NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Intelligence

Reporting – Third Delivery

Nicolas, Reyes number: 20230779

Peter, Falterbaum, number: 20230956

Luis, Penteado, number: 20230441

Pedro, Adonis, number: 20230994

Group 24

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

9 Jun, 2024

INDEX

1. Introduction 1
2. Business Understanding 2
   2.1. Business Problem 2
   2.2. Business Questions 2
3. Process discovery 2
4. Data modelling methodology 3
5. Dimensional Model 4
   5.1. Describe the creation process 4
   5.2. Table relationships and model schema 4
6. Main components and Hierarchies 5
7. Extract, Transform and Load 7
8. Describe data engineering 8
   8.1. Data Warehouse Creation 8
   8.2. Data Lakehouse Creation and Initial Data Preparation 8
   8.3. Dataflow for Preprocessing 8
   8.4. Data Quality and Transformation Issues 9
   8.5. Data Pipeline Configuration 9
   8.6. Data Transfer to Data Warehouse and outputs 10
9. Model optimization 10
   9.1. Configured Hierarchies 10
   9.2. Date Table Usage and Calculated Columns 10
10. Measures and Calculated Columns 11
   10.1. DAX Measures 11
   10.2. DAX Time Intelligence Measures 13
11. Dashboards 13
   11.1. Overview 13
   11.2. Pharmacy 14
   11.3. Location 14
   11.4. Product 15
   11.5. Sales 15
12. Analysis and results 16
   12.1. Data Analysis 16
   12.2. Answers to Business Questions 17

12.3. Answer to Business problem.....18
13. Conclusions.....18
14. References.....20

1. Introduction

Pharma4all is a pharmaceutical chain with several units spread around different cities, grouped in two distinct sub-companies. One of the next goals is to elaborate how to enhance the decision-making process through building a new BI system. With about three years of historical data, collecting data daily from their main systems and extracting them into csv files, they're strongly convinced that building a robust Business Intelligence System can be a game-changer to keep boosting the business performance and follow the market dynamic. The challenge faced is mostly focused on sales themselves, across different dimensions like timeframes, products, location and other KPIs derived by them like quantity of products and revenue. Having different business approaches, looking into these analyses aim to uncover growth opportunities, improve sales strategies, and gain deeper market knowledge. Regarding this brief description, the real obstruction lies in mining the available data efficiently and making the decision-making process more smoothly according to the business needs. A gap that can be easily filled by the implementation of a sophisticated BI system.

2. Business Understanding

2.1. Business Problem

Pharma4all has been facing informational challenges with their current extraction of csv files from its main servers. The manual process slows down the decision-making and increases the likelihood of errors.

This necessitates a solution that democratizes data access across organizational tiers. This solution should not only accelerate the analysis process for faster decision-making, but also implements a governance framework to secure sensitive information by taking out bottlenecks on the IT team to extract and provide the data. Furthermore, the Data team becomes responsible for adding a data quality layer and centralizing key performance indicators under uniform assumptions, which is essential to guarantee that all stakeholders have a consistent, accurate basis for their decisions.

They're also facing a complex business challenge in optimizing their sales strategy. The company aims to analyze sales data across various dimensions, including different pharmacies, suppliers, location, profiling sales based on product type, specifically branded versus unbranded products, all of them in terms of quantity and amount. Additionally, there's a need for comparative analysis between the current and previous fiscal years to identify growth patterns and potential areas for improvement.

2.2. Business Questions

This research stratifies the business problem into 5 complementary questions which can help to build strategies to solve it.

- What are the total sales per year?
- What are the total sales compared year after year?
- What is the top Manufacturer in terms of sales by year?
- Which team of operators sell more per year?
- What are the top 5 Brands in each district by month?
- What are the sales by location and which location has the highest and lowest sales monthly?
- What product (medication) has the highest sales?
- Which product sells more per year, the generic ones or not?
- What are the top districts with higher sales amount by year?
- What is the pharmacy who sell more by month?

3. Process discovery

The company's discovery process enables us to understand the main information about the operation and deeper knowledge of how their supply chain works. The datasets Location and Point of Supply

provide an overview of the geographical spread of pharmacies, identifying their sites and the critical supply points to the logistical network. In the retail aspect, the two datasets for Pharmacies A and B reveal the market coverage operational scope, not only in various locations.

Going further to a different dimension, the products catalogue offers a wide array of medications with manufacturer, brand, and drug types. This information is useful for improvements in the inventory management leading to better stock availability and hence better customer satisfaction.

The human element is captured in the Operators dataset, which delineates the workforce composition, roles, and organizational structure, essential for understanding the operational dynamics and employee engagement within the company.

Finally, the Sales (Pharmacy) table, centralizes transactional sales data, including details on timing, product IDs, pharmacy IDs, points of supply, locations, quantities sold, and financial figures. This table is the core around analytical queries and enables the correlation of sales data with the other dimensions.

4. Data modelling methodology

When designing an enterprise-wide data warehouse, two distinct methodologies can be applied. On one hand, the Inmon strategy focuses on a top-down approach, starting the design from a single, centralized enterprise-wide data warehouse that represents all data as a single source for the whole company. Data marts which are separating the concerns of users and applications can be derived in a following step to accompany the needs more efficiently and effectively. Especially the requirement to oversee and understand the whole data landscape in sources and needs, can lead to immense effort in implementing it.

On the other hand, the Kimball approach starts by separating several areas into data marts from the beginning on. These marts can be joined in a subsequent step to an enterprise-wide data warehouse but are not necessarily the initial focus. This approach leverages the benefit of defining data marts in alignment with the specific business requirements (subject-oriented), with the option to summarize everything into one wholesome organization.

This study uses the bottom-up strategy by Kimball to define a meaningful data warehouse, following four distinct steps to formalize the implementation done. Firstly, the business process needs to be identified and captured. For Pharma4All, this is the sales of the products across different locations of the company. Separated into two sub-companies, each sell transaction is registered and target of the presented analysis. Subsequently, the granularity – the highest level of detail in the dataset – needs to be considered. The granularity was set for each sale by time on a daily level (with more abstract hierarchies), by product, the pharmacy, and the location. In other words, each transaction can be uniquely differentiated with the listed attributes, which makes the sales id obsolete. Thirdly, several dimensions shall be employed to offer several perspectives on the data. Dimensions for products, operators, pharmacies (A and B), locations, point of supply and the date have been initialized. Finally, the measures and metrics need to define which can be used to quantify transactions. Sales records of

the described granularity have been considered as transactions, whereas quantity and amount are the quantifiable measures.

By following the Kimball method and its distinct steps, a supporting and effective data warehouse can be developed to enrich data-driven decisions in business.

5. Dimensional Model

5.1. Describe the creation process

To evaluate the situation of the company and business performance, it was needed to create a structure to read and understand the data that Pharma4All gives. To do that, some alterations were made to the initial tables for model improvements.

- We appended the tables Pharmacies A and B to create a new table named dim_pharmacy. It was also created a new column called pharmacy_a_or_b to differentiate which was the pharmacy.
- A new dimensional table was created and called dim_date that represents the date, year, and month information from the column Datetime that was taken out first on the sales table. So, we dropped the Datetime column from the fact table and created a new one called date_id, which is the number of years, month, day and hour all joined together to make a relationship between this new dimensional table with the fact one.
- For currency, as it is a dimension that has a one-to-many relationship with the sales table and can also give further information about different tables from the model, we created a table.

5.2. Table relationships and model schema

After that, we looked for each table and created relationships to establish how the information would go through the database.

In this context, the star schema was chosen for our model because of its effectiveness in terms of performance, given that the presented problem has initially a simple structure with only one fact table, and the simplicity making it easier to understand and to access for user writing queries. After that, we organized them into two different types: dimensional tables and a fact table. To connect them, one-to-many (1:n) relationships between the dimensions and the fact table were made with the surrogate key (“sk_”) in the former and a foreign key (“fk_”) was used in the latter. So, the following relationships have been established:

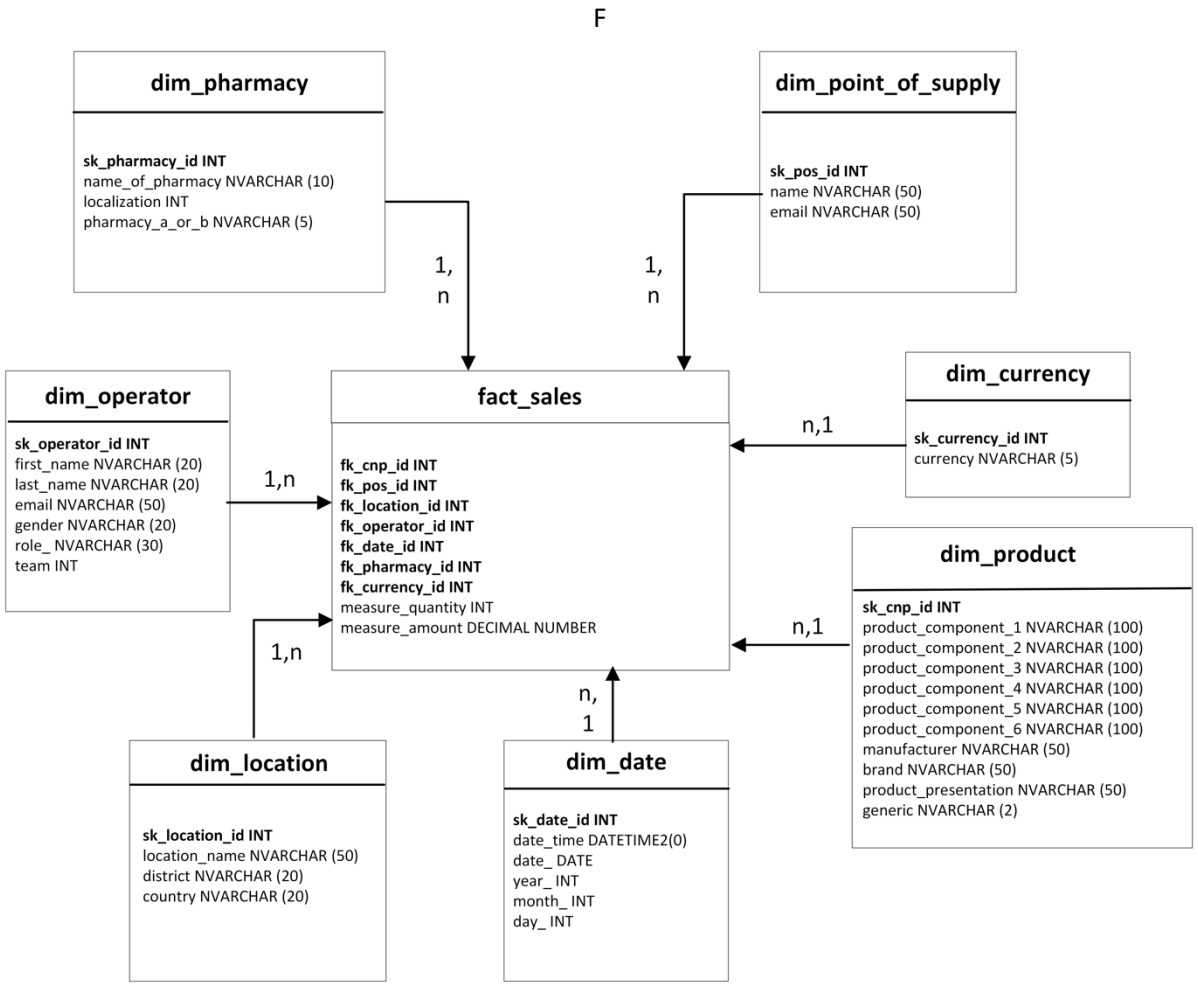

Figure 1: dimensional model showing tables from the Pharma4all business.

- dim_product and the fact_sales: sk_cnp_id ➞ fk_cnp_id
- dim_point_of_supply and the fact_sales: sk_pos_id ➞ pos_id
- dim_location and the fact_sales: sk_location_id ➞ fk_location_id
- dim_operator and the fact_sales: sk_operator_id ➞ fk_operator_id
- dim_date and the fact_sales: sk_date_id ➞ fk_date_id
- dim_pharmacy and the fact_sales: sk_pharmacy_id ➞ fk_pharmacy_id
- dim_currency and the fact_sales: sk_currency_id ➞ fk_currency_id

6. Main components and Hierarchies

Regarding the model created for the understanding of the data, and for bringing new solutions in how Pharma4All should democratize the information, it can be observed that in the structure of the model, there are some tiers and layers of relevance. For example, the model works with tables and each of them is populated. In this case, it is considered primary keys to distribute the database.

It is important to know that a fact table is a central component of a data warehouse that stores quantitative information for analysis, surrounded by dimension tables, and contains measures and

foreign keys linking to dimension tables. And on the other hand, the dimension tables provide the context and background information for the measures in the fact table, containing descriptive attributes like product details or customer information, and are linked to fact tables through primary keys.

In the following lines, this is how can be expressed the order of the tables and their components. Also, the dimension of those components and their hierarchies, expressed in the following structure A>B>C.

Hierarchies:

Apart from that, dimensional hierarchies were created for future Drill-down/Drill-up analysis, organization and navigability of the dimensional model. For that, we created the following multilevel structure:

- Periods Hierarchy - Year > Month > Day

  This hierarchy allows analysis of data over time at different levels of granularity. It starts with the broadest time period, the Year, then narrows down to the Month, and further to the Day. This structure enables users to analyze data annually, monthly, or daily, facilitating time-based trends and patterns identification.

- Manufacturer Hierarchy - Manufacturer > Brand > Product Name

  The Manufacturer one is designed to analyze data related to products. It begins with the broadest category, the Manufacturer, then drills down to the Brand associated with that manufacturer, and finally to the specific ProductName. This enables an analysis that can start at a high level (looking at all products from a particular manufacturer) and drill down to a very granular level (analyzing a specific product).

- Location Hierarchy - Country > District > Location Name

  This last hierarchy is about geographical data. It starts with the broadest geographic entity, the Country, then narrows down to the district within that country, and finally to a specific Location Name (which is the city). It allows for geographic analysis at various levels, from broad (country-level) to specific (location-level).

7. Extract, Transform and Load

The next step after defining the dimensional model is to build the data warehouse itself. This process consists of extracting the files provided from the company, submitting them to a data cleansing process and loading them to the data warehouse. This process contains three main steps:

Extract: The extraction step involves pulling data from various source systems, which in the Pharma4All will be csv files. The key challenges here are to ensure the data is extracted efficiently without disrupting the source systems and that the data captured is relevant and timely for the analysis needs of the organization.

Transform: During transformation, the data undergoes several necessary processes to ensure it is accurate, consistent, and usable. The outcome should match the definitions created for the dimensional model. This includes cleansing activities like parsing to break data into standard formats, correcting errors and inconsistencies, matching to identify and merge duplicates, standardizing to unify different formats and scales, and consolidating to combine data from various sources into a coherent dataset.

Load: The final step is loading the transformed data into the data warehouse. This process can be done in two ways: incremental loading, where only new or changed data is added, saving time and resources, or full loading, where the entire dataset is loaded, which is useful for initial setups or major refreshments.

The chosen tool to support this process is Data Factory from Microsoft Fabric that offers a unified solution for data engineering, integration, and analytics. Following the software guidelines, the step-by-step process to build the data warehouse was implemented in the following order:

1. Creating a Workspace: Set up a Workspace, which provides tools to manage and analyze data effectively, encapsulating this project to its own environment.
2. Building a Lakehouse: A Lakehouse is a platform created inside the workspace that combines the flexibility and scalability of the storage benefits of a data lake and has the SQL-based analytical capabilities of a data warehouse.
3. Uploading CSV Files: Upload local csv files to the Lakehouse.
4. Creating Dataflows (Gen2), based on Power Query: the dataflows are created to ingest and transform the initial csv files. This is the step where a sequence of activities is defined to transform the data to follow the rules and requirements outlined in the cleansing and integration steps. They can be further incorporated into data pipelines.
5. Building Pipelines: Dataflows are combined to a logical sequence of steps commonly used to automate the whole ETL process, starting from the file ingestion, passing through the transformations in the dataflow and loading the final data architecture to the data warehouse. You can define either to run it manually or schedule the run.
6. Loading Data into the Data Warehouse: This tool is responsible for storing, organizing and managing a large volume of data. With strong SQL capabilities, the data can be easily accessed from here to be used for data preparation, analysis, and reporting.

8. Describe data engineering

Following the initial ETL cycle, in which the data is extracted, transformed and loaded into the data warehouse, our data engineering efforts focused on better handling of the data within our business intelligence framework. Here we discuss the specific activities that optimize the ETL process.

8.0. Data Warehouse Creation

The data warehouse creation process began with setting up the environment in Microsoft Fabric, where we executed SQL scripts to structure our warehouse. These scripts defined both dimension tables, such as dim_pharmacy and dim_product, and the fact table, fact_sales. With that in mind, dimension tables stored specific attributes like pharmacy id and Localization, facilitating the ETL process, while the fact table captures the foreign keys and measures. Apart from that, we also employed SQL data types like VARCHAR, DATE, INT, and DECIMAL to ensure data consistency through the process. With that, our data warehouse integrated efficiently with the outputs from our dataflows and pipelines, allowing correct data loading.

8.1. Data Lakehouse Creation and Initial Data Preparation

We started by creating a Lakehouse to host all initial CSV files collected from Pharma4All's operational systems. This central repository simplifies data management by allowing us to store raw data in a scalable and cost-effective manner. Each CSV file represents a source system dataset, including sales, products, locations, supply, and the operators working for the company.

8.2. Dataflow for Preprocessing

In our data engineering efforts, we have developed specific Dataflows in Microsoft Fabric to handle preprocessing tasks such as cleaning, validating, and preparing data for further processing. An example is adding column names to the tables for Products and Locations. This step ensures that data entering our pipelines is accurate, correctly formatted, and matches the datatypes defined during the creation of the Warehouse, thereby reducing errors and inefficiencies in the processes.

As part of these Dataflows, we have also focused on creating tables specifically designed to address the business problems and questions at hand. Integrating table creation within the Dataflows allows for better results and aligns with the previously made database design. The tables we've developed include:

- **Dataflow Pharmacy:** the process combines data from Pharmacies A and B to provide a consolidated view, which is crucial for comprehensive performance analysis across the network. So, it was needed to merge both tables. Also, in the merged table, a column with the type of pharmacy was created, so it was used add new conditional column with a separation greater than ID 7, would be pharmacy type B, so instead of that, would be A. The destination table in the warehouse is dim_pharmacy.

- **Dataflow Currency**: We created this table to handle financial metrics and currency conversions that the main fact table could not accommodate, including columns that are not measures or foreign keys.

- **dim_date**: This table supports time-based analysis, enabling us to track trends from hourly to annual levels, depending on the required granularity.

- **dim_product:** Is important, given that it is the principal object that the customer receives. It is stored and sold in various locations. A critical issue addressed is the components of each product. In the beginning it was considered just the product, but now with the sk_product_id, it can be considered more in detail by separating it into 6 components for each product. This was changed also with a split column by delimiter '+'.

These tables ensure that the data is not only clean and structured but also meets the primary needs of business operations, thereby improving the efficiency of our data processes.

8.3. Data Quality and Transformation Issues

To ensure the integrity and utility of the data in our Business Intelligence system developed, we faced and addressed data quality and transformation issues. With that in mind, the main inconsistencies found were handling incorrect values and data merging/missing values. This chapter details the methods employed to resolve these issues, ensuring that the final data warehouse is reliable and efficient to support correct analysis and informed decisions.

Handling incorrect values

1. We transformed the "S" from the Generic products to "Y".
2. We transformed the name columns that were in Portuguese to English. For example: The column "Nome da farmácia" became "Name of the pharmacy".
3. We filter the outliers from "sales amount" of the fact sales.

Handling missing values

1. To address missing values regarding whether the products were generic or not, we used brand names for identification. We checked these brands on the Drugs.com website. This approach helped us fill in the gaps for drug-related information.
2. Additionally, we identified the districts corresponding to various cities by manually searching online.
3. We handled the missing emails from the employees

These steps ensured that all necessary data was complete and correctly integrated, effectively targeting the main problems and creating a more robust system.

8.4. Data Pipeline Configuration

Our data pipeline configuration helped the efficiency of our data processing architecture. It begins with clearing out old data from both dimensional and fact tables to ensure the data warehouse is updated with only the necessary information. This step is relevant to maintain the cleanliness and relevance of our data.

The pipeline includes essential "Wait" steps like "Wait to Delete Data" and "Wait to Load Dims". These steps help to manage the sequence of operations, making sure that data integrity is preserved by preventing data from being loaded prematurely. Each dimension such as location, product, operator,

and date have its own dataflow that processes data by transforming and standardizing it, preparing it for analysis.

After processing, the data is loaded into the appropriate tables in our data warehouse. This approach helps ensure data is accurately organized, which is critical for effective analysis and the performance of our systems. This configuration enhances our ability to conduct an effective analysis.

8.5. Data Transfer to Data Warehouse and outputs

After preprocessing, the data is transferred to the data warehouse and loaded into structured tables designed to optimize the database in the same way our previous database design. This setup enhances quick query processing and enables analysis across various dimensions. By integrating processed data into this repository, we facilitate effective analysis. Also, this final output allows us to read the data in the Power BI report tool, which should be the next step of the project.

9. Model optimization

9.1. Configured Hierarchies

The hierarchies considered to structure the data frame were considered between two important tables. The principal ones were from the dimension tables Product, Location and Date. The main idea of having hierarchies is because it brings an opportunity to structure the data, the visualizations and the filtering levels. In the following points there will be possible to read the structure of each hierarchy, considering the symbol > as a path of the hierarchy from the general to the granular level.

- Product (Dim_Product), has the following structure starting with the Products Components, they can go from 1 to 6. PC1 >PC2 >PC3 >PC4 >PC5 >PC6. This is because not all the products could have 6 components, so it is a rare situation.
  On the other hand there is another hierarchy, regarding this table: Manufacturer > Brand, because multiples brands could be done from the same manufacturer.
- For the Location (Dim_Location) it was considered like this: Country>District>Locality. Which provides a good opportunity to filter the information for each Pharmacy.
- Date (Dim_Date), this was important to the comparison between years and months, and the suggested way was: Year>Quarter>Month>Day.

9.2. Date Table Usage and Calculated Columns

The date table was created to consolidate columns related to dates, such as year, quarter, and datetime, into a single dimension. This setup enabled us to create Time Intelligence measures and specify date-related information in a more granular way for better analysis. Below is the latest output from the date table in Power BI.

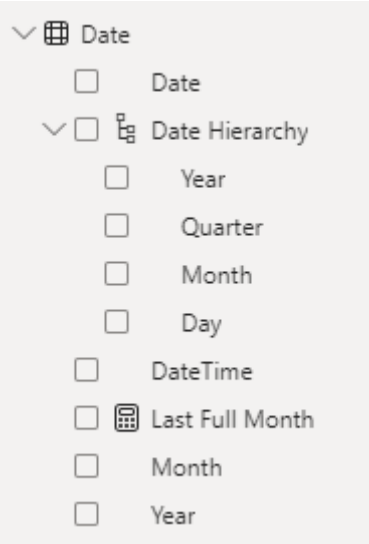

Figure 2: Date table showing the hierarchy

Apart from that, the calculated columns were created in the Microsoft Fabric Environment, inside the Dataflows. This approach eliminated the need to create them directly in Power BI, streamlining our data processing workflow. By doing this, we enhanced the efficiency of our data model and allowed us to leverage the Power BI for analysis and visualization.

10. Measures and Usage of DAX

10.1. DAX Measures

In this section, we detail the specific DAX measures created for our data model, which were created to better analyze and understand the sales performance of Pharma4All.

1) **Average Sales Amount per Month**
    a) **Description:** Computes the average sales amount for each month.
    b) **Purpose:** This measure was useful for identifying monthly sales trends and comparing them across different periods.
2) **Predicted Growth**
    a) **Description:** Calculates the predicted growth in sales based on historical data and trends.
    b) **Purpose:** This measure is used to make scenarios for forecasting future sales and planning accordingly.
3) **Quantity**
    a) **Description:** Sums up the total quantity of items sold.
4) **Ratio of Generic on Quantity Sales**
    a) **Description:** Calculates the ratio of generic product sales to total sales quantity.
    b) **Purpose:** Useful for analyzing the market share of generic products compared to branded ones.
5) **Ratio of Pharmacy A on Total Sales**
    a) **Description:** Computes the ratio of sales contributed by Pharmacy A to the total sales.

b) **Purpose:** This measure helps in understanding the contribution of Pharmacy A to the overall sales performance.

6) **Top Manufacturer by Sales**
   a) **Description:** Identifies the manufacturer with the highest sales.
   b) **Purpose:** Useful for supplier performance evaluation.
7) **Total Sales Amount**
   a) **Description:** Sums the total sales value amount.
   b) **Purpose:** Provides a view of total sales performance.
8) **Total Sales Quantity**
   a) **Description:** Sums the value of total sales quantity.
   b) **Purpose:** Used for understanding overall sales volume.
9) **YoY Percentual Difference**
   a) **Description:** Calculates the year-over-year percentage difference in sales.
   b) **Purpose:** Provides relevant information on the growth rate and trends over the years.
10) **Target Sales 2023**
   a) **Description:** Sets the sales target for the year 2023. Which is the growth of the Portuguese pharmaceutical industry from 2022 to 2023 multiplied by the sales in 2023.
   b) **Purpose: This measure evaluates performance against the set target for 2023.**
11) **Target Sales YTD**
   a) **Description:** Measures the year-to-date sales against the target.
   b) **Purpose:** Provides a better understanding into whether the sales are on track to meet the annual targets
12) **Total accumulated Sales by Month**
   a) **Description:** Sums the total sales amount accumulated monthly.
   b) **Purpose:** Useful for understanding the sales accumulation over the months.
13) **Total Sales 2023**
   a) **Description:** Sums the total sales amount for 2023.
   b) **Purpose:** Used for current year sales tracking.
14) **YTD CY**
   a) **Description:** it shows the total sales amount of Year-to-date sales for the current year.
   b) **Purpose:** Tracks the sales performance within the current year up to the present date.
15) **YTD CY Quantity**
   a) **Description:** Presents the total quantity number of Year-to-date quantity for the current year.
   b) **Purpose:** Tracks the sales volume within the current year up to the present date.
16) **YTD LY**
   a) **Description:** Shows the year-to-date sales for the last year.
   b) **Purpose:** Provides a way of comparing the current year's performance with the same period last year.
17) **YTD LY Quantity**
   a) **Description:** Year-to-date sales quantity for the last year.
   b) **Purpose:** Enabled us to compare the current year's sales volume with the same period last year.

10.2. DAX Time Intelligence Measures

It was also used Time intelligence measures for analyzing data over different periods. By doing that, it can compare and analyze trends over time. Below are the DAX time intelligence measures used in our project:

1) **Quarter over quarter total sales**
    a) **Description:** It calculates the total sales for the current quarter, comparing it to the previous quarter.
    b) **Purpose:** This measure helps in analyzing sales performance on a quarterly basis, identifying trends, and making informed decisions based on quarter-over-quarter changes.
2) **Rolling 12 months sales amount**
    a) **Description:** Sums the total sales amount over the past 12 months, being updated as time goes by.
    b) **Purpose:** This measure provides a view of sales performance over the last year, explaining better the trends over a longer time period.
3) **Average sales amount per pharmacy month over month**
    a) **Description:** It calculates the average sales amount for each pharmacy, comparing it through month.
    b) **Purpose:** This measure is useful for tracking the performance of individual pharmacies, identifying growth patterns month after month.

11. Dashboards

11.1. Overview

The dashboard provides an overview of Pharma4all's sales performance, featuring several key metrics designed to inform better decision-making. It begins with a "Year to Date" comparison of total sales against the previous year, offering information on the company's current trajectory and its position relative to historical benchmarks. This is complemented by the average sales amount per month, which helps in understanding the consistency and reliability of revenue streams.

Additionally, the dashboard includes ratios for pharmacies and generics in relation to total sales. These ratios offer valuable insights into the contribution of individual pharmacies to the overall sales and the popularity of generic products compared to branded ones, potentially influencing pricing strategies and inventory management.

Further, the accumulated yearly sales comparison and forecast sections provide a forward-looking perspective, allowing for proactive planning and budgeting based on anticipated sales trends. The top 5 total sales by district and manufacturer highlight areas of robust performance and potential opportunities for expansion or optimization.

Lastly, the total sales per pharmacy metric offers granular insights into the sales contributions of individual outlets, facilitating targeted support and resource allocation to high-performing pharmacies.

Together, these elements form a cohesive picture of the company's sales landscape, empowering leadership to make data-driven decisions aimed at maximizing profitability and operational efficiency.

11.2. Pharmacy

The pharmacy page of the dashboard focuses on performance metrics for individual pharmacies, offering insights that are crucial for optimizing operations and enhancing customer satisfaction. At the core of this page is the "Monthly Pharmacy Sales Comparison," which juxtaposes the sales performances of Pharmacy A and Pharmacy B month over month. This feature allows for direct comparison, highlighting trends, and identifying periods of increased or decreased sales activity, which can be pivotal in adjusting marketing strategies or addressing operational inefficiencies. Additionally, there are some tooltips in this graph to show the Pharmacy type, the total Sales amount and the year/month, just as an extra-help to understand the data.

Another key aspect of the dashboard is the "Staff Sales Performance" section, which evaluates the sales contributions of staff members over the last three months. This evaluation includes a performance indicator with rates, providing a clear assessment of whether staff members are performing better or worse than their historical averages. Such insights are invaluable for identifying high-performers who may benefit from additional training or recognition, as well as for addressing underperformers through targeted coaching or restructuring efforts.

Together, these components of the pharmacy page offer a nuanced view of each pharmacy's sales dynamics and staff productivity, equipping managers with actionable intelligence to drive improvements in sales outcomes and overall operational excellence.

11.3. Location

The Location page of the dashboard is meticulously designed to provide a geographical and temporal analysis of sales performance across various districts and brands, leveraging interactive visuals to convey complex data in an accessible manner. Central to this page is a map visualization that displays the total amount of sales across various locations, with the size of the bubbles indicating the volume of sales. This innovative representation not only aids in quickly identifying high-sales regions but also facilitates comparisons between distant or numerous locations.

Accompanying the map is a bar chart detailing the average amount per sale by district, offering insights into the transaction sizes prevalent in each area. This metric is crucial for understanding the purchasing behavior of customers in different locales and can influence pricing strategies or promotional activities tailored to local markets.

The dashboard also features a line chart that tracks the total sales amount by year, month, and brand, providing a comprehensive view of sales trends over time. This tool is particularly useful for identifying seasonal patterns, brand-specific sales cycles, and long-term growth trajectories, enabling strategic planning and forecasting.

Additionally, a bar chart presents the total sales amount by district, highlighting the relative performance of different geographic areas. This can help in allocating resources more effectively, targeting marketing efforts, and optimizing store layouts to maximize sales.

Moreover, a line chart of total sales amount by year, month, and point of supply offers a detailed breakdown of sales channels, revealing which methods of supply—such as online, in-store, or through third-party distributors—are most effective. This insight is vital for streamlining operations and focusing on the most profitable sales channels.

Finally, the page incorporates flexible filtering options, including relative time and date filters with both basic and advanced settings. These filters allow users to drill down into specific periods, compare sales across different time ranges, and analyze trends over various intervals, enhancing the dashboard's utility for ad-hoc analyses and strategic planning.

11.4. Product

The Product page of the dashboard is dedicated to analyzing sales performance at the product level, offering insights that are crucial for inventory management, product positioning, and strategic planning. At the heart of this page is a bar chart displaying the total sales quantity for each product name, providing a clear overview of which items are generating the most revenue. This visualization allows for easy identification of top-selling products and can guide decisions on stock levels, promotions, and new product introductions.

In addition to product-level analysis, the dashboard also breaks down sales quantities by manufacturer and generic status. This dual classification helps in understanding the contribution of different manufacturers to the overall sales mix and the impact of pricing strategies on sales volumes. Insights gained from this comparison can inform negotiations with suppliers, pricing policies, and the selection of products to feature prominently in stores or marketing campaigns.

One of the standout features of the Product page is a diagrammatic representation of the composition of each product, showcasing the amount of sales associated with each component. This unique visualization allows users to delve deeper into the sales data, identifying even minor contributors to a product's overall sales. For instance, looking at a sixth component might reveal a previously overlooked factor that significantly influences sales. This detailed breakdown supports more nuanced inventory management, allowing for adjustments to be made based on the actual sales composition of each product.

Overall, the Product page equips stakeholders with insights into product performance, enabling informed decisions that can optimize sales, improve inventory turnover, and enhance customer satisfaction by aligning product offerings with market demands.

11.5. Sales

The Sales page of the dashboard is designed to provide a comprehensive overview of sales performance across various dimensions, including time, supplier, brand, and product. This page is equipped with a range of time filters, allowing users to select specific dates or periods for analysis, making it possible to conduct historical comparisons, trend analyses, and forecasts.

At the center of the dashboard is a Key Performance Indicator (KPI) for the average quantity sold per supplier per year. This metric is relevant for assessing the efficiency and reliability of each supplier, guiding decisions on contract renewals, negotiation terms, and potential supplier diversification strategies.

Another KPI featured on the page is the Quantity Year-to-Date (YTD) comparison with the previous year. This metric provides immediate insights into the current sales performance relative to historical benchmarks, offering early warnings of potential shortfalls or surpluses in sales volumes.

The dashboard also includes a line chart of accumulated sales by year, offering a long-term perspective on sales trends. This visualization helps in identifying patterns, such as seasonal fluctuations or gradual growth, and in setting realistic sales targets.

A bar chart of rolling sales by year and month offers a dynamic view of sales performance, allowing for the detection of short-term changes and the anticipation of future trends. This can be particularly useful for adjusting inventory levels, launching promotional campaigns, or modifying pricing strategies in response to changing market conditions.

Additionally, the page features separate bar charts for the top brands and top products by total sales. These visualizations highlight the most successful offerings in the market, informing decisions on product placement, marketing investments, and new product development.

Lastly, a line chart depicting month-over-month sales per pharmacy by year and month provides insights into the performance of individual pharmacies. This metric is essential for identifying high-performing locations, addressing underperformance, and optimizing resource allocation across the network.

Overall, the Sales page of the dashboard serves as a powerful tool for monitoring and managing sales performance, supporting strategic decision-making and operational excellence in the pursuit of maximum profitability.

12.Analysis and results

12.1. Data Analysis

From the dashboards created, here is the most relevant information by main topic we were able to find.

Sales Performance and Trends:

- The Year-to-Date (YTD) sales for the current year (2023) have reached 2.05 million, a 17.94% increase compared to the goal, which was the sales from the first semester of 2022 multiplied by the increase of the Portuguese pharmacy sector from 2022 to 2023, found on the Statista website.
- The accumulated yearly sales show a steady increase from 2020 to 2023, with a forecast indicating continued growth into 2024.

Top Performers:

- Districts like Braga, Beja, and Aveiro lead in total sales. Additionally, Roche, Sanofi, and Novo Nordisk are the top manufacturers.

Pharmacy Analysis:

- By analyzing the plots, Pharmacy B represents most of the company's sales.
- It is noticeable that in the winter months (especially January) of each year there is a spike in sales, with January 2023 being the top month for both pharmacies in the entire time series.

Location Analysis:

- Braga and Beja are the top-performing districts in sales. On the other hand, Lisbon, which has a significantly larger population, is only in the 6th position, and Porto is not even among the top-performing districts.
- The allergy brands Clarityn-D and Zyrtec-D are the top sellers in all districts over the last year of data.

Product Analysis:

- Although sales are well balanced, Clonazepam and Loratadine represent the product components with the highest sales. Loratadine, a component for anti-allergic medication, aligns with the top-selling nature of allergy brands.
- The average amount by generic and brand shows that most product sales are generic, which makes sense given that the regions with the highest sales do not have a GDP per capita as high as Lisbon or Algarve, for example.

12.2. Answers to Business Questions

The process through Data Factory and Power BI can effectively address the business questions outlined by Pharma4all through its robust data visualization and reporting capabilities. Here is a brief explanation question by question:

- **Total Sales Per Year**: By the KPI Year-to-Date you can see the total of sales from this year. Considering the ratio of the difference between the expected and the real amount, it is a simple and effective way to show the performance updated. There is a total of 2.05M in sales, a +17.94% on last year.
- **Total Sales Comparing Year After Year:** A line chart was implemented to compare sales figures year-over-year, highlighting trends, and identifying periods of growth or decline. It maintains a consistency in quantity, but at the end of 2023, it decays.
- **Top Manufacturer in Terms of Sales by Year:** With a bar chart it is ranked manufacturers based on their sales performance annually, offering insights into market dominance and shifts in consumer preferences. The ones with more sales are Roche, Sanofi and Novo Nordisk.
- **Operators' Sales Performance:** It was created a table visual to compare sales performance across different operator teams, allowing for easy identification of top performers and areas needing improvement. In the last month, the best performance comes from pharmacy B, from Floor Manager and Director. Considering that this is a comparison MoM.

- **What product (medication) has the highest sales?** A bar chart shows the sales performance of medications monthly, helping in identifying top-selling products. The one with more sales is Ácido Acetilsalicílico.
- **Which product has more total sales, the generic ones or not?** This is compared using a stacked bar chart, providing insights into consumer preferences and pricing strategies. Also, with the card in the overview page, there is a rate of quantity sales between Generic or not Generic. In this case, the one which has more is the Generic with 55%.
- **What are the top districts with higher sales amount by year?** A geo-spatial bubble chart was used to visualize sales performance across different cities, with bubbles representing the magnitude of sales, enabling the identification of top-performing cities. The location with the most sales is Braga and the lowest is Faro.
- **What is the pharmacy who sell more by month?** With a bar chart, it is possible to see that every month, pharmacy B has more sales.

By leveraging Power BI's advanced visualization tools, Pharma4all can gain deep insights into their sales data, enabling informed decision-making and strategic planning.

12.3. Answer to Business problem

Pharma4all is facing challenges due to the manual extraction of CSV files from its servers, leading to slower decision-making and higher error rates. To address these issues, they require an automated solution that improves efficiency, minimizes human involvement, and maintains data accuracy. By utilizing Power BI for data visualization and Data Factory for automated data extraction, Pharma4all can streamline the process of accessing and analyzing data. Data Factory excels in connecting to various data sources, extracting data in real-time, and loading it into destinations like Power BI, where it can be visualized and analyzed. This combination enables Pharma4all to achieve quicker data availability, decrease manual errors, and allocate resources more effectively towards strategic initiatives.

13.Conclusions

The report outlines a strategy for Pharma4all to implement a Data Warehousing system, addressing manual data extraction issues and decision-making inefficiencies. The project is aligned with the company’s immediate needs, offering a path to quickly benefit from BI capabilities. This method ensures that data management efforts are both practical and aligned with the strategic goals of Pharma4all, setting a clear direction for the project.

With that in mind, the report lays out a detailed plan for developing a dimensional model and a BI implementation methodology that is expected to provide insights into sales, products, and operations, thereby improving business outcomes.

In the second part, the process of Extract, Transform and Load to build a Lakehouse architecture was elaborated. This part facilitates the integration and analysis of large volumes of structured and unstructured data from various sources. Implementing a lakehouse supports the scalability of data

analytics efforts, ensuring that Pharma4all can leverage advanced analytics capabilities for deeper insights and more informed decision-making.

The final step was leveraged by the semantic model provided by Fabric and previous ETL steps to create a Power BI Dashboard. The tool helped to answer key business questions by presenting data in an intuitive and interactive manner. This dashboard not only enhances data visibility but also empowers stakeholders with actionable insights, as required by the scope of the project.

14.References

Gallo, A. (2014, October 29). The value of keeping the right customers. Harvard Business Review. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

McIlroy, A., & Barnett, S. (2000, December 1). Building customer relationships: Do discount cards work?. Managing Service Quality: An International Journal.

Kukreti, D. (2023, 07 26). Dimensional Modeling with Kimball. Medium. https://medium.com/@deeksha.kukreti18/kimballs-dimensional-modeling-f64c559cdfed

Drugs.com. (2024, June 2). Prinivil + HCT + Amlodipina. Drugs.com. https://www.drugs.com/search.php?searchterm=Prinivil+%2B+HCT+%2B+Amlodipina

Instituto Nacional de Estatística (INE). "GDP per capita by NUTS III." June 9, 2024. Available at: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0009975&selTab=tab0

Statista. "Pharmaceuticals - Portugal." June 9, 2024. Available at: https://www.statista.com/outlook/hmo/pharmaceuticals/portugal#revenue