance of the results.

---

1. DATA LOADING AND CLEANING

The project began with the importation of three datasets:

- account_df: containing all transactions executed in 2024;
- symbols_df: including metadata about each traded asset (symbol, sector, industry, country);
- country_df: listing countries along with geographic and regional metadata.

The first step was to clean the data, which involved:

- Removing empty or unnecessary columns;
- Dropping rows with missing values;
- Standardizing formatting by removing trailing/leading white spaces;
- Converting all symbol names to uppercase to ensure consistency during joins;
- Parsing dates in European format using dayfirst=True.

These steps were essential to guarantee the success of the merges and to build reliable dimension tables.

---

2. CONSTRUCTION OF THE STAR SCHEMA

Once the datasets were cleaned, I proceeded to build a star schema. This structure was composed of:

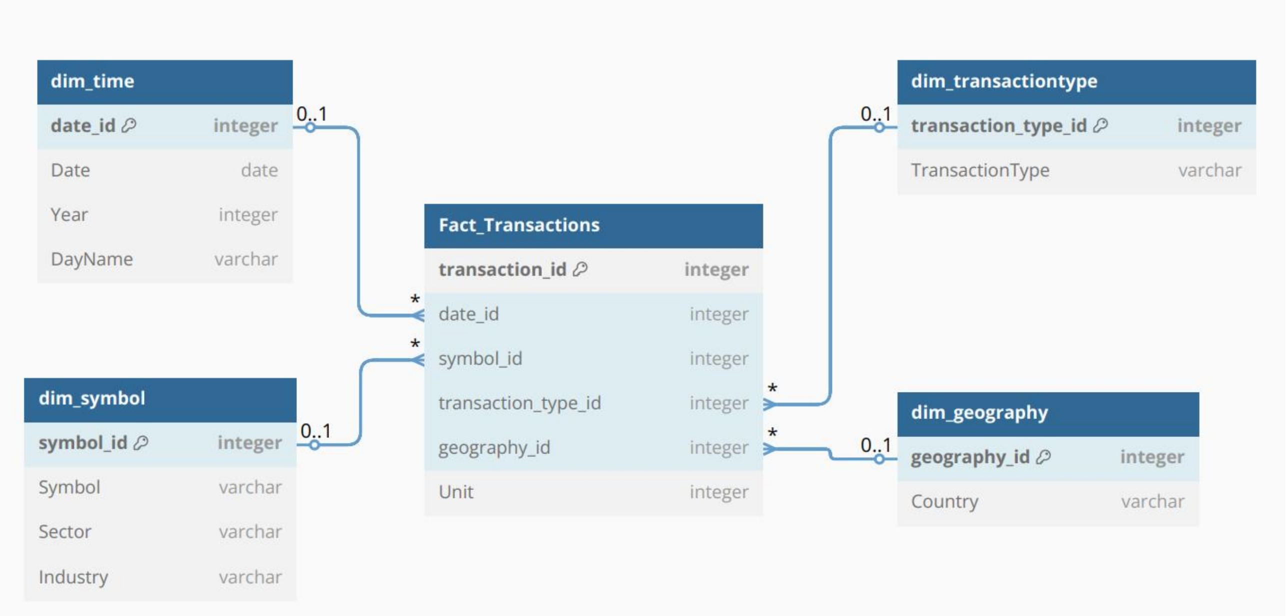
- A central **fact table** (Fact_Transactions) capturing all transactional events and linked to dimensions via surrogate keys;
- Four dimension tables:

- dim_time: including the date, year, and day of the week of each transaction;
- dim_symbol: containing information about each traded symbol, its sector, and industry;
- dim_transactiontype: listing the possible types of transactions (BUY and SELL);
- dim_geography: listing countries involved in the transactions.

Each dimension was carefully constructed by extracting distinct values, cleaning the strings, and assigning surrogate keys starting from 1. Only essential attributes were retained, ensuring the model remained both expressive and efficient.

The fact table was then built by merging the cleaned transaction dataset (account_df) with the dimensions, linking each transaction to its associated symbol, transaction type, time, and country.

This resulted in a **coherent and consistent star schema**, validated through direct inspection and test queries.



3. ANALYTICAL QUERIES AND VISUALIZATIONS

Once the star schema was complete, I proceeded to answer three analytical questions based entirely on the fact table and its linked dimensions:

1. **What are the top 5 sectors by number of SELL transactions in China during 2024?**
2. **What are the top 3 sectors by number of units sold on Mondays across 2024?**
3. **What are the top 5 countries by number of units traded in Financial Services companies?**

For each question, I wrote a dedicated query using pandas and produced a matplotlib bar plot to visualize the results. The charts were enriched with descriptive titles, axis labels, and numerical annotations for clarity.

4. STREAMLIT DASHBOARD – PAGE 1: TIME ANALYSIS

To complete the project, I developed a first interactive dashboard page using **Streamlit**. The application allows users to:

- Filter transactions by a custom date range (defaulting to the full year 2024);

- View a line chart showing the evolution of the number of transactions over time;
- Explore three bar charts showing:
    - The top 3 symbols by number of transactions;
    - The top 5 sectors by number of transactions;
    - The top 5 industries by number of transactions.

The dashboard is powered by a pre-merged dataset (final_joined_dataset.csv), optimized for speed and interactivity. Streamlit allowed the project to move from a static analysis to a more interactive, visual, and accessible data exploration experience

CONCLUSION

This project provided a comprehensive exercise in data transformation, modeling, and visualization. The entire pipeline was implemented using Python and followed best practices in data engineering. The result is a compact yet scalable star schema, capable of supporting a wide range of analytical queries.

The work was carried out with a high level of **consistency and logical structure**, from the initial ETL process to the final Streamlit dashboard. All elements were tested for correctness, and unnecessary attributes were removed to maintain a clear and focused data model.

The result is a solid foundation for business intelligence and analytics.