جامعة الأخوين
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵏ ⵓⵅⴰⵡⴰⵢⵏ
AL AKHAWAYN UNIVERSITY

Fall 2022

CSC 5301
Advanced Database Systems and Data Warehousing

Project Final Report:
Car Sales

Name	ID
Othmane Atanane	**76576**
Yassir Benabdallah	**76865**
Meryem Abdallah	**75657**

Supervised by:

Dr. Nasser Assem

Table of Contents

I. Introduction

Our business is a car sale company called “Kolchisog. It includes many agencies that help customers buy vehicles according to their needs and requirements. Its primary activities include the import, purchase, and sale of new and used automobiles. Kolchisog used to keep track of all sales-related information in an excel sheet.

During the last 15 months, Kolchisog noticed a drop in its sales, thus she decided to hand in its data for analysis. The analysis should give the company an idea of the reasons behind that drop and help them determine a new strategy to boost sales and increase profit. Moreover, the company should be able to extract and visualize meaningful patterns.

II. Project Objective

This project will involve the design and implementation of a data warehouse that will serve as the central data repository for all operational data collected over the years. This project's main goal also includes running analytical queries on the data warehouse based on precise client requirements, and finally producing user-friendly statistics that will have a critical effect in the decision-making process.

III. Requirements Gathering

In order to discover patterns and connections within their data set that gets updated daily in addition to the fact that it is historical, Kolchisog would like to store their data in a data warehouse. The current data of the company, which is collected from different branches distributed across many regions of the United States, is stored in the form of operational database, which summarizes our task in converting this huge data in the form of data warehouse. The goal behind this operation is to assist Kolchisog in making their decisions, by providing clearer visibility of the process of selling cars over all the branches. Therefore, they would like to:

- make the data warehouse updated regularly, preferably on a daily basis.
- examine the sales performed in different branches, cities, and time frames.
- have an idea about how the human resources (salespersons) are performing.
- generate a recommendation system based on the purchase patterns of each customer.

IV. Requirements Specification

- The data warehouse should extract data from a database of the Kolchisog agencies.
- The data brought from the databases should be cleaned and structured before being uploaded to the data warehouse.
- The data warehouse should contain all the information related to car sales, by the corresponding agency, to a client, on a specific date.
- The data warehouse should be updated daily.
- The data warehouse should be queried by the users using a tool with an easy-to-use interface.
- The user shall be able to generate reports in standard formats such as pdf and excel.
- The user shall be able to dynamically switch back and forth from high levels of details to summarized views.

V. Project Timeline

The following figure shows we managed our time to come up with this final project. The figure is illustrated using Gant Chart.

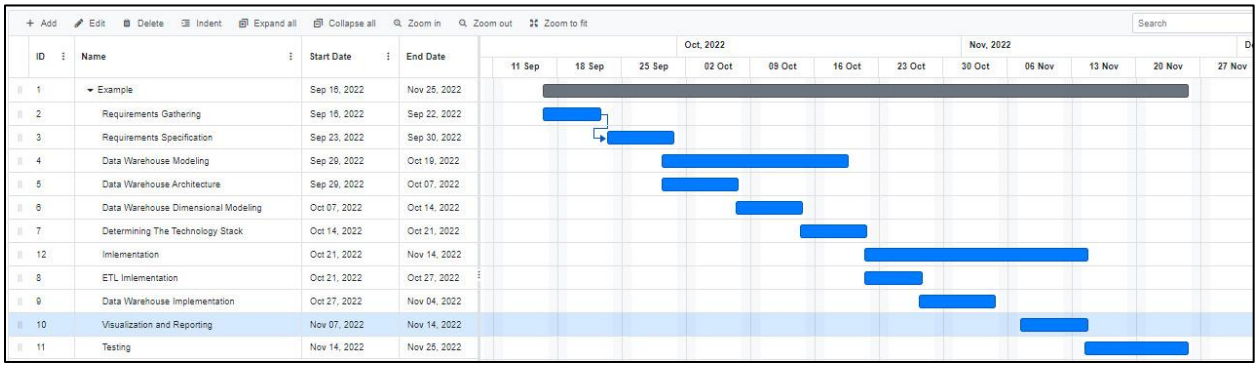

Figure 1: Gant Chart

VI. Design and Architecture

1. Entity Relationship Diagram

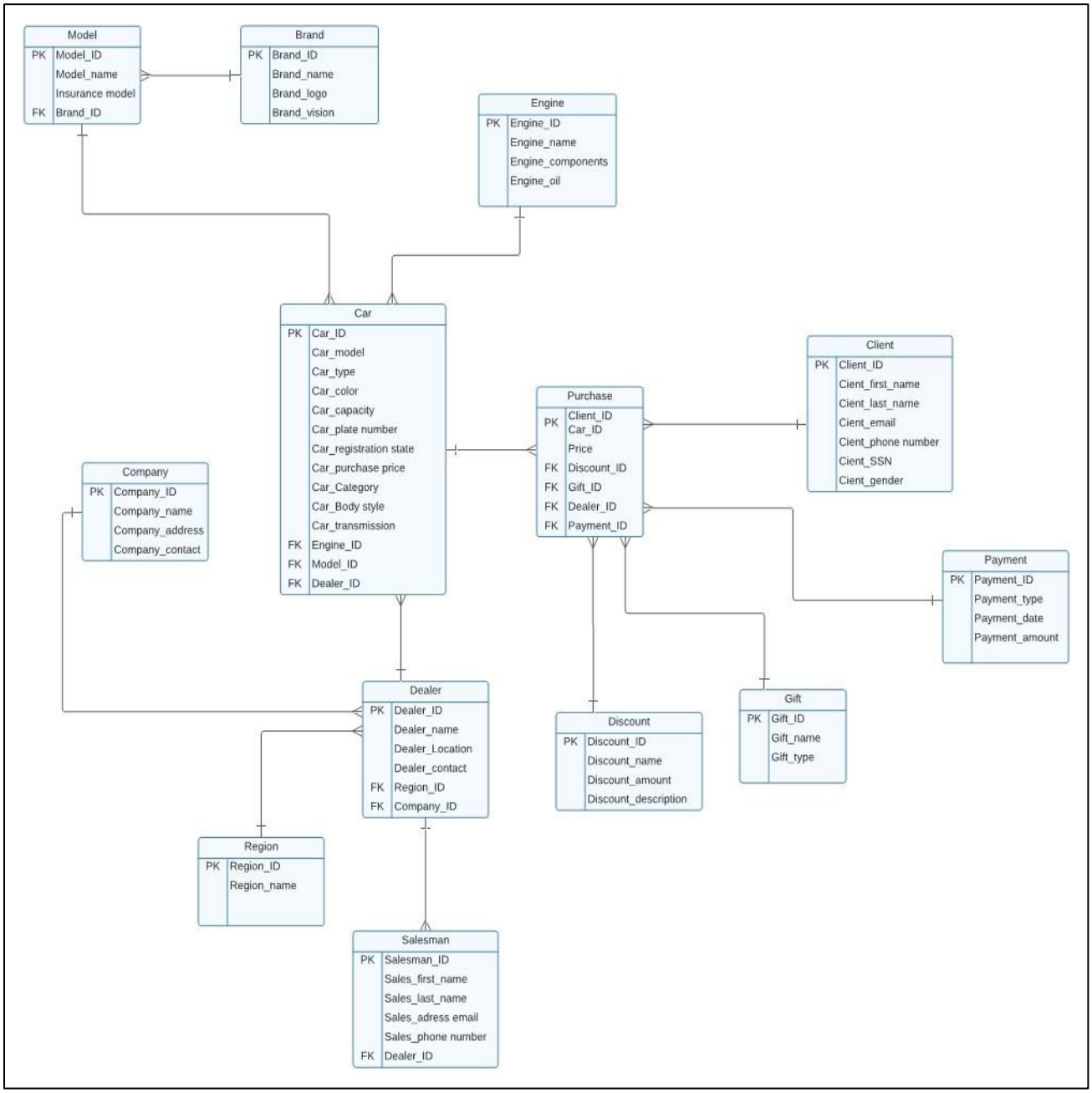

Figure 2: Entity Relation Diagram

2. Data Warehouse Architecture

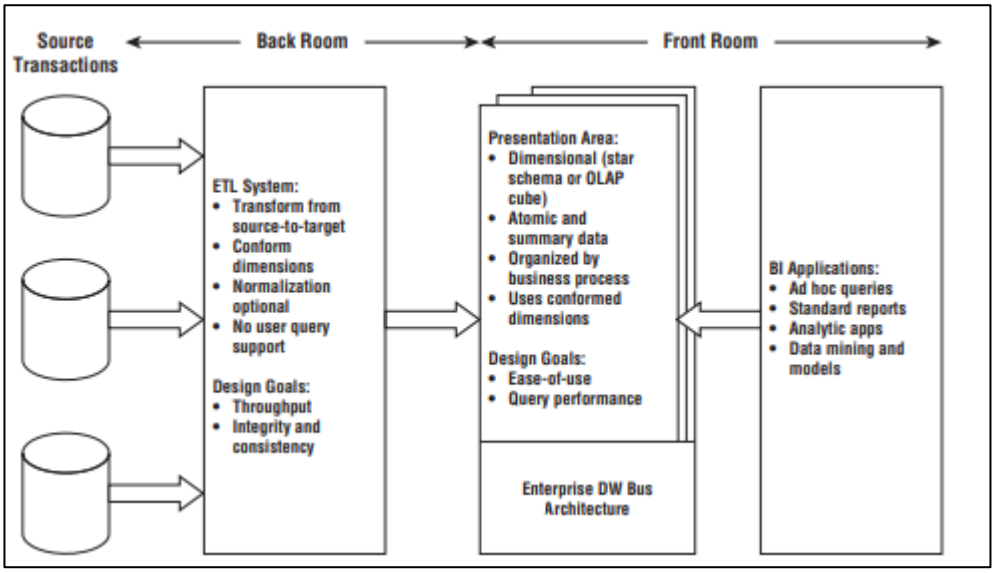

Figure 3: Core elements of the Kimball DW/BI architecture.

The Kimball data warehouse architecture, developed by Ralph Kimball, is a method for designing and building a data warehouse. It is based on the idea of a "bus architecture," in which a central data warehouse stores data from multiple sources, and a set of dimensional tables are used to organize and structure the data for querying and analysis.

One key feature of the Kimball architecture is the separation of the ETL (extract, transform, load) process from the front-end tools used for querying and analyzing data. The ETL process is responsible for extracting data from various sources, cleaning and formatting it, and loading it into the data warehouse. This process typically occurs in a "back room" environment, separate from the tools used by end users to query and analyze the data.

The decision to follow the Kimball architecture in our project may have been made for a number of reasons. The denormalized ETL system and the separation of the ETL process from the front-end tools may have been seen as beneficial for the specific needs of the project. It is also possible that the decision was influenced by the fact that the Kimball architecture is a well-established and widely used method for designing data warehouses.

3. Dimensional Model

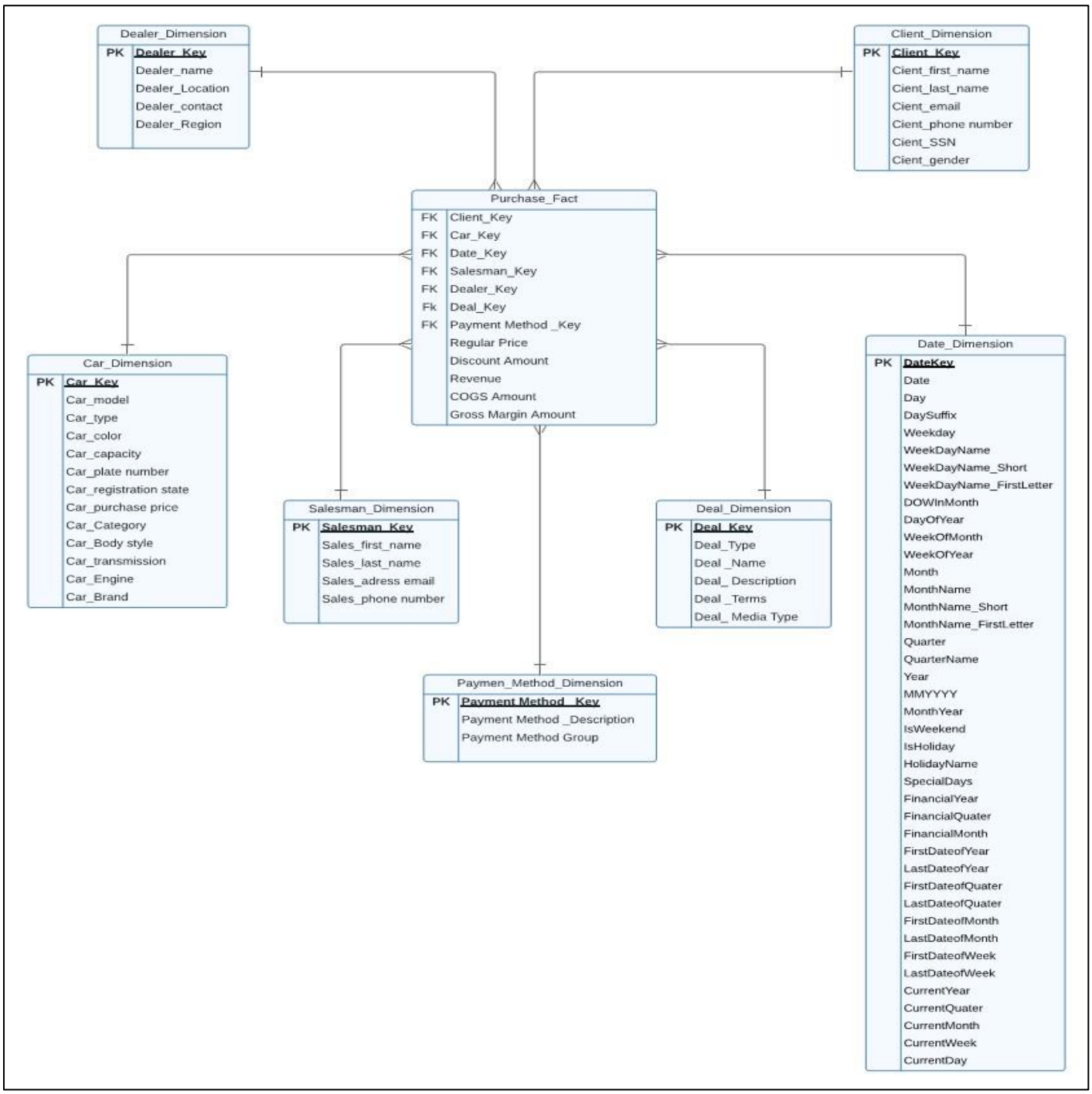

Figure 4: Dimensional Model

4. Granularity

Our dimensional model was designed to abide by the following granularity: each row of the in the fact table represents a purchase of a car by a client from a salesman's dealer on a certain day using a certain payment method.

VII. BI Tools

For this project, the following tools will be used:

ETL: Microsoft SQL Server Integration Services (SSIS) will be used to build the ETL system. SSIS is a data warehouse tool that performs extract, transform, and load (ETL) operations, as well as offering a range of built-in tasks.

Data Warehouse: Microsoft SQL Server will be used to build the data warehouse. SQL Server is a relational database management system (RDBMS) that is compatible with SSIS.

Reporting: Microsoft Power BI will be used as a visualization and reporting tool. Power BI is a business intelligence and data visualization tool that helps to convert data from various sources into interactive dashboards and reports and offers a range of software connectors and services.

VIII. Implementation

1. Star Schema Implementation Using SQL Server

We can now implement a physical data warehouse after finishing the dimensional modeling. We used SQL Server as relational database management system in order to create the start-schema as well as its corresponding relations. The dimensions created are empty, and it will be the mission of the ETL (next part of the project) to transform our relational database model to a data warehouse model.

The figure below shows the data warehouse dimensions.

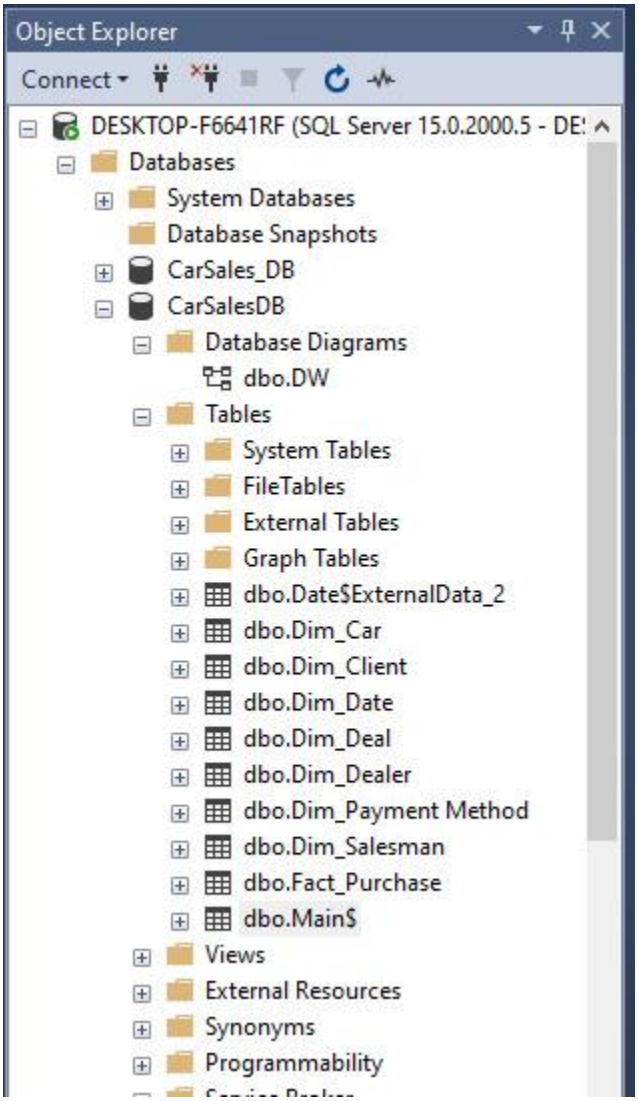

Figure 5: List of the Datawarehouse Dimensions in SQL Server

2. ETL System

ETL refers to Extract, Transform and Load. This is a pipeline that aims to extract the data from the operational sources that has been already loaded into the SQL server. Then we transform this data in a way that matches our data warehouse schema. Finally, we load the data to its destination, which is the tables created on SQL Server.

Figure 6 shows an errorless control flow of this ETL process using SSIS (SQL Server Integration Services).

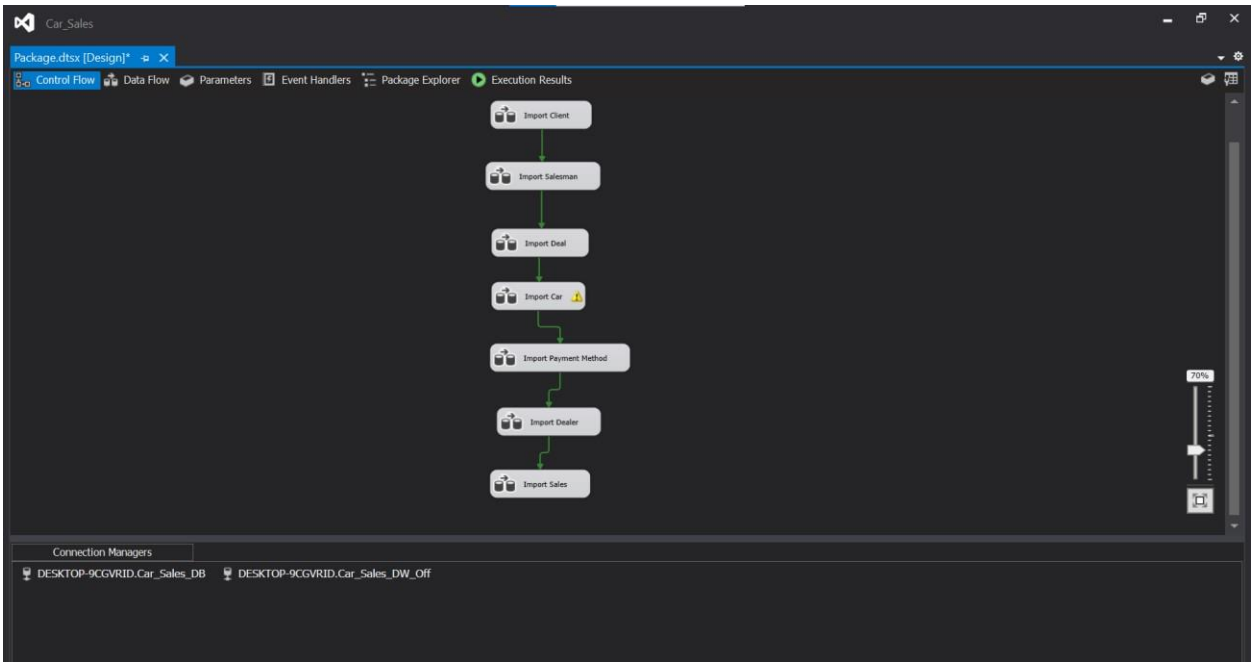

Figure 6: ETL Flow

a. Client Import

This task consists of extracting data from the client table in the operational database then transform it to fit into the data warehouse schema, then load it to the client dimension.

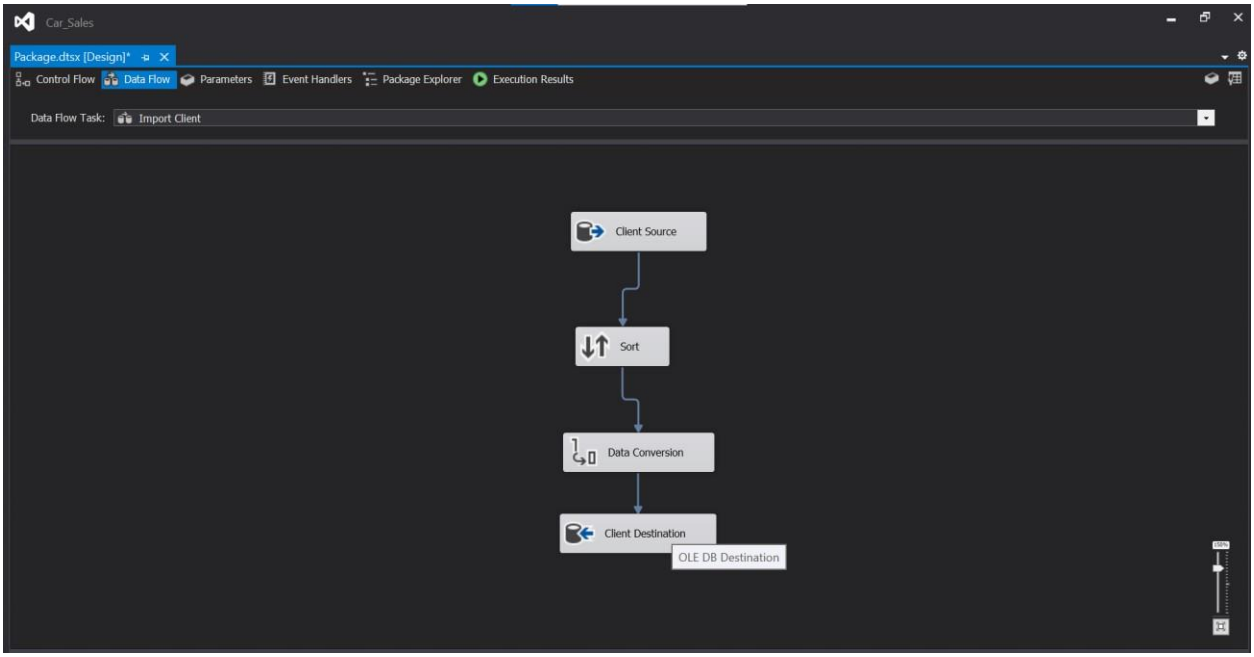

Figure 7: Client Import Data Flow

b. Salesman Import

Like the importation of the client, the ETL pipeline took care of migrating the data of salesman table to the salesman dimension.

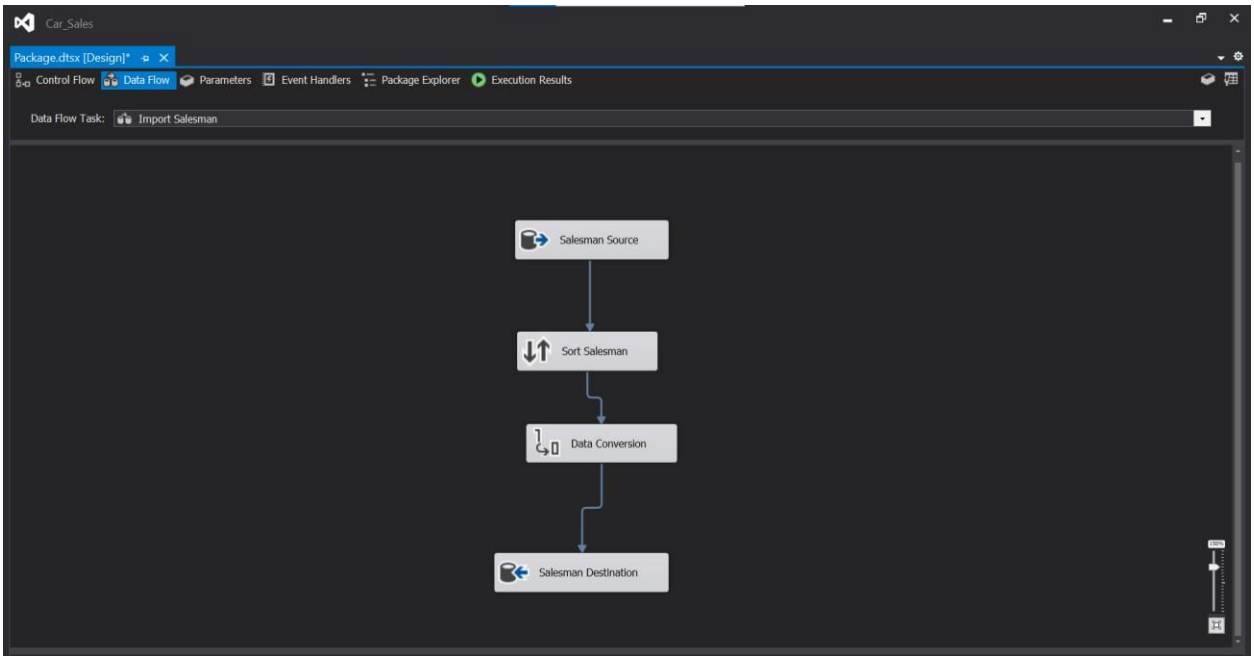

Figure 8: Salesman Import Data Flow

c. Deal Import

Unlike the previous tasks where the migration of data was smooth, this task required data denormalization in the sense that we had to join the discount table and the gift table in order to produce a new dimension called deal dimension.

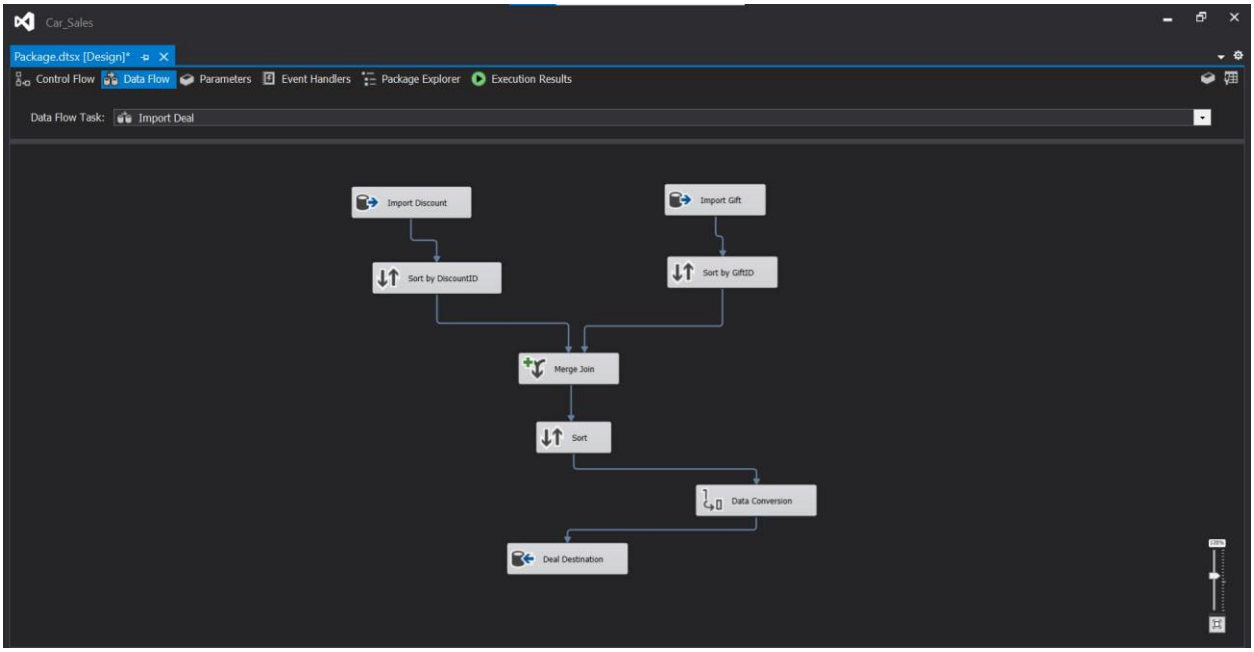

Figure 9: Deal Import Data Flow

d. Car Import

Again, we were obliged to perform data denormalization on multiple tables (Brand, Model and Engine) to produce new Car dimension.

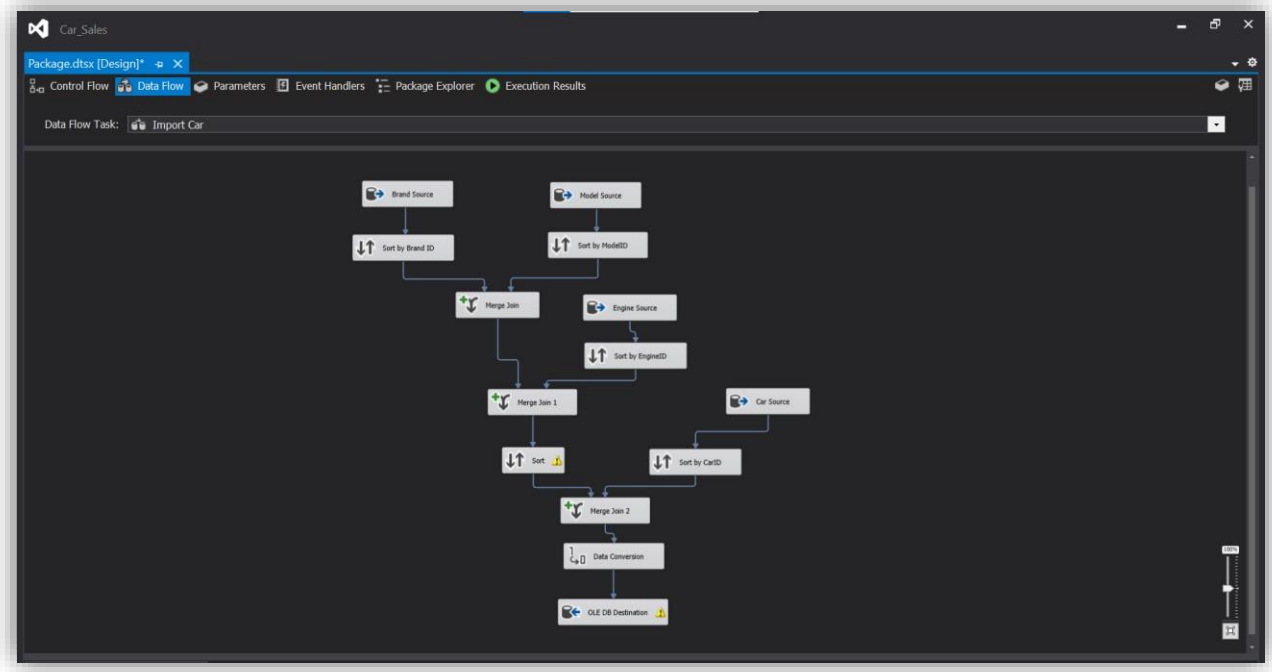

Figure 10: Car Import Data Flow

e. Payment Method, Dealer and Sales Imports

The data migration in these cases was very smooth, as it did not require any data denormalization.

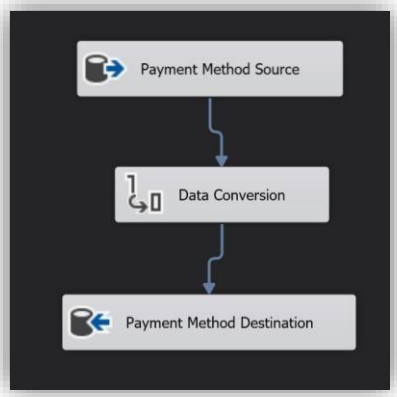

Figure 11: Payment Method Import Data Flow

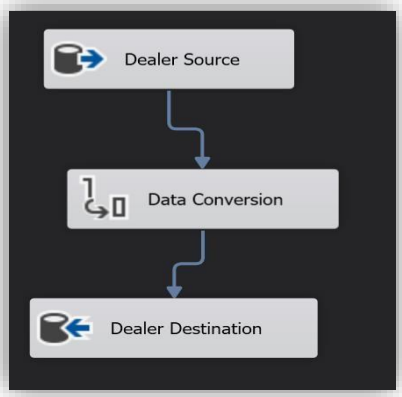

Figure 12: Dealer Import Data Flow

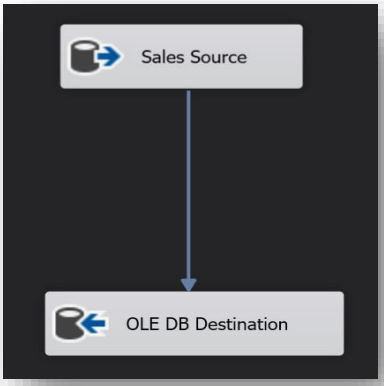

Figure 13:Sales Import Data Flow

3. Example of Star Schema Population Using ETL Systems

After running the ETL pipeline successfully, it was observed that the data warehouse had been correctly populated and that all of the constraints imposed by the design were respected.

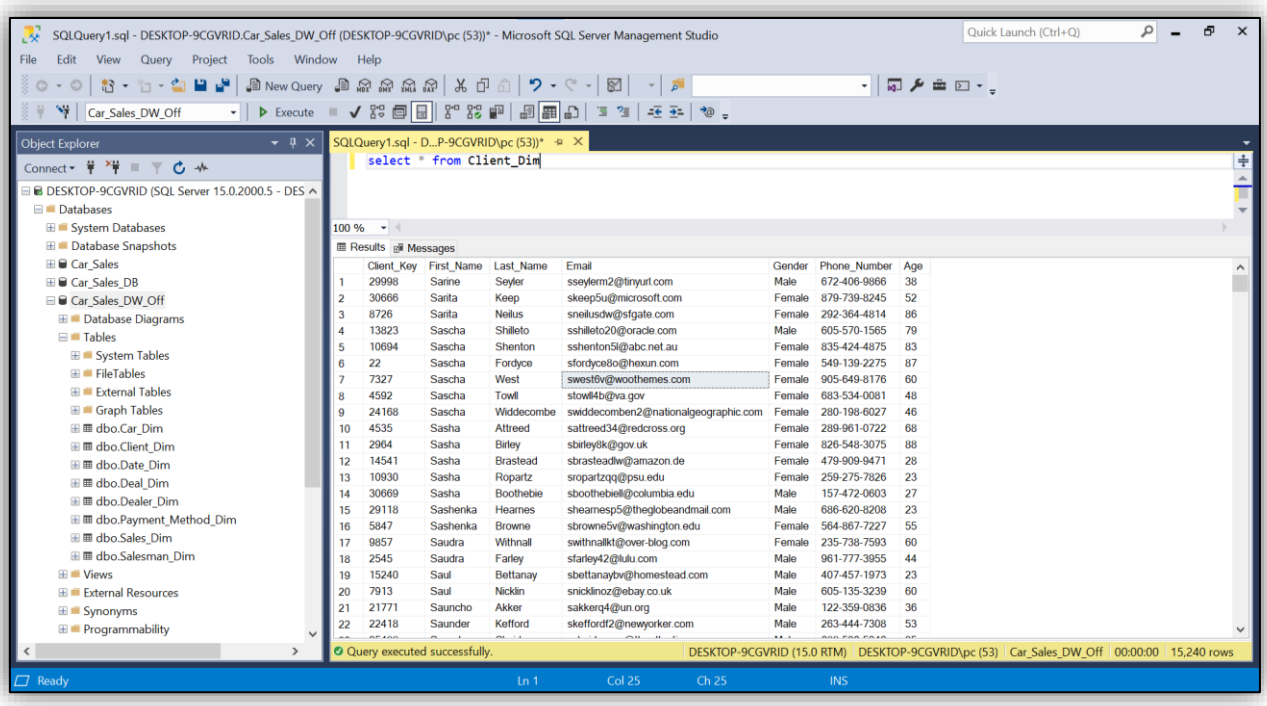

Figure 14: Schema Population Using ETL

4. OLAP Cubes Implementation

OLAP (Online Analytical Processing) is a type of database technology that is designed to support efficient querying and analysis of data. It is often used in business intelligence and data warehousing applications to allow users to easily analyze and understand large amounts of data.

An OLAP cube is a multidimensional data model that is optimized for querying and analyzing data. It is organized into a series of dimensions, which represent different aspects of the data, and measures, which are numerical values that can be aggregated and analyzed. The OLAP cube allows users to quickly perform various operations on the data, such as drilling down into specific details, rolling up to higher levels of aggregation, slicing the data to focus on specific subsets, dicing the data to look at multiple dimensions at once, and pivoting the data to change the way it is organized.

Microsoft SQL Server Analysis Services (MSSAS) is a tool that can be used to create and deploy OLAP cubes. It allows users to extract the cube schema from a data warehouse's star schema, which is a type of data model that is optimized for fast querying and data retrieval. The

cube schema can then be created within a Visual Studio project, which is a software development tool that allows users to design and build applications.

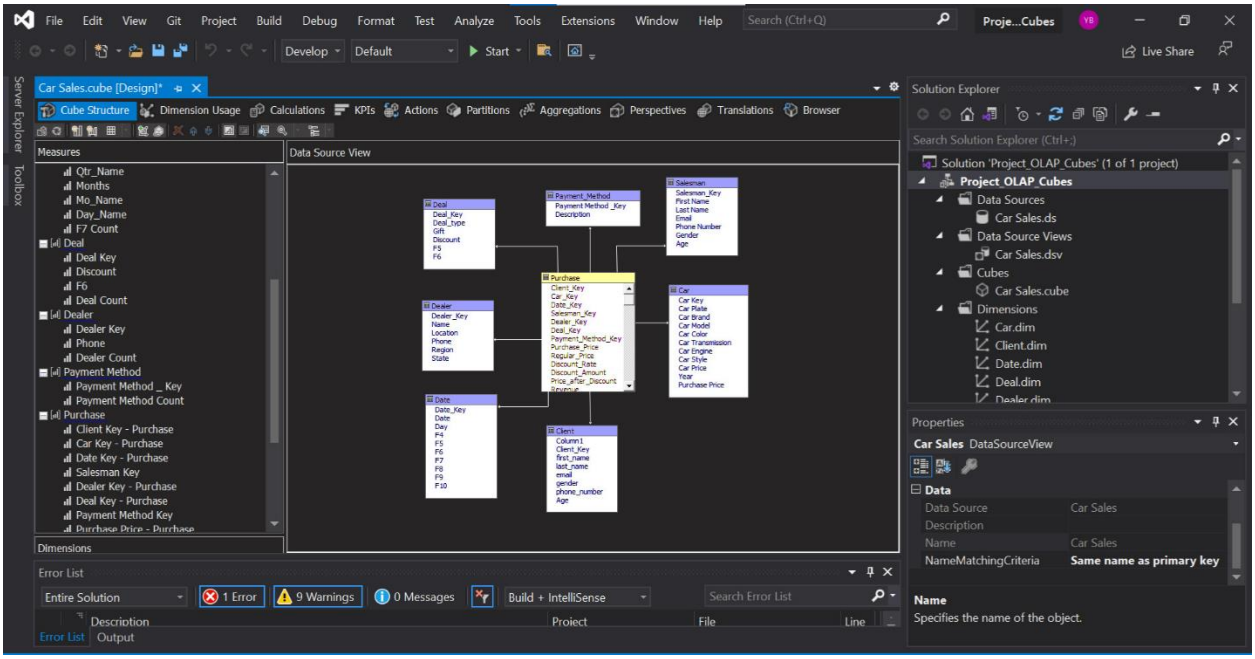

Figure 15: The Cube Schema

The figure bellow shows the different components created for the cube dimensions:

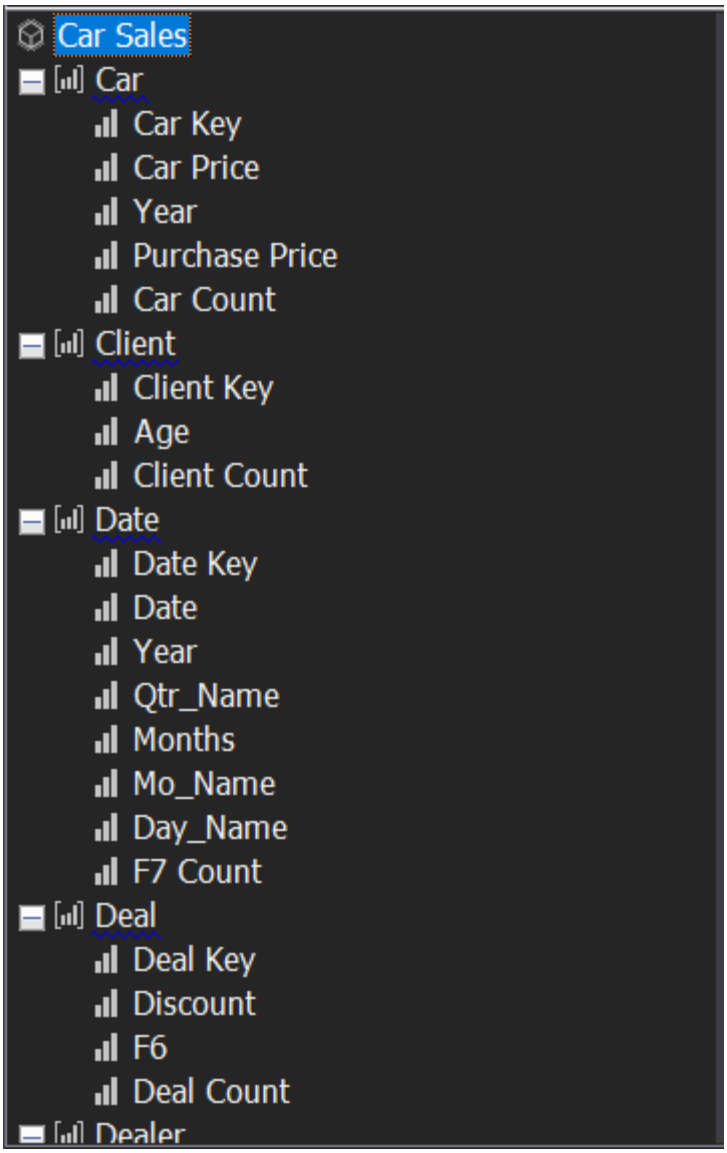

Figure 16: Components Visualization Using Cube Dimensions

IX. Visualization of Data Using Power BI

The aim of this part is to introduce a visualization tool called Power-BI that would allow the user, especially the dealers, to track their sales in a simple manner that does not really require knowledge about querying or coding. Moreover, thanks to its optimized and live ability to connect to SQL Server, Power BI will help us in easily generating charts, reports, and dashboards by directly working with a large amount of data.

The following figures show some visuals that illustrate the way Power BI provides the user with charts that can be easily understood.

1. Sales Based on Location

The following figures depict a business process that aims to display the number of sales made by the company based on geographic locations. The first figure presents data for only one country, Texas, due to limitations in the available data. However, for the purpose of demonstrating the capabilities of Power BI, we have applied a Slice operation to the data, even though we only have data for one country. This serves to showcase the software's support for OLAP operations and its user-friendly interface.

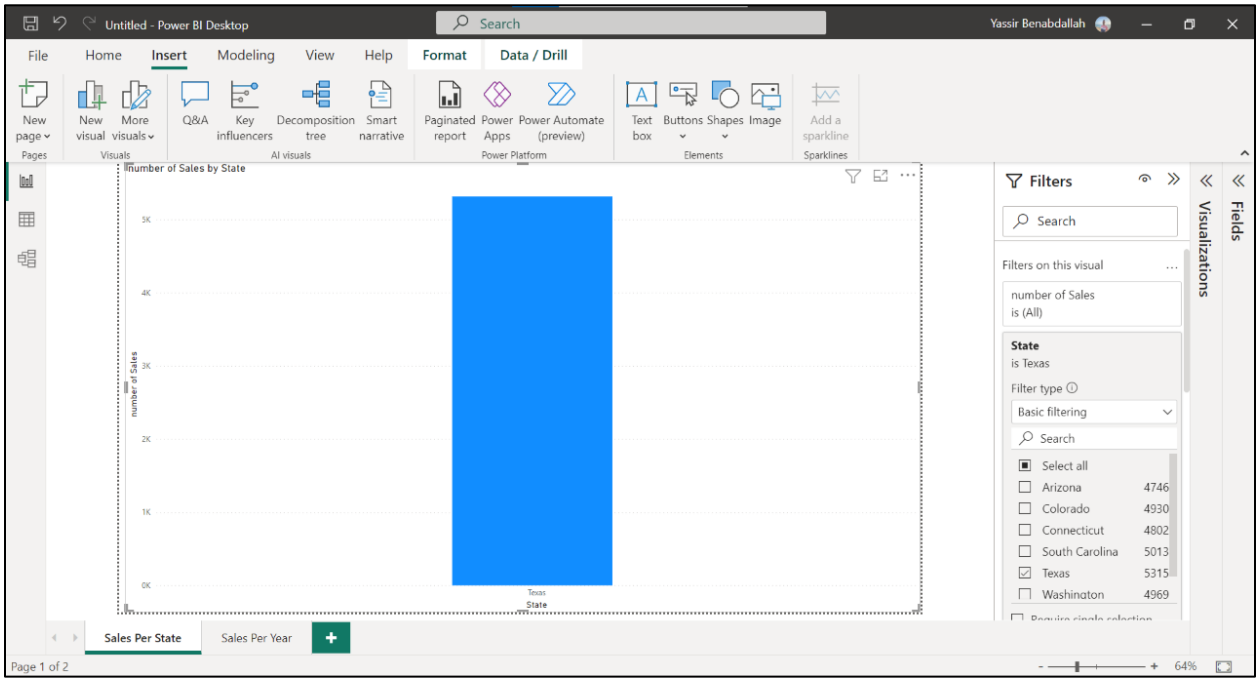

Figure 17: Sales Based on One Location

We can visualize more on many other states by choosing all of them on the filters. This would allow the managers to have clear and concise ideas about how the selling is going in different geographic locations, and to make quick comparisons without the need to perform Mathematical operations.

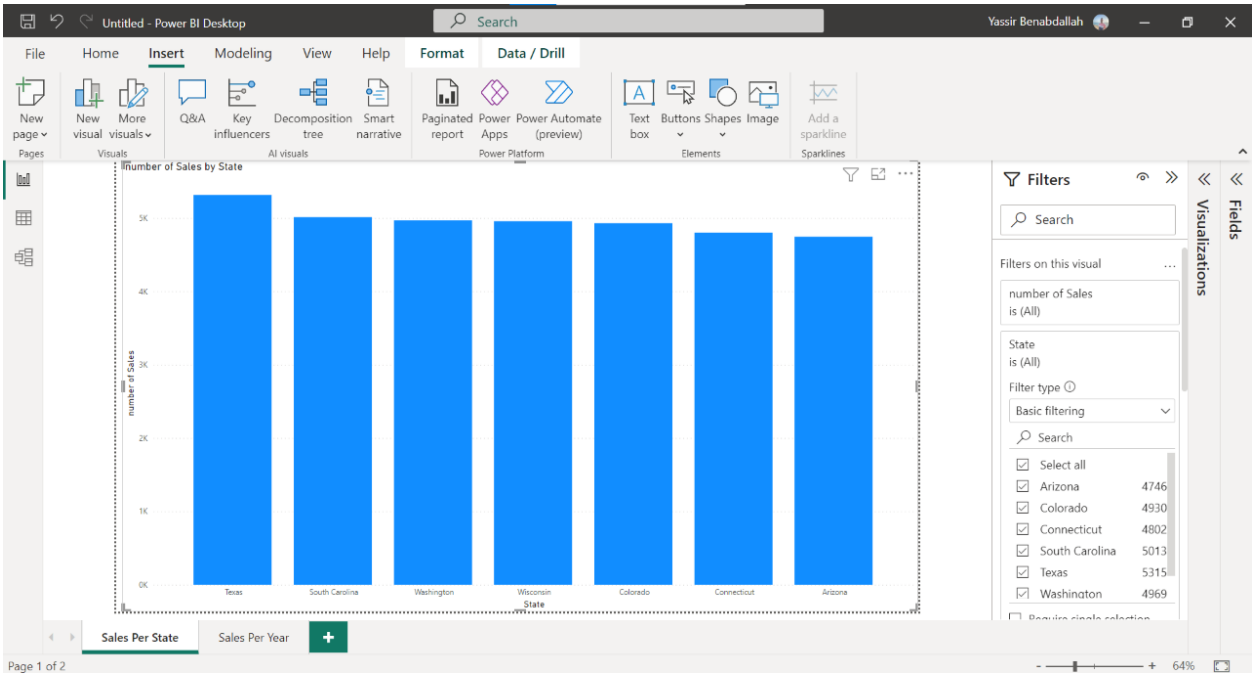

Figure 18: Sales Based on Location

2. Sales Based on Date

Again, the manager would like to access their sales by date in the sense that they would make analysis on which year they make high or low sales. The hierarchy concepts gets repeated in this section of the profit, by analyzing within one year which quarter makes the high number of sales.

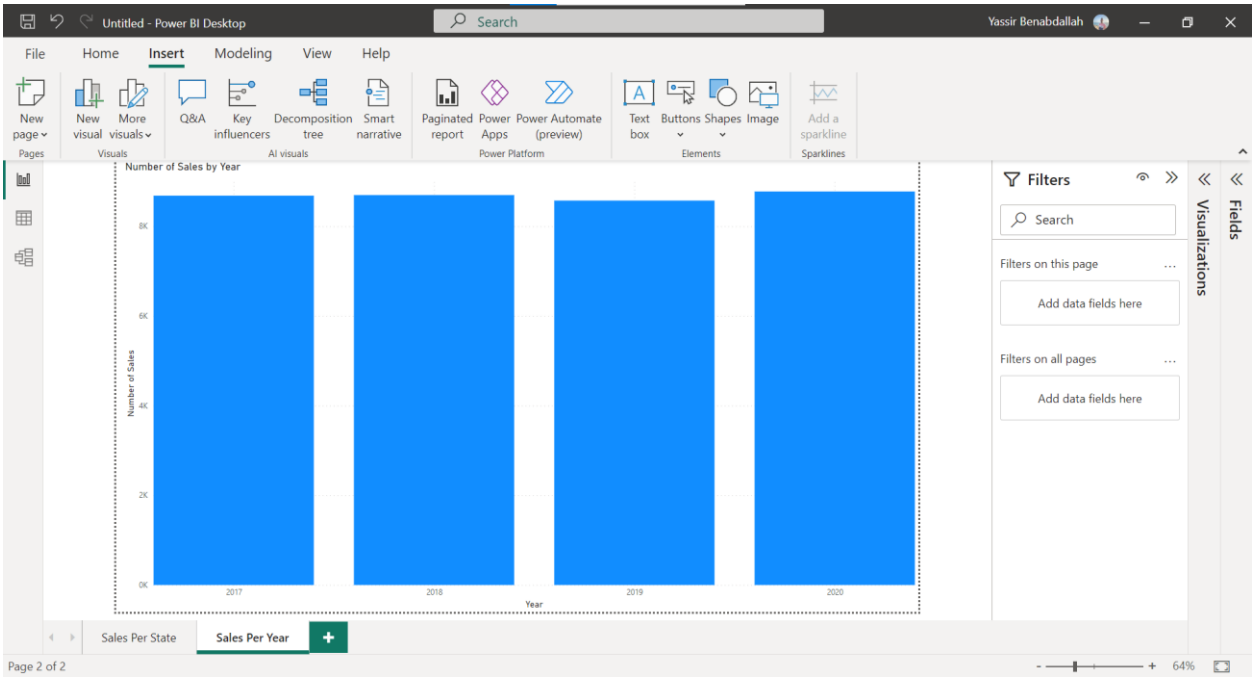

Figure 19: Sales Based on Years

The following figure shows the number of sales per quarters in 2020.

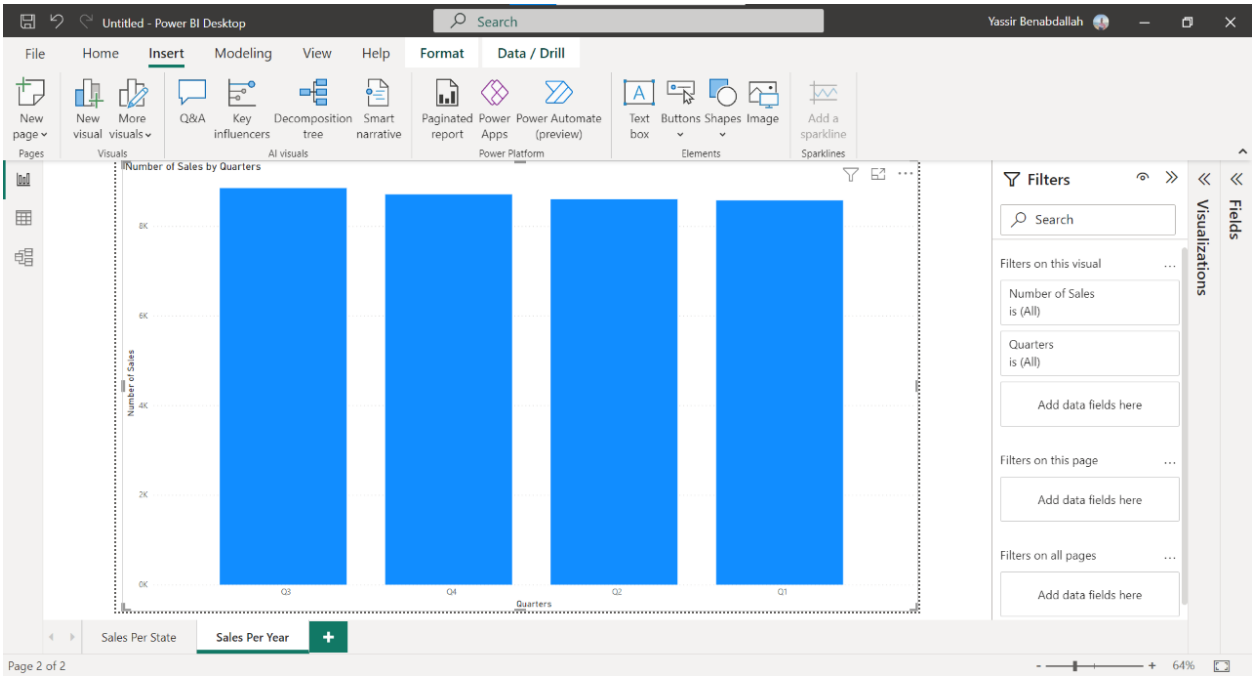

Figure 20: Sales Based in the Quarters of 2020

3. Revenue Per Location

For the following analytic view, we aim to show a graph that illustrates the revenue made by each state. The revenue on the Y axis is scaled in millions of Dollars.

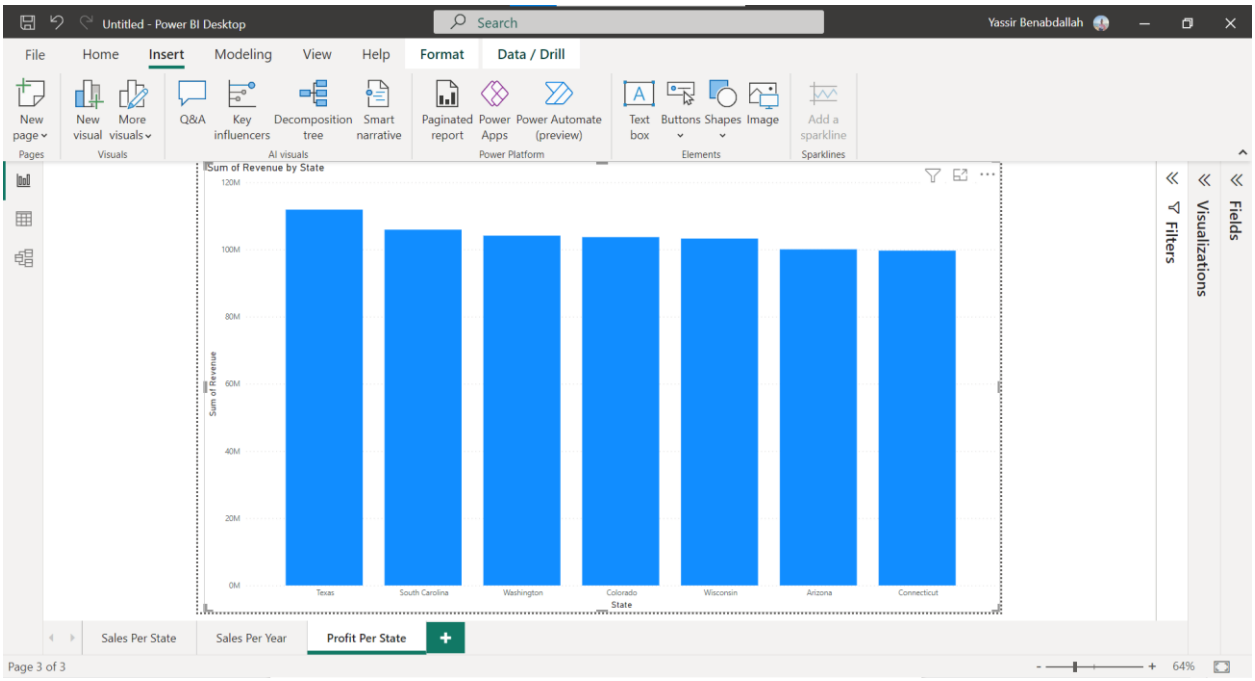

Figure 21: Revenue per Location

4. Revenue Per Date

The last analytic view shows the revenue made in each year. This gives an idea to the manager on which year produces the highest revenue, so that they develop the marketing, financial and managerial techniques that allowed them to make this profit. Also, avoid any problem the faced in the same year.

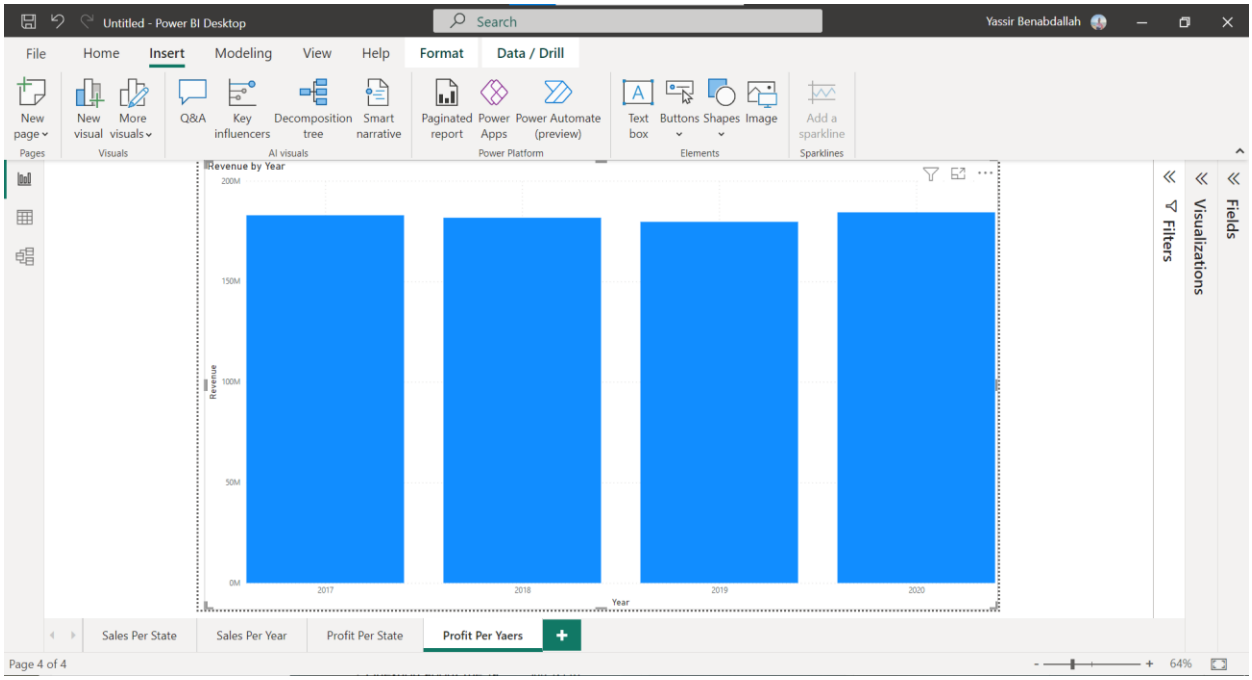

Figure 22: Revenue per Date

X. Conclusion

Data warehousing (DW) and business intelligence (BI) are processes and technologies used to collect, integrate, and analyze data from various sources to support business decision-making. In this project, the goal was to help decision makers at the company by identifying patterns in the company's data from the past 20 years. To achieve this, the project team followed the steps for DW/BI, which typically involve gathering requirements from the client, designing and building a database or data model to store and organize the data, extracting, transforming, and loading (ETL) the data from various sources into the database or model, and using the data to support business decision-making through reporting and analysis tools.

In this case, the team built a multidimensional model, which is a type of database designed specifically for data analysis and reporting. The model was populated using an ETL pipeline, which is a set of processes for extracting data from multiple sources, transforming it into a suitable format for the multidimensional model, and loading it into the model. The team also applied key business rules during the ETL process to ensure that the data was consistent and accurate.

Once the data was loaded into the multidimensional model, it was used as the source for an OLAP (online analytical processing) cube, which is a pre-calculated multidimensional database that allows users to perform fast querying and analysis of data. The OLAP cube

supported dynamic querying through Power BI, a business intelligence and data visualization tool, which allowed the decision makers to evaluate their business and make informed decisions. Overall, the project aimed to provide decision-making support by deriving meaningful patterns from the company's data and presenting them in a way that was useful and actionable for the decision makers.

References

Gagandeep16. (n.d.). Car sales [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/gagandeep16/car-sales

Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The complete guide to dimensional modeling (3rd ed.). John Wiley & Sons. Retrieved from https://aatinegar.com/wp-content/uploads/2016/05/Kimball_The-Data-Warehouse-Toolkit-3rd-Edition.pdf

Microsoft Learn. (n.d.). ETL [Web page]. Retrieved from https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl

Microsoft Learn. (n.d.). Defining and deploying a cube [Web page]. Retrieved from https://learn.microsoft.com/en-us/analysis-services/multidimensional-tutorial/lesson-2-defining-and-deploying-a-cube?view=asallproducts-allversions