Tunis Business
University of Tunis

Brazilian e-commerce Dataset
Business intelligence mini project

FINAL PROJECT REPORT

Submitted to : Prof. Manel Abdelkader
January, 2024

TABLE OF CONTENTS

1. INTRODUCTION

Welcome to the culminating report of our Business Intelligence and Database Management Systems mini-project, a comprehensive exploration undertaken with the goal of unraveling insights within the realm of a Brazilian e-commerce. Our journey involves the strategic integration of powerful tools such as Talend for ETL processes, MySQL for robust data storage, and Power BI for OLAP and visualization.

At the heart of our study is a dataset sourced from a prominent Brazilian e-commerce platform. Our focus lies in dissecting the intricacies of various influencing factors – be it sellers, products, categories, or customer locations – to decipher their impact on product ratings. By delving into these relationships, we strive to unearth valuable insights into consumer behavior and preferences, providing a roadmap for our e-commerce partner to enhance profitability, elevate earnings, and catalyze sustainable business growth.

Methodology:

Utilizing Talend for Extract, Transform, Load (ETL) processes ensures seamless data integration, cleansing, and transformation. The dataset is then stored in MySQL, a robust database management system, providing a solid foundation for our analytical endeavors. The Power BI tool serves as our lens into the data, enabling On-Line Analytical Processing (OLAP) and visually representing complex relationships, trends, and patterns.

Key Objectives:

Understanding Influential Factors: Our primary objective is to discern the impact of sellers, products, categories, and customer locations on product ratings.

Strategic Insights for Business Improvement: Through our analysis, we aspire to equip the e-commerce entity with strategic insights to improve overall customer satisfaction, refine operational strategies, and foster sustained business success.

Enhancing Profitability: By unraveling the dynamics of consumer behavior, we aim to guide the e-commerce platform towards strategies that enhance profitability and earnings.

2. Work Explanation

2.1. Data Gathering

Our comprehensive data exploration embarks on the utilization of the Brazilian E-Commerce Public Dataset by Olist ,sourced from the Kaggle website. In a meticulous selection process, we curated essential files aligning with our Key Performance Indicators (KPIs). This amalgamation of datasets seamlessly combines both CSV and JSON formats, presenting a diverse and rich source of information.
The dataset we worked with contain this combinaison of data files :

- olist_customers_dataset.json
- olist_geolocation_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.json
- product_category_name_translation.csv

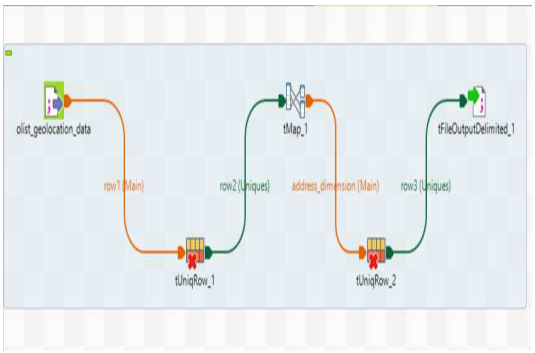
2.2. ETL Process

2.2.1. Data Extraction

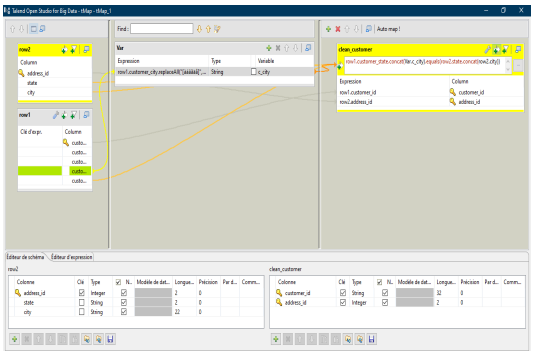
Our data extraction process was facilitated using Talend, a powerful ETL (Extract, Transform, Load) tool. The objective was to obtain relevant information from our dataset by excluding unnecessary columns and including only the essential ones. This involved the utilization of key Talend components such as tMap, tSort, and other relevant methods.

As an illustrative example, let's delve into the extraction of data from the geolocations file. Our focus here was to retrieve specific fields like address-id , state and city. The Talend job

employed a series of transformations and mappings to streamline the data extraction process efficiently.



(a) Data Extraction Transformation in Talend



(b) Output Configuration for Extracted Data

Figure 2.1: Screenshots Illustrating the Data Extraction Process

In Figure 2.4a, you can observe the Talend job configuration for data extraction, where transformations like tMap were employed to filter and manipulate the columns. Figure 2.4b showcases the final output configuration, ensuring that only the required fields are included in the extracted data.

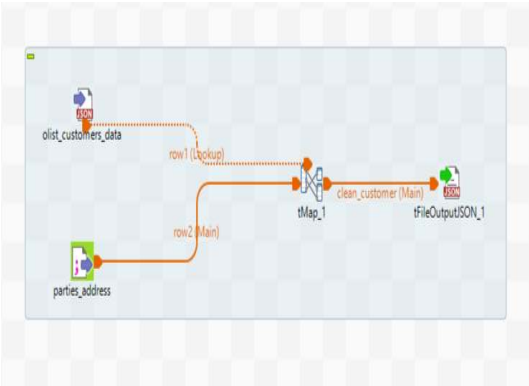
This meticulous extraction process lays the foundation for subsequent stages in our ETL pipeline, setting the stage for meaningful analysis and insights in the later phases of our project.

2.2.2. Data Transformation

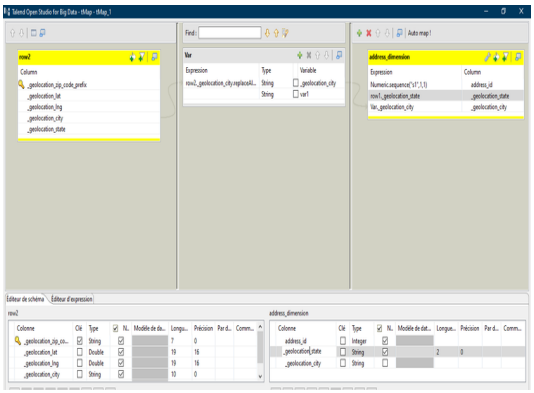
In the data transformation phase, it became imperative to optimize certain data for more efficient processing, and Talend played a pivotal role in achieving this objective. Our focus was on manipulating data to ensure a standardized and user-friendly format, creating a unified and usable dataset.

Let's delve into the transformation of customer data files as an illustrative example. One of the critical transformations involved converting the "prefix-zip-code" column into "address-id," ensuring a unique identifier for each state-city combination. Additionally, we addressed the issue of city name variations (e.g., Sào Paolo, sao PAolo, SAO PAOLO) by standardizing them to a singular and consistent form, such as Sao Paolo.

The screenshots below provide insights into this transformation process:



(a) Talend Job Configuration for Data Transformation



(b) Data Transformation Output Configuration

Figure 2.2: Screenshots Illustrating the Data Transformation Process

In Figure 2.2a, you can observe the Talend job configuration designed for transforming the customer data. This involved specific transformations to handle the conversion of zip codes,

city standardization, and other necessary modifications. Figure 2.2b showcases the final output configuration, ensuring that the transformed data adheres to the desired format which a .csv form.

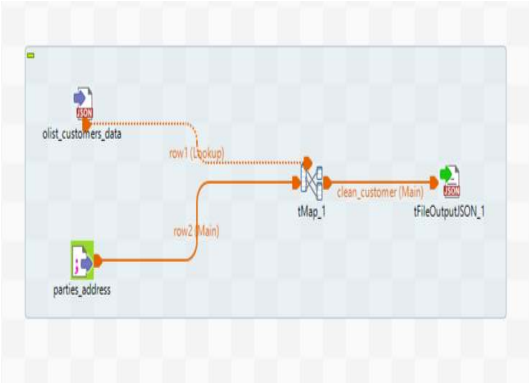
This meticulous transformation process enhances the quality and usability of our dataset, laying the groundwork for robust analysis and interpretation in subsequent stages of our project.

2.2.3. Data Loading

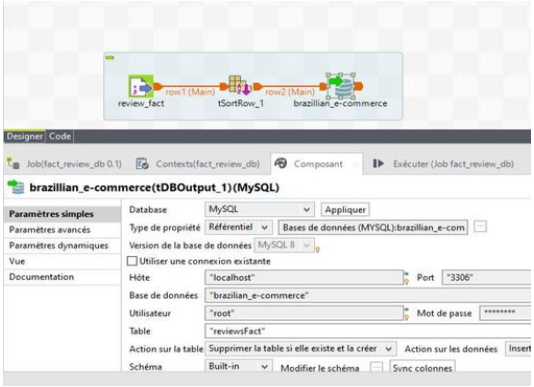
In this crucial step, our aim was to consolidate the transformed data into a unified format. We adopted two distinct approaches for data storage and accessibility: utilizing .csv files for seamless manipulation within Power BI and implementing SQL storage through MySQL server for robust data warehousing.

The .csv format proved highly advantageous, facilitating easy data manipulation and analysis within Power BI. Meanwhile, our exploration into SQL storage aimed at creating a structured and organized data repository for enhanced data conservation.

To illustrate this process, the following screenshots provide insights into our data loading configuration:



(a) Loading Data into .csv Format



(b) MySQL Server Configuration for Data Loading

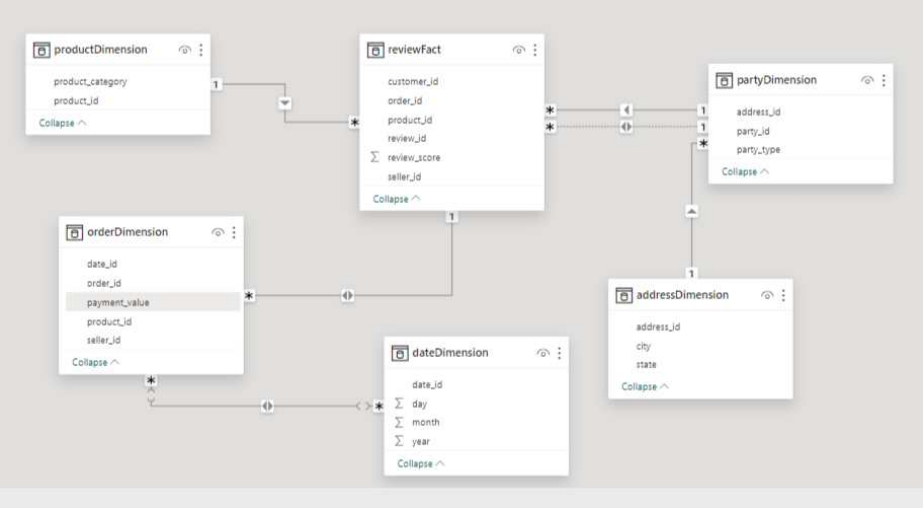
Figure 2.3: Screenshots Illustrating the Data Loading Process

In Figure 2.3a, you can witness the Talend job configuration designed for loading data into the .csv format, optimizing it for subsequent analysis in Power BI. Figure 2.3b showcases the MySQL server configuration tailored for efficiently loading and conserving data.

This dual approach ensures flexibility in data utilization, offering the advantages of both file-based manipulation and structured database storage.

2.3. Data Modeling

In this section, we delve into the data modeling process, focusing on structuring our dataset to enable efficient analysis and reporting. The primary components of our data model include the Fact Table, Dimensions, Measures, and the chosen schema. This careful design aims to provide a comprehensive foundation for deriving valuable insights from the Brazilian E-Commerce dataset.



2.3.1. Fact Table: Reviews

The central element of our data model is the **Reviews** table. This Fact Table captures essential information about customer reviews, including Review_id, associated Customer_id, Seller_id, Order_id, Product_id, and the corresponding review scores. These components collectively form the core of our analytical focus, providing valuable insights into customer feedback and product performance.

2.3.2. Dimensions

To enhance the depth of our analysis, we utilize several Dimension tables, each providing a unique perspective on the data. These dimensions include:

- **Product Dimension:** Describes the products involved, featuring attributes like Product_id, and Category.
- **Party Dimension:** Represents parties involved in transactions, distinguishing between customers and sellers. Key attributes include Party_id, Party_type, and Party_address_id.
- **Address Dimension:** Provides geographical details with attributes like Address_id, City, and State.
- **Order Dimension:** Represents the orders made by customers, including attributes such as Order_id, Product_id, Payment_value, Date_id, and Seller_id. This dimension adds a crucial layer to our analysis, connecting the Reviews Fact Table with the specifics of each order.
- **Date Dimension:** Contains temporal information with attributes such as Date_id, Year, Month, and Day. It enables time-based analysis.

2.3.3. Measures

Our analysis requires quantitative metrics, known as Measures, to quantify performance and trends. In the context of our data model, key Measures include only :

- **Ratings:** Derived from the Ratings table, capturing the review scores given by customers to orders.

2.3.4. Schema Choice: Snowflake Schema

We opted for a **Snowflake Schema** to organize our data efficiently. This schema's normalized structure facilitates easy maintenance and reduces data redundancy. It fosters a clear separation between dimensions and the central fact table, promoting scalability and flexibility in our analytical endeavors.

2.3.5. Purpose of Schema and Model

The Snowflake Schema's choice aligns with our goal of conducting in-depth analyses of the Brazilian E-Commerce dataset. By structuring our data in this way, we aim to streamline complex queries, enhance performance, and provide a solid foundation for reporting and visualization tools.

Through this data model, we seek to uncover intricate relationships between various dimensions and measures, shedding light on factors influencing customer behavior, product performance, and regional trends. This insight, in turn, empowers the e-commerce platform to make informed decisions, optimize strategies, and ultimately enhance overall business performance.

2.4. OLAP and Data Visualization Process

To harness the insights gleaned from our comprehensive data model, we employed Power BI as our OLAP and Data Visualization tool. This powerful tool allowed us to create interactive and insightful reports, transforming raw data into meaningful visualizations. The key steps in this process include:

2.4.1. Data Connectivity

Power BI seamlessly connects to a variety of data sources. We linked our data warehouse, hosted on MySQL server, to Power BI, ensuring real-time access to the most up-to-date information. The direct connectivity facilitated smooth data extraction for reporting.

2.4.2. Data Import and Transformation

Once connected, we imported data from our MySQL server into Power BI. Leveraging Power BI's transformation capabilities, we refined and shaped the data to meet the specific requirements of our analysis. This step included handling any necessary data cleansing, filtering, or transformation to enhance the quality of our visualizations.

2.4.3. Data Modeling in Power BI

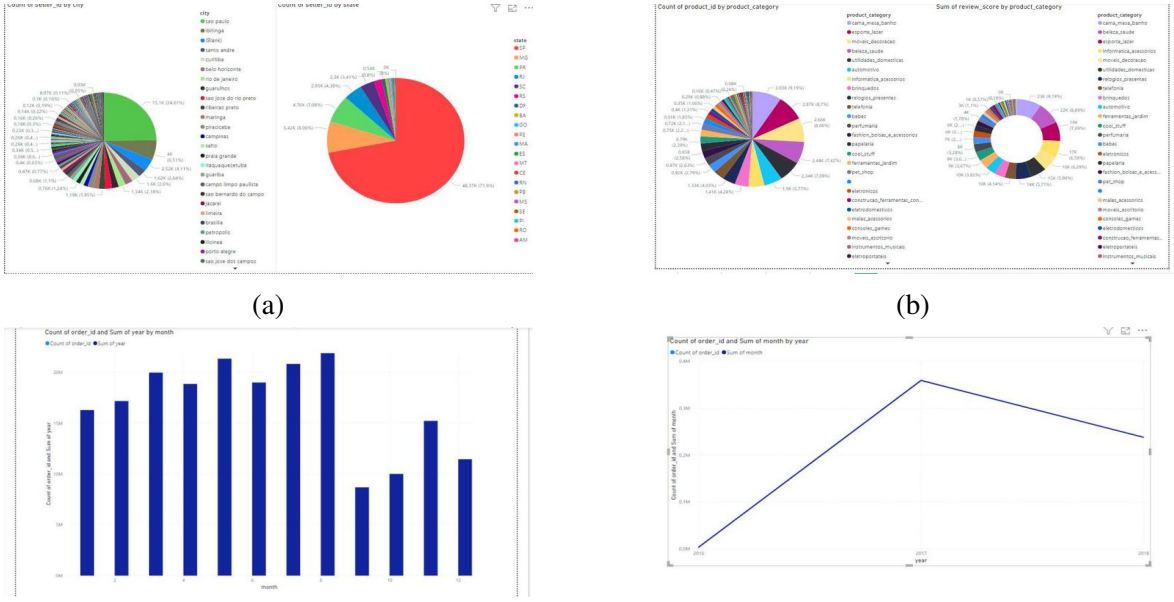
Power BI's robust data modeling features allowed us to define relationships between our Fact and Dimension tables. We mapped out the associations between the Reviews Fact Table and the various Dimension tables, such as Date, Product, Party, Address, and the newly introduced Order Dimension.

2.4.4. Measures and Metrics

In Power BI, we defined measures and metrics based on our business goals. These measures provided key performance indicators (KPIs) to evaluate the success of the e-commerce platform. Examples of measures include average review scores, sales trends, and customer satisfaction metrics.

2.4.5. Interactive Dashboards and Reports

Power BI's drag-and-drop interface enabled us to create interactive dashboards and reports effortlessly. We designed visualizations such as line charts, bar graphs, and geographic maps to convey insights effectively. Users can explore the data dynamically, gaining a comprehensive understanding of the Brazilian e-commerce landscape.



(a) (b)

(c) (d)

Figure 2.4

2.4.6. Insights and Decision Support

The culmination of our OLAP and Data Visualization process is the generation of valuable insights. Decision-makers can use these insights to refine strategies, enhance customer satisfaction, and make informed business decisions. The interactive nature of Power BI empowers users to explore data autonomously and extract actionable intelligence.

Through this OLAP and Data Visualization process, we've transformed raw data into a dynamic and insightful tool for decision-making, enabling stakeholders to navigate the complexities of the Brazilian e-commerce domain.

A strategic application of a variety of metrics in data visualization was utilized to understand the data and derive valuable insights.

Geospatial Analysis Olist has achieved extensive geographic coverage, spanning 21 states in Brazil. The majority of the customer base is concentrated in Sao Paulo (SP), Minas Gerais (MG), Parana (PR), and Rio de Janeiro (RJ). The top three cities for both customers and sellers are Sao Paulo, Ibitinga, and Santo Andre.

Product Categories Analysis The product categories with the highest demand are Bed Table Bath (cama_mesa_banho), Sports & Leisure (esporte_lazer), Furniture & Decor (moveis_decoracao), Beauty & Health (beleza_saude), Housewares (utilidades_domesticas), and Auto (automotivo). The most reviewed product categories are Bed Table Bath, Beauty & Health, and Sports & Leisure.

Orders Analysis The number of orders peaked at 0.36 million in 2017, but declined to 0.24 million by the end of 2018. The highest customer order trends are observed in August, May, and July, while the lowest trends are seen in September, October, and December

3. CONCLUSION

Summing up, to overcome mishaps such as the decrease in orders, and the concentration of customers and sellers in the urban areas only, etc, Olist has to implement the following solutions:

- Focusing more on the high-demand categories to drive sales and revenue growth.
- Entering new states and cities to broaden the market coverage.
- Developing targeted marketing strategies and inventory planning to align with customer order trends, especially during peak months.
- Focusing on promoting the top-performing products.
- Increasing the collaborations with new sellers to expand the market diversity