

Coffee Bean Business

SALES ANALYSIS

AHMED KAHLAOUI

CHEDHLI BEN AMARA

BA/IT

Introduction

Our project focuses on developing a Business Intelligence (BI) dashboard to provide actionable insights for a coffee bean business analyzing its sales and store performance over its lifetime.

Goals

We aim to understand customers' buying habits regarding our product, and get a clear vision on where to implement our resources to achieve better numbers.

The dashboard will highlight key performance indicators (KPIs), including:

- Sales by region
- Profit/quantity ordered by coffee type
- Customers demographic distribution
- Yearly/Monthly sales trends
- Sales trends by region
- Product features sales



Process Summary

The process of creating this project starts with data gathering, ensuring data metrics are relevant to the business' KPIs. Used ETL method for data preparation. Designed the data warehousing schema (star schema) with a ROLAP process. For data analysis and visualization we used POWER BI for an interactive dynamic dashboard, allowing filtering and drilling down for better insights.



ETL Process

1-Extract:


Collecting data from various sources .

2-Transform:

Cleaning, standardizing, and reshaping the data to ensure consistency and usability.
(handling missing values, removing duplicates...)

3-Load:

Storing the transformed data into a target system.



EXTRACTING

```
import pandas as pd

orders=pd.read_csv("./Raw Data/Orders.csv", sep=';',usecols=range(5))
customers=pd.read_csv("./Raw Data/Customers.csv", sep=';')
product=pd.read_csv("./Raw Data/Product.csv", sep=';')
```

This python code extracts the data from the different sources in their native form (csv: comma separated values) using the pandas library and loads them into a pandas data frame, which is a tabular data structure for data manipulation and analysis.

TRANSFORMING

	order id	order date	customer id	product id	quantity	customer name	email	phone number	address line 1	city	...
0	13644-699	2022-06-03	46296-42617-OQ	R-D-1	4	Fernando Sulman	fsulman10@washington.edu	+1 (828) 464-2678	45 Village Terrace	Asheville	...
1	ABK-08091-531	2020-10-30	53864-36201-FG	L-L-1	3	Tess Benediktovich	tbenediktovichmv@ebay.com	+1 (505) 523-8113	1068 Sutherland Plaza	Albuquerque	...
2	ABO-29054-365	2019-01-19	00256-19905-YG	A-M-0.5	6	Stanislaus Valsler	No email	+353 (479) 865-9222	95 Southridge Alley	Castlebridge	...

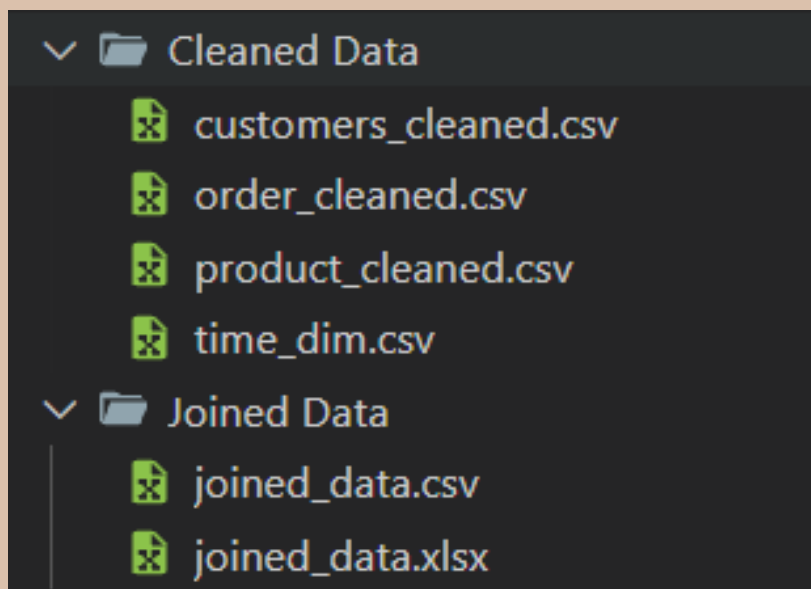
3 rows x 23 columns

Cleaned the data with python using different rules and functions to handle various problems such as missing values, duplicated data, inappropriate data types etc...

Joined the different data into a singular cleaned table containing all the needed information.

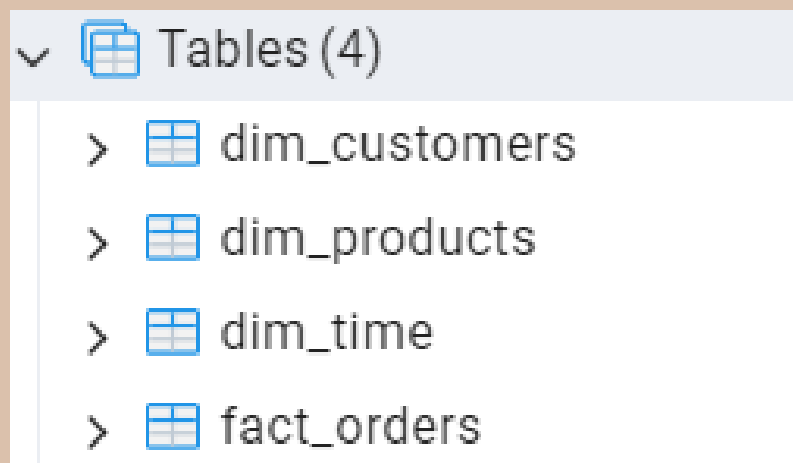
LOADING

The transformed data is saved into multiple formats for flexibility and further use. The processed data was exported as CSV and excel files. Additionally, the data was loaded into a PostgreSQL database for structured storage and advanced querying.



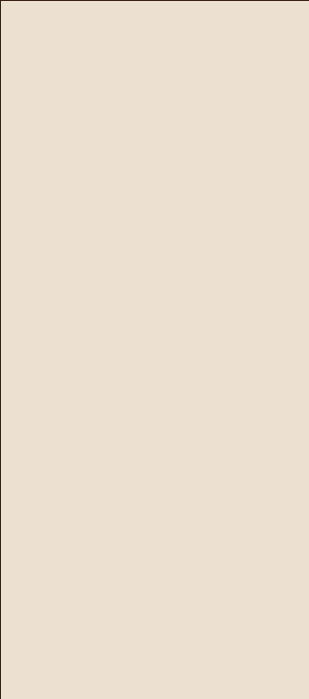
CSV AND EXCEL

POSTGRESQL DATABASE



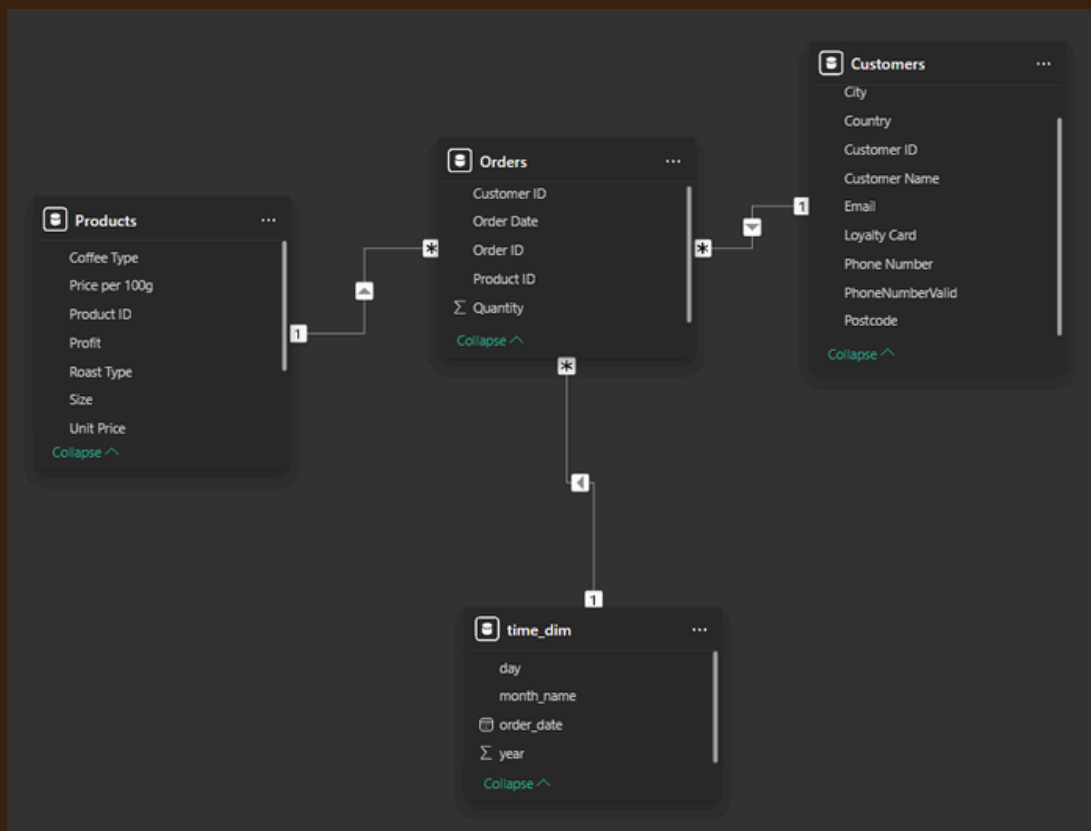


Data modeling and storage design

- IDENTIFY THE FACT AND DIMENSION TABLES
 - REPRESENT THE USED SCHEMA FOR MODELING
 - REPRESENT THE OLAP PROCESS
- 

1. Identification of Fact and Dimension Tables

- Fact Table:
 - Orders: the central fact table, containing transactional data such as Order ID, Product ID, Customer ID, Quantity, and Order Date. It captures the core metrics for analysis.
- Dimension Tables:
 - Products: contains details about the products, It provides context for the products sold.
 - Customers: stores customer-related information, such as Customer ID, Customer Name, Country.... It helps analyze customer behavior and demographics.
 - Time: captures time-related data, including order_date, day, month_name, and year. It enables time-based analysis, such as sales trends over time.



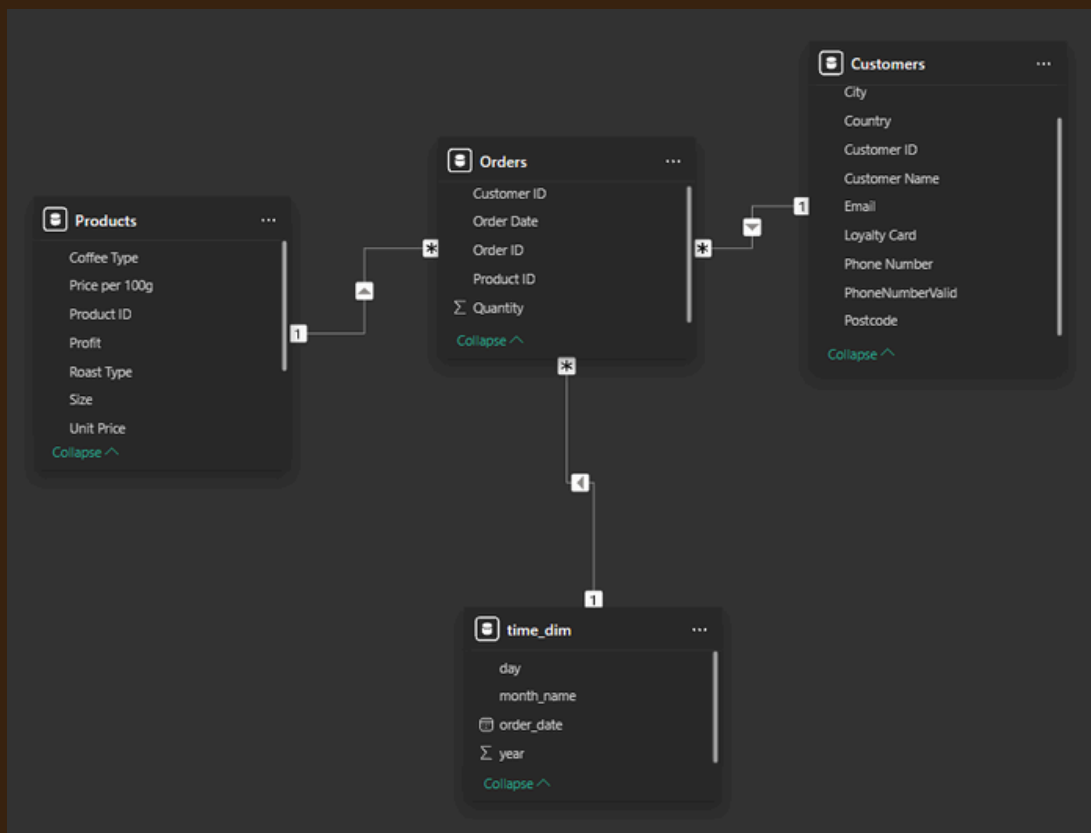
2. Schema Representation:

The schema was designed using a Star Schema.

In this schema:

- The Orders table acts as the central fact table.
- The Products, Customers, and time tables are connected to the fact table through foreign keys, forming the "star" structure.

The star schema was chosen for its simplicity, ease of querying, and performance benefits for analytical workloads.



3. OLAP Process

To support online analytical processing (OLAP), the ROLAP (Relational OLAP) approach was selected. ROLAP was chosen because:

- It leverages the existing relational database structure.
- It supports complex queries and large datasets efficiently.
- It is well-suited for the star schema design.

Data visualization



Data Visualization phase is critical to presenting insights in a clear manner. This involves designing and implementing a Business Intelligence displaying KPIs.

Aiming for simplicity, supporting decision making and interactivity we opted for using POWER BI. In our dashboard we used several visual components including:

- Bar chart to show sales distribution across different regions and coffee types
- Line chart to illustrate sales trends over time
- KPI and metric displays to showcase key numbers such as profit and quantity ordered
- Maps to indicate the demographic distribution of our customers

All in a unified layout with consistent color schemes, labels and legends for easier interpretation.

The implementation of this BI project has provided a solution for analyzing the coffee bean business's performance over their lifespan. By integrating data from multiple sources into a unified Data Warehouse and visualizing key performance indicators through an interactive dashboard, helped achieve better understanding of our business and its sales trend.

During the project, several challenges were encountered, such as ensuring consistency across all data sources, being unfamiliar with visualization tools and techniques. Which were addressed by data cleaning techniques and using star schema with a ROLAP process, learning and mastering data visualization tools (POWER BI) respectively.

The dashboard allowing us to visualize KPIs resulting in several recommended enhancements such as

- Pinpointing high-demand regions, in our case the US and focusing on them
- Focusing on ARA type as it generates the most profit and more demanded than the others
- Addressing weaknesses during the summer season as the quantity ordered is at its lowest during those months, while amplifying the strengths during the winter season as the quantity ordered is at its highest in general.
- There was no significant difference between quantity ordered by fidelity card customers and others so we should allocate less resources in that field.
- The process revealed that the dataset's suitability for effective CRM analytics was limited due to lack of variety, which limited segmentation.

(based on the POWER BI file attachment)

Conclusion

Difficulties and enhancements