DataCo Supply Chain Data Warehousing

Data Warehouse

Master's Degree in Data Science and Engineering - FEUP

April 2023

Group 5

Carlos Miguel Veloso (up202202463)

Cátia Teixeira (up200808037)

Luís Henriques (up202204386)

Rojan Aslani (up202204382)

Professor: Gabriel David

---

1. Introduction

Data warehouses are centralized data repositories that integrate data from various systems, applications, and sources. The data warehouse is an environment separate from the operational systems, and it is completely designed to provide for analytical and ad-hoc reporting, queries, and data analysis [1].

Supply chain encompasses all of the facilities, functions and activities involved in producing a product or service from suppliers to customers. The modern-day supply chain processes generate enormous amounts of data every day that is often overlooked by most enterprises. According to experts, this data holds immense value that may be useful to train BI tools and provide advanced analytical insights. Owing to the veracity of business intelligence tools, the SCM data may prove significant to enhance business productivity [2].

Nevertheless, to drive intelligent decision-making, BI tools require agile access to business data as well as effective means for data storage and retrieval. Considering the sheer volume of data, traditional database systems might not suffice to address the growing requirements of an enterprise. To serve this purpose, enterprises require data warehousing solutions that are conducive to storing and analyzing large volumes of data [2].

Data warehousing is of foremost importance in supply chain management as it collects and stores essential data related to vendors, supplies, shipments, and finished goods.

Supply chain functions include purchasing, inventory, production, scheduling, facility location, transportation and distribution. All these functions are affected in the short run by product demand and in the long run by products and processes and changing markets. Forecast of product demand determines how much product to make and how much material to purchase from suppliers to meet forecasted customers' needs. Due to the ever-increasing global competition, sales forecasting plays a prominent role in supply chain management. The availability of huge data demands tools that can extract information from data. Recent research has shown that advanced forecasting tools enable improvements in supply chain performance [3]. In this work, we shed light on the importance of data warehousing solutions for supply chain management. - To be used for descriptive analysis.

The requirements of this work included a reasonable size of dataset (10K facts) including at least one additive measure. Moreover, it was necessary to have aggregated facts and at least 4 dimensions including one temporal.

Kimball Lifecycle methodology was used which provides the overall framework that ties together the various activities of a Data Warehouse/Business Intelligence implementation. The overall Kimball Lifecycle approach to DW/BI initiatives is illustrated in Figure 1.

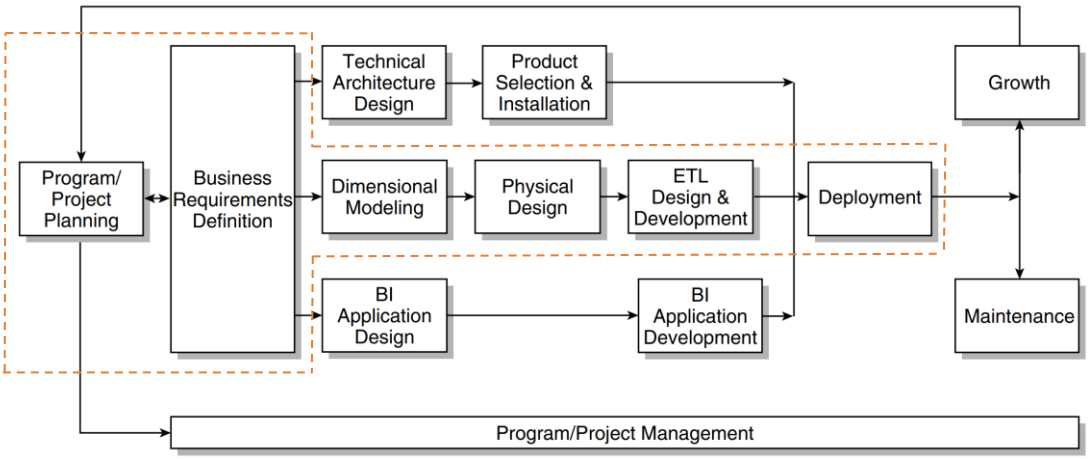


Figure 1. The Kimball Lifecycle diagram [4].The three parallel tracks (infrastructure, data, applications) start at the same time after the business requirements definition. This report focuses on the project planning, business requirements, and the data track, as shown by dashed line in figure.

The document is organized in four sections. In Section 2 (Materials and Methods) we introduce the chosen dataset, the design process of the dimensional model and the methods used to implement the system. Then, Section 3 (Results and Discussion) presents the results obtained by integrating the warehouse with power BI and others. Finally, Section 4 (Conclusions and Future works) concludes this work referring to some future challenges and perspectives in this area.

2. Materials and Methods

The main **business requirements** in the case of this supply chain system project are listed below:

- Monitoring and planning (shipment details, order details, etc.)
- Order processing and inventory management
- Transportation and management
- Supply chain analytics

Having these requirements, we decided to implement the data warehouse for more efficient querying and organization of the data. To achieve this goal, the following tasks were done in the presented order:

1. Acquiring data source
   a. Data understanding and cleaning
2. Design of relational model (Entity relationship)
3. Dimensional model (DM)
   a. Bus matrix definition
   b. Definition of dimensions and fact dictionaries
   c. Design of DM
4. Implementation: Extraction, transformation, and loading (ETL)
5. Data Analytics

Each one of these steps is explained in detail in this section of the report.

2.1 Data Source

(single source project – we don't have info about the stock of products for ex.)

For this project the data was obtained from kaggle website [5]. This dataset was made available by Politécnico de Leiria university, and includes supply chain data obtained from the DataCo Global company. The provided dataset includes 2 CSV files:

1. **Data**: transactions – each record represents one sold item
2. **Metadata**: Description of each of the attributes in the data file

The original data file has 53 attributes, providing detailed information about the customers, shipments, orders, products, and departments. With a size of 91 megabytes (MB), the data file has 180519 transaction records.

Data understanding and cleaning

For this project, one of the requirements being to have a dataset of about 10000 facts, we decided to take a smaller sample of the data. Data from the last 6 months of 2017 (n = 27128) was selected and irrelevant attributes were removed.

Irrelevant attributes included attributes that are repeated (order_benefit and order_profit) attributes that were described incorrectly in the metadata file and their real definition was inconclusive (customer_state, customer_street), and attributes that did not contain information (customer_email, customer_password, product_description).

2.2 Relational Model

The relational model was designed to better understand the connections between the attributes and tables. The model is presented in Appendix 1 – Relational Model.

2.3 Dimensional Model

The dimensional model is a data modeling technique that is widely used in data warehousing to organize and represent data for reporting and analysis. It is a design pattern that is based on two main concepts: dimensions and facts. Dimensions are attributes or characteristics of the data that provide context for analysis, while facts are the numeric values that represent the measures or metrics being analyzed.

a) Dimensional bus matrix

To plan the design of the dimensional model, initially a bus matrix (presented in Table 1) was designed. By using a bus matrix, the process of adding new data sources or modifying existing ones without having to make major changes to the overall system is much easier. It also ensures that the data is integrated in a consistent and accurate manner, which is essential for generating meaningful insights and making informed decisions.

A bus matrix provides a standardized way of integrating disparate data sources. It defines a set of data elements, or dimensions, that are common to all data sources. These dimensions include Time, Location, Product, and Customer.

To design the dimensional bus matrix, 4 main steps were carried out. Initially the importance of data marts was evaluated, and the authors decided that various data marts were unnecessary. At the 2nd step the granularities of the fact tables were defined. In step 3, the dimensions were defined, and finally, the measures for each fact table were defined. The resulting matrix on the four described steps is presented in Table 1.

The bus matrix shows the 2 fact tables (order and order items) of the Datawarehouse and the Sales table, an aggregation used to facilitate the view of some relevant data in the organization.

Table 1. Dimensional bus matrix. In this work, only one data mart was implemented for the whole organization.

			Dimensions			
Stars	**Measures**	**Granularity**	Time	Location	Product	Customer
Order	Number_of_items, total_discount, total_price	1 / customer / time	x	x		x
Order_items	Quantity, Sales, order_item_total	1 / product / time	x		x	
Agg_Sales	Total_sales	1 / time / product / location	x	x	x	

Customer dimension

Customer dimension was one obvious dimension to make to store the information of the customer. The dimension has a level key for the segment of the customer.

Product dimension

Like customer dimension, product dimension is also an obvious one to have. It is important to highlight that the product dimension contains two levels – category and department. The category level is lower than the department level and can include categories such as “cleats”, “women’s”, “Cameras” and “Shop by sport”. The department level is more generic and provides information about the department to which the product belongs, for example “Footwear”, “Apparel”, “Golf” and “Technology”.

Location dimension

A location dimension was also included to store data about the delivery location of the order. Bear in mind that DataCo is an online store business, and its customers can be in every part of the globe, meaning that it is important to register not only the city and state, but also the country, region and market of the order. Hence, for each of these attributes a level key was considered for easier manipulation.

Time dimension

The temporal dimension is used in all the stars. Note that the time dimension has a wide spectrum of granularity. For each of the attributes a level key was generated.

This table could indeed be divided into two, one for minute, hour, day, month, and year, and another table with data regarding analytic attributes such weekday and quarter, as well as day, month, and year. This was avoided as the two tables will have many common attributes and the addition of only minute and hour to the latter table will allow it to serve as a common time dimension for all cases. This simplifies the data model as it reduces the number of tables needed. It also improves the query performance by reducing the number of joins to access data and provides easier analysis of temporal data by providing a unified view of the data. To finalize, this architecture provides a better support for time-based hierarchies, allowing to create hierarchies and support a drill-down analysis.

Order star

In the dimensional model the **Order** fact table is supported by three dimensions (time, location and customer). The objective of the Order star is to have an overview of the orders by date, which has a granularity from minute to year, meaning that with simple queries it will be possible to check for example which month has the higher

number of orders or even what is the day of the week where the company receives more orders. The same line of thought can be made for locations of the order delivery. Overall, this star can improve the understanding of customer behavior by analyzing order time and location, helping to optimize marketing campaigns and offerings.

Regarding the measures, the Order fact table has 3 additive measures – additive measures can be summed across all dimensions. Number of items is the number of items per order. This measure can also be used to relate the number of items per country or city and per date. The total discount is the sum of the all the discounts on the order. The discount is applied by product and not by order, it can also be used to analyze the evolution/effect of the discount in locations and a given date. This measure can be used to compare periods with discount or not and the number of items per order. The third measure is the total price of the order, a sum of the quantity plus the price of each item.

Order item star

Each Order may have different items. Those items can be registered and used for analysis. Hence, the second star uses Order items fact table, and it is considering the order per item, individually, each row in the fact table corresponds to a particular item that was sold in a specific order.

This star has the support of product and time dimensions. By connecting the order items fact table to the product dimension, additional information about each product sold (product name, product category and other product attributes) can be obtained. Individualizing each item in the order allows for detailed analysis. Having the time dimension supporting the order items fact table means that the user can consult which month or day is having more quantities of a certain product sold.

This fact table has three additive measures - Usually, measures that are connected to the dimension *Product* are not additive because product names and descriptions cannot be summed, however in this scenario all these measures are additive because they can be summed across the category attribute or department. These measures can produce interesting information namely *quantity*, *sales*, and *order item total*. The quantity represents the quantity of that item that was bought in an order. *Sales* measure multiplies the quantity of a certain product by its price. Since a discount can be applied to any product, it needs to be reflected in the star, so another measure was created to provide this information. The *order item total* uses the sales and subtract it by the discount.

In sum, the connection of product and time to the order items fact table will provide information about the products sold, its name, category, and department as well as the period which the sales occurred and even the hour.

Sales aggregation

The Sales star is an aggregation. It is represented in the bus matrix and also in the dimensional model to provide a clear view of the design of the Datawarehouse. The idea of having an aggregation is to improve query performance and provide quick access to pre calculated summarized data in the data warehouse. This information should be one of the most requested as it will reduce computational costs. By pre-calculating this summary information, the aggregation table allows for faster querying and reporting of sales data, as the pre-calculated values can be retrieved more quickly than calculating the values on-the-fly. This star is supported by location, product, and time dimensions. Information on the customer segment was included in the aggregation for information purposes, however the connection to customer dimension was not considered since it was concluded not be relevant to consult the customer itself at this level of analysis.

Time is an important dimension because it allows analysis of sales trends over different granularities (daily, weekly, monthly, quarterly, or yearly). By analyzing sales data over time, businesses can identify patterns and trends, such as seasonal fluctuations, and adjust their strategies accordingly.

Product is a critical dimension as it allows the analysis of sales by different products, departments or categories. An analysis of sales data by product will give a better view of product trends, or top-selling products.

The sales aggregation contains one additive measure, total sales. Once more, due to the support of the location, time and product a rollout like monthly sales in a given location of a given product can be executed faster than a normal query. The provided measures will be some of the most requested by manager or C-level directors.

b) Dimensions and facts dictionaries

A data dictionary is an essential tool for managing the vast amounts of data generated by an online sales company like DataCo Global. It provides a comprehensive list of all the data elements used in the company's operations, along with their definitions and relationships to other data elements. With a well-maintained data dictionary, DataCo Global can ensure that everyone working with its data has a clear and consistent understanding of its meaning and usage. This, in turn, can help to improve the accuracy and efficiency of its data management and analysis processes, and support the company's growth and success. The dictionaries of dimensions and fact tables are presented in Table 2.

Note that the dimensions have different types of Slowly Changing Dimension (SCD). The choice of which type of SCD to use depends on the specific requirements of Datawarehouse/Project, the nature of the data being tracked, and the frequency and type of changes expected to occur in the data over time. Nevertheless, different tables have different requirements and consequently different SCDs. Dimensions with data that are relatively static and change infrequently (Time and Location) the SCD Type 1 may be sufficient as it overwrites the old values with new values whenever a change occurs leading to the loss of historical data. The SCD Type 2 involves creating a new record for each change that occurs in the data and preserving the original data and for that reason, the Type 2 is a good option for the Product dimension. The product dimension contains information of the prices that should be preserved. The customer dimension does not have any attribute that should change frequently, at least not so frequently as the prices can change. Maybe the email, location or country can change once in a while and the data should be preserved but not sufficiently as the Product attributes, so the SCD Type 3 may be the right option for this dimension.

There are a few attributes that could be in the same dimension, for example customer and location, however it wouldn’t work as designed, they represent different entities with different attributes and relationships and location is relevant for both. The locations dimension table includes information about each order such as delivery address, on the other hand, the location attributes in customer are the information of the physical location of the customer. Storing the location information in separate tables allows for efficient querying and analysis of location-specific data in the different entities as well as the flexibility in how the date is organized and aggregated.

Table 2. Dictionary of dimensions and fact tables. a) Location dimensions; b) Customer dimension; c) Product dimension; d) Time dimension; e) Order fact table; f) order_item fact table.

a)

Name	Description	SCD		Date:29/03/2023	
Location		Type 1			
Attribute	Description	Level	Key	Type	Size
location_id	location identifier		PK	ID	
city_id	city identifier	City	LK	ID	
city	city name	City	UK	Varchar	32
state_id	state identifier	State	LK	ID	
state	state name	State	UK	Varchar	32
country_id	country identifier	Country	LK	ID	
country	country name	Country	UK	Varchar	32
region_id	region identifier	Region	LK	ID	
region	region name	Region	UK	Varchar	32
market_id	market identifier	Market	LK	ID	
market	market name	Market	UK	Varchar	32

b)

Name	Description	SCD		Date: 29/03/2023	
Customer		Type 3			
Attribute	Description	Level	Key	Type	Size
customer_id	customer identifier		PK	ID	
first_name	customer first name			Varchar	32
last_name	customer last name			Varchar	32
country	customer's country			Varchar	32
city	customer's City			Varchar	32
zip_code	customer's ZIP CODE			Number	
latitude	customer latitude			Number	15
longitude	customer longitude			Number	15
segment_id	segment identifier	Segment	LK	ID	
segment	customer segment	Segment	UK	Varchar	320

c)

Name	Description	SCD		Date: 29/03/2023	
Product		Type 2			
Attribute	Description	Level	Key	Type	Size
product_id	product identifier		PK	ID	
name	product name		LK	Varchar	32
image	product image		LK	Varchar	256
price	product price		UK	Number	8
status	Status of order		LK	Varchar	32
category_id	category identifier	Category	LK	ID	
category_name	product category name	Category	UK	Varchar	32
department_id	department identifier	department	LK	ID	
department_name	product department name	department	UK	Varchar	32

d)

Name	Description	SCD		Date: 29/03/2023	
Time		Type 1			
Attribute	Description	Level	Key	Type	Size
date_id	time identifier		PK	ID	
minute_id	minute identifier	minute	LK	ID	
hour_id	hour identifier	hour	LK	ID	
hour	hour	hour	UK	Number	2
week_day_id	week day identifier	week_day	LK	ID	
month_day_id	month day identifier	month_day	LK	ID	
month_id	month identifier	month_day	LK	ID	
quarter_id	quarter identifier	quarter	LK	ID	
year	year	year	LK	Number	4

e)

Star	Order	Date: 29/03/2023
Granularity	one per customer per order_date (time)	
Dimensions		
customer_id	customer	
order_location_id	order location	
shipping_date	order location	
order_date	order date	
payment_type	method of payment	
days_shipping_real	real # of days of shipping	
days_shipping_schedule	predicted # of days of shipping	
delivery_status	delivery status	
late_delivery_risk	risk	
profir_per_order	profit per order	
order_status	status	
shipping_mode	shipping mode	
Measures		
number_of_items	Number of item in the order	
total_discount	Total order discount	
total_price	Total order price	

f)

Star	Order item	Date: 29/03/2023
Granularity	one per product per order_time_id	
Dimensions		
product_id	product	
order_time_id	order time	
discount	discount	
discount_rate	rate	
product_price	price	
profit_ratio	profit ratio	
Measures		
quantity	quantity	
sales	Amount of sales in $	
order_item_total	total order items	

c) Dimensional Model Design

The final design of the DM is presented in Figure 2.

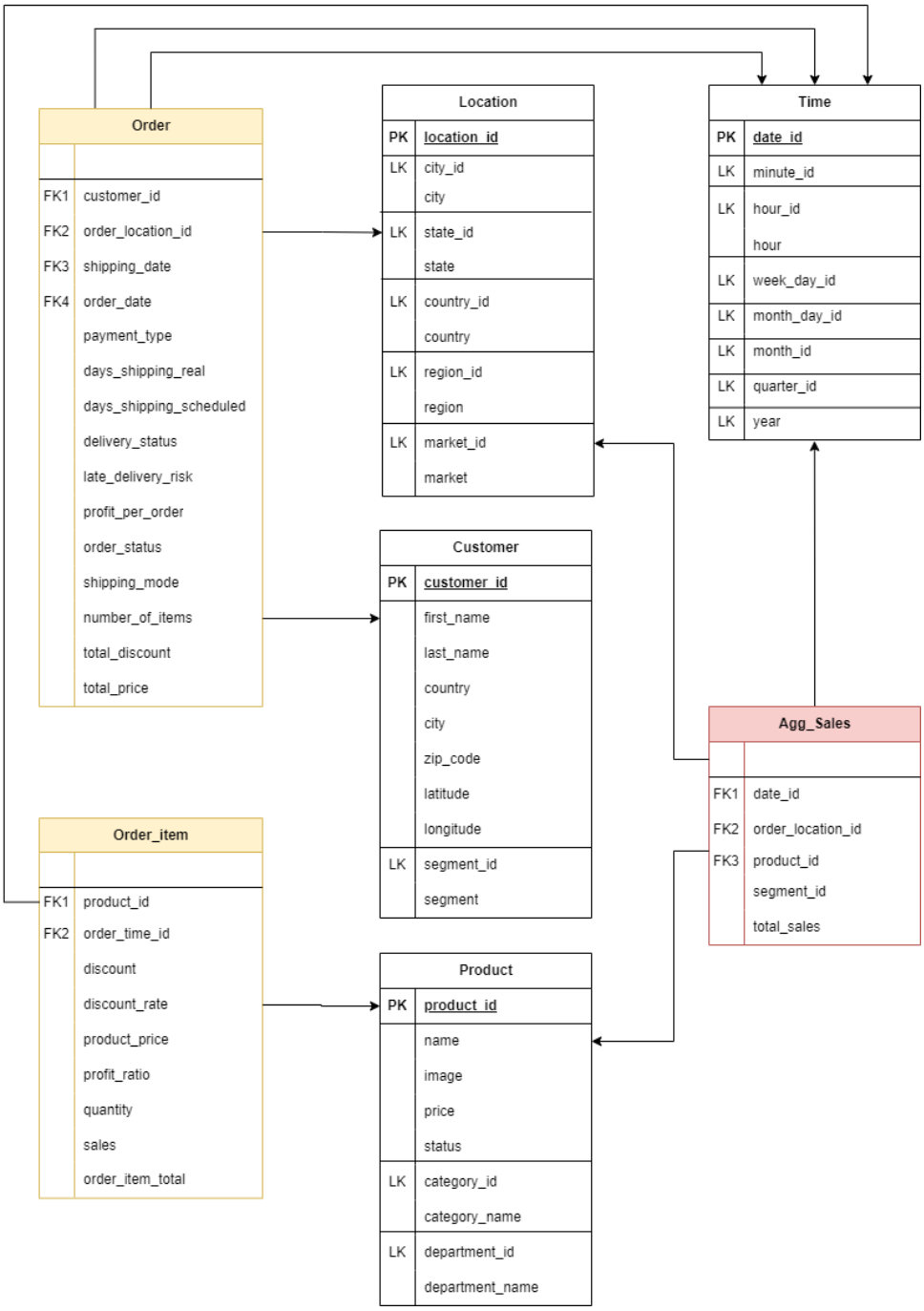


Figure 2. Dimensional model of supply chain data, designed by the authors.

2.4 Implementation: Extraction, Transformation, and Loading (ETL)

To implement the data warehouse, Pentaho from IBM was used. The server used to locally store the data is PostgreSQL. Pentaho Data Integration (PDI) (also called Spoon) allows you to create two basic file types: transformations and jobs. Transformations describe the data flows for ETL such as reading from a source, transforming data and loading it into a target location. Jobs coordinate ETL activities such as defining the flow and dependencies for what order transformations should be run or prepare for execution by checking conditions. Using PDI, the input data was cleansed, formatted, and categorized.

ETL Design and Development

A transformation file was developed for each one of tables (dimensions and fact tables) according to their specific attributes, limitation, and requirements. Starting with the **Product dimension**, the pipeline of the ETL process is demonstrated in Figure 3. The process starts by reading the input file (DataCo supply chain date of the second semester of 2017). After selecting the product attributes from the input file (*Select values*), the rows were sorted alphabetically (*sort rows)*. This step is a requirement for the *Unique rows* to correctly select the unique values and attribute unique ids to them. *Table exists* confirms whether the table exists in the server's database, and if not, it runs the *Execute SQL script* to create the table with the required attributes. Finally, the selected data is loaded to the database (*Table output*).

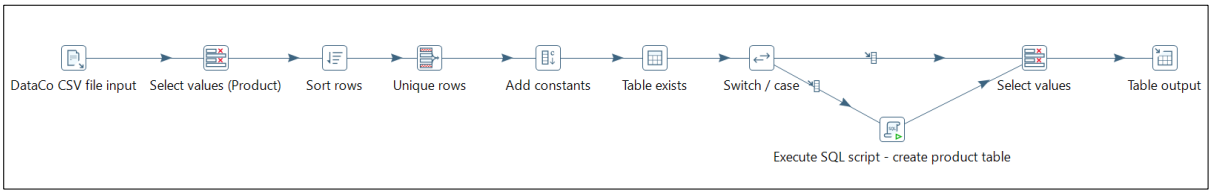


Figure 3. Product dimension ETL process – made in Pentaho Spoon.

For the **Customer dimension**, the pipeline of the ETL is demonstrated in Figure 4. The process is quite similar to that of Product dimension, with the difference that in the customer dimension, segment was chosen as level key, but this attribute does not have an id in the input data file. Therefore, we had to add parallel sequence process which checks the unique values of segment and attributes an id to them. In Figure 4 we can see that the process starts by reading the input file, followed by creating two separate tables, one for segment, and one for the customer information. The upper pipeline generates level keys for the segment, and the lower pipeline selects the unique keys for each unique customer (since the original file already contained customer ids). After unique keys were created for segment, the two tables were joined, with the segment id added using stream lookup.

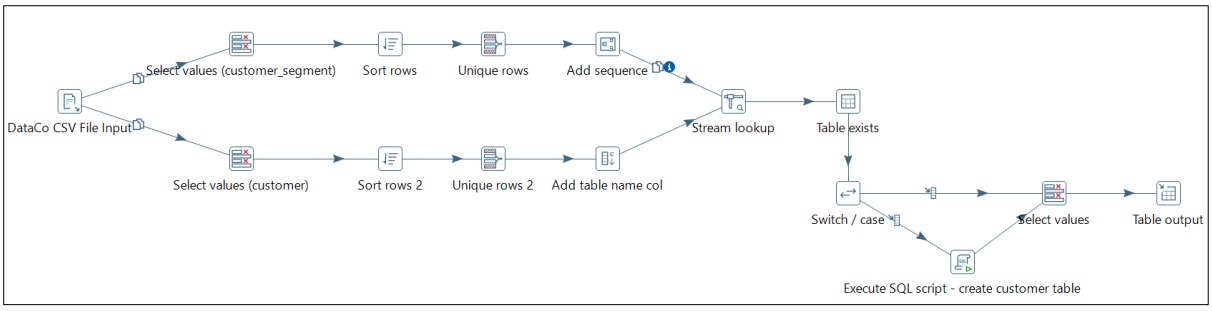


Figure 4. Customer dimension ETL process – made in Pentaho Spoon.

For the **Location dimension**, a similar pipeline was used (Figure 5). Like customer dimension, this dimension also has levels that do not have ids attributed to them, hence, it has to be generated in the ETL process. For each of the attributes city, state, country, region, and market, a level key was generated and attributed to them. Moreover, a unique key for each location was made.

For the **Time dimension** after reading the input file and selecting the correct columns, some additional steps were needed to replicate the designed dimensional model (Figure 2). The input date and time data only provides a timestamp type of data including year, month, day, hour, minute, and second. Besides the mentioned attributes (except second), the dimensional model also includes data regarding the weekday and quarter, besides having the necessity of attributing level keys to some of the attributes. To add level keys the same procedure as the past dimensions was used, and to add weekday and quarter, *Calculator* was used, which automatically generates the desired attributes for a given date.

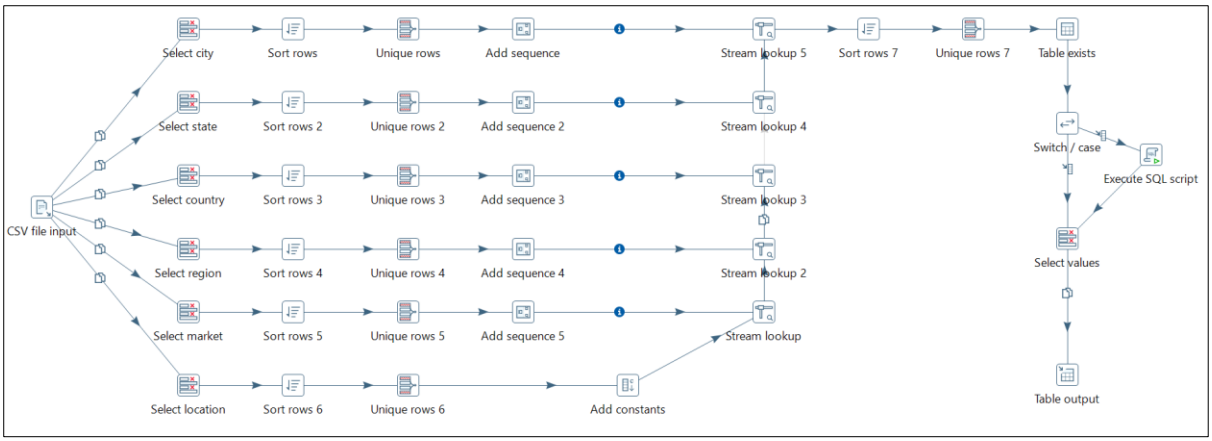


Figure 5. Location dimension ETL process – made in Pentaho Spoon.

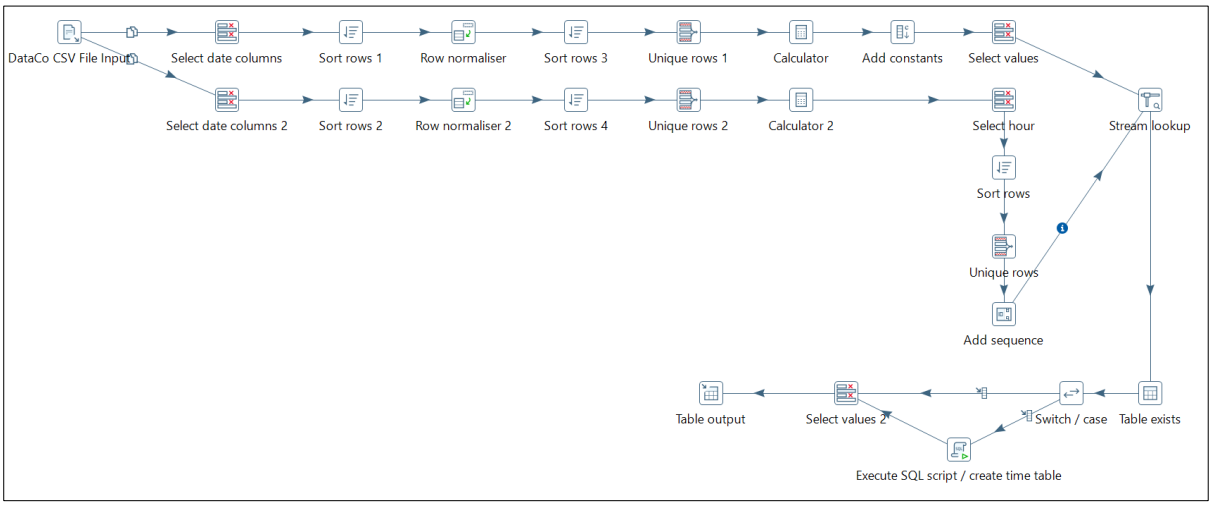


Figure 6. Time dimension ETL process – made in Pentaho Spoon.

After having all the dimensions created and loaded in the database, the fact tables were created. For the **order fact table** several input sources were considered (Figure 7). Besides the data CSV file, which includes information about shipping and other order details, customer, delivery location, shipping time, and order time were considered as foreign keys and were uploaded directly from the dimension tables from the database.

The process starts by importing the original csv and adding the foreign keys from *Customer*, *Location* and *Time* dimensions already created in the previous steps. This creates a raw table with the necessary columns to populate *Order*.

In parallel, the aggregated calculations necessary for the Order table are prepared (since our original csv contains one record per order item). In the bottom stream, the original file from the csv is imported and sorted by order. In the next step, a group by order operation is necessary where the total discount, total price, profit by order and number of items are calculated by summing order item discount, order item total, order profit per order and order item quantity, respectively. With all the necessary calculations performed, the stream lookup operation allows for completing the table with all necessary keys by matching them with the previously parallel prepared table.

The rest of the process is done similarly to that of the other tables explained earlier.

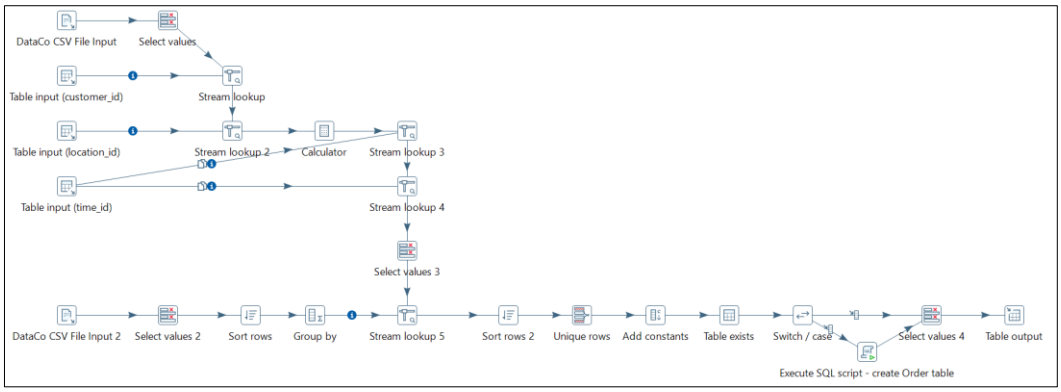


Figure 7. Order fact table ETL process – made in Pentaho Spoon.

For the **order_item fact table**, similarly, several input sources were selected to match the records with its foreign keys for time and product. The data was loaded to the database in a similar procedure as the previous tables, as shown in Figure 8.

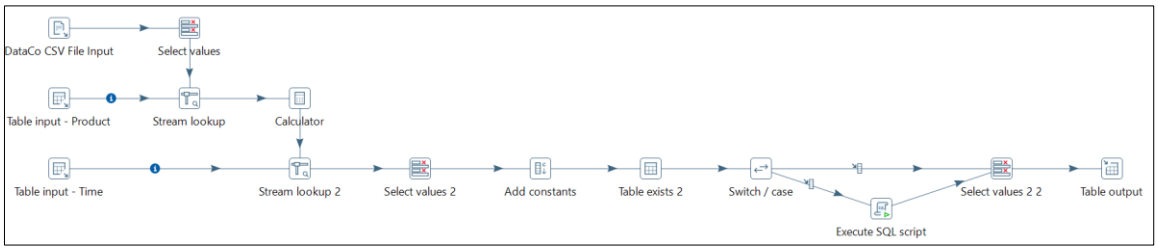


Figure 8. Order item fact table ETL process – made in Pentaho Spoon.

Regarding the **Sales aggregation table**, likewise, various input sources were considered. It is interesting to mention that from the customer table, segment id was read, and not the customer id. This is due to the fact that we do not want to form aggregates for each customer, but for the segments. This can be an interesting insight to provide for the analysts to understand, for example which products are most popular in a specific segment of customers.

After adding the necessary keys from different tables, a sort and group by process was applied. The records are sorted by time id, product id, location id and segment id. They are grouped by the same attributes in order to obtain the sum of sales.

The rest of the process is done similarly to that of the other tables explained earlier.

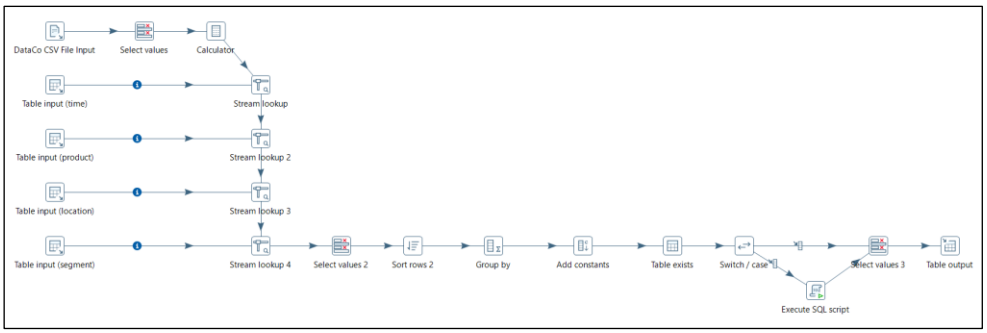


Figure 9. Sales aggregation table ETL process – made in Pentaho Spoon.

2.5 Querying and Data Analysis

For analytics of the data in the implemented warehouse, Power BI platform was used for visualization, and PostgreSQL (pgadmin4) to run queries.

Pentaho Schema Workbench

It is possible to create a Mondrian schema using Pentaho Schema Workbench. A Mondrian schema is a XML file that performs the mapping of a multidimensional model. The multidimensional model contains the logical model consisting of cubes, hierarchies and members and a mapping of this model onto the physical model. The Mondrian schema was developed but was not used in the analytics section, due to lack of appropriate software, as the *Pentaho Analysis* (Saiku) software has been discontinued[1,2].

SQL queries

Several queries were done for each of the stars. Each SQL query and the screen shot of the respective results was saved in a folders. To see the code for queries and their results in detail, in the appended folders visit *Analytics&Queries\Queries* folder.

1. **Order items table**

   For the order items star, three main queries were done (snapshots of the resulting tables are presented in Appendix 2 – Query results - 0Order item):

   - **(Query 16) Category's profit margin**: the profit margin was calculated considering the sales and the order_item_total, and then grouped using a CUBE for the category, and ordered by profit margin, from the highest to the lowest.
   - **(Query 17) Department profits per quarter**: we sum the profits, multiplying the profit margin by the order_item_total to get them, and then a CUBE aggregation function was applied.
   - **(Query 18) Revenue, discount, and profit variations for each category per month**: the AVG aggregation function, was used to calculate the average profit, revenue, and discount per item order item of a certain category and in each month, and then variations are calculated for the same category, considering the average values of the given month and of the previous month. If a given month is the first in the data, then the output gives a zero value. The output was sorted by category and month.

   It is interesting to mention that to remove the null values COALESCE was used in the SELECT section of the query. This was used for the majority of the queries (for example: order_items \ department_per_quarter_profit).

2. **Orders table**

   For the orders star, six queries were done:

   - **(Query 19) average scheduling shipment days:** CUBE was used in the GROUP BY clause to get the average scheduled shipment days per country, and then ordered by those averages, from the highest to the lowest. Part of the output is in Table 6.
   - **(Query 20) average real shipping days**: similar to the first query, however now we use the average real shipping days instead of the average scheduled shipment days. Part of the output is in Table 7.

[1] http://downloads.meteorite.bi/saiku3/saiku-plugin-p7.1-3.90.zip
[2] www.meteorite.bi

- **(Query 21) discount per payment type:** the average order discount was calculated and checked what happened when we grouped that data by payment type, using CUBE. The output is shown in Table 8.
- **(Query 22) payment type per country:** the number of times a certain payment type was used in each country was counted using the COUNT aggregation function, and then sorted by that same metric, from the highest to the lowest value. Part of the output is in Table 9.
- **(Query 23) percentage of completed orders per country, region and market:** the SUM aggregation function was used and the CASE keyword to check the total number of completed orders and the total number of orders made for each market, region, and country. Contrasting with the previous queries, here the ROLLUP function was used instead of the CUBE function, and that was done because the extra lines that the CUBE function gives add no relevant information and knowledge to extract from. Part of the output is in Table 10.
- **(Query 24) total profit per city and per country where the orders were shipped to:** the SUM aggregation function was used to sum the order's profit (here we have the monetary value, contrasting with the previous star where that value is derived) and then we used the ROLLUP function on the country and city, ordering by the profit, from highest to lowest. Part of the output is in Table 11Table 10.

3. Sales aggregation table

Several insights can be obtained using this aggregation sales table. Several queries were performed in order to obtain general understanding of sales data. In Appendix 2 the detailed results can be consulted.

Sales by market, by customer segment and by product:

- **(Query 1)** Grouping the total sales results by market, Europe has the higher value of sales, followed by Pacific Asia and Latin America.
- **(Query 3)** Grouping the total sales results by segment, Consumer has the higher value of sales, followed by Corporate and Home Office.
- **(Query 2)** Grouping total sales results by product, the highest selling product is Field & Stream Sportsman 16 Gun Fire Safe, followed by Dell Laptop and Perfect Fitness Perfect Rip Deck. More details can be found in appendix 2.

Sales by time:

- **(Query 6)** Grouping total sales results by quarter, Q3 has the highest value of sales followed by Q4 and Q2.
- **(Query 5)** Grouping total sales results by month, the top 3 selling months are September, August and July (in highest selling order).
- **(Query 7)** Grouping total sales results by weekday, the top 3 selling days of the week are Saturday, Sunday and Friday (in highest selling order).
- **(Query 4)** Grouping total sales results by hour, the top 3 hours are 5 am, 10 am and 00 am (in highest selling order).

Sales by quarter by market, by region, by customer segment and by product (Query 8-12):

Regarding Q2:

- The highest selling market in Q2 is Latin America, followed by Europe.
- The highest selling customer segment is Consumer followed by Corporate and Home Office.
- The top 3 selling regions are Western Europe (Europe), South America (Latin America) and Central America (Latin America).

Regarding Q3:

- The highest selling market in Q3 is Europe.
- The highest selling customer segment was Consumer followed by Corporate and Home Office (as in Q2).
- The top 3 selling regions were Western Europe (Europe), Northern Europe (Europe) and Southern Europe (Europe).

Regarding Q4:

- The highest selling market in Q4 is Europe followed by Pacific Asia.
- The highest selling customer segment was Consumer followed by Corporate and Home Office (as in Q2 and Q3).
- The top 3 selling regions were Western Europe (Europe), Southeast Asia (Pacific Asia) and Northern Europe (Europe).

Sales of highest selling product per month and per quarter:

- **(Query 13)** The overall highest selling product was Field & Stream Sportsman 16 Gun Fire Safe.
- **(Query 14)** Field & Stream Sportsman 16 Gun Fire Safe has the higher value of its sales in Q3 followed by Q2 and Q4.
- **(Query 15)** The top 3 months of sales of Field & Stream Sportsman 16 Gun Fire Safe are September, August and July (in highest selling order).

Power BI

Power BI is a data visualization tool that allows its users to connect to a database such as PostgreSQL. By connecting PostgreSQL database to Power BI, it can provide significant benefits and powerful data visualization tools to help to understand data and make informed decisions. It is also possible to automate the refresh process and ensure that the share insights with stakeholders are always up to date. Two dashboards were created, an overview for each star.

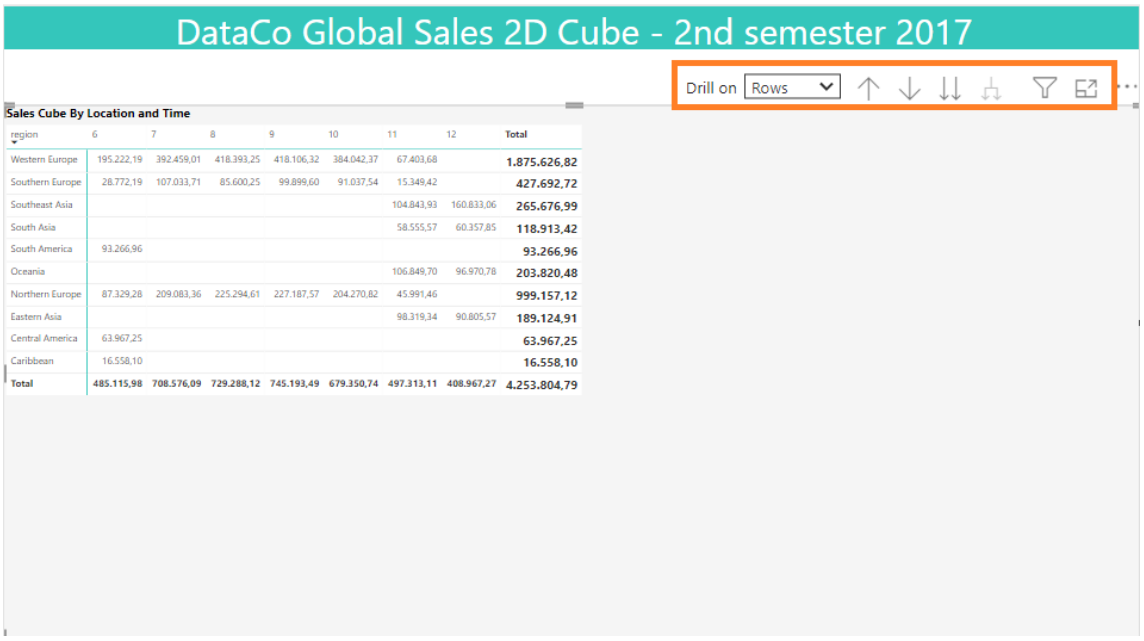
The first dashboard *Orders* (presented in Figure 11) shares the data for the Order star focusing on *quantity, profit and shipping status* and it is even possible to select the timeline, there are 3 slicers for quarter, month and weekday.

The second dashboard is the *Order Items* overview (presented in Figure 12), very similar to the *Orders* dashboard but focusing in *sales, discount and profit ratio* and it is also possible to select the timeline (quarter, month and weekday). The *Order Items* dashboard provides an overview of the above metrics by country and department. There are also the top 7 products sold in selected timeframes, and it can be related to the discount and profit dashboards.

3. Results and Discussion

Regarding the queries, it was clear that the queries in the dimensional architecture, in comparison to the relational architecture, were much more compact. More specially in the aggregation table.

The designed dimensional model and the implemented data warehouse allow the creation of data cubes. However, cubes with more than two dimensions were not developed in this project due to lack of software systems. A data cube for the aggregation table was done with two dimensions (product and time) in PowerBI (presented in *Figure 10*). The file is available in *Analytics&Queries* folder.



region	6	7	8	9	10	11	12	Total
Western Europe	195.222,19	392.459,01	418.393,25	418.106,32	384.042,37	67.403,68		1.875.626,82
Southern Europe	28.772,19	107.033,71	85.600,25	99.899,60	91.037,54	15.349,42		427.692,72
Southeast Asia						104.843,93	160.833,06	265.676,99
South Asia						58.555,57	60.357,85	118.913,42
South America	93.266,96							93.266,96
Oceania						106.849,70	96.970,78	203.820,48
Northern Europe	87.329,28	209.083,36	225.294,61	227.187,57	204.270,82	45.991,46		999.157,12
Eastern Asia						98.319,34	90.805,57	189.124,91
Central America	63.967,25							63.967,25
Caribbean	16.558,10							16.558,10
Total	**485.115,98**	**708.576,09**	**729.288,12**	**745.193,49**	**679.350,74**	**497.313,11**	**408.967,27**	**4.253.804,79**

Figure 10. Two dimensional (product and time) data cube made in PowerBI. Rollup and drilldown can be done using the arrows on top right.

Regarding PowerBI, the implementation was more straight forward in comparison to uploading a CSV file. Having the connection to data warehouse and with the pre-defined inter-tabular connections, the implementation of the dashboards and cubes was much simpler. The dashboards are presented in Figure 11 and Figure 12 for Orders and Order items stars, respectively.

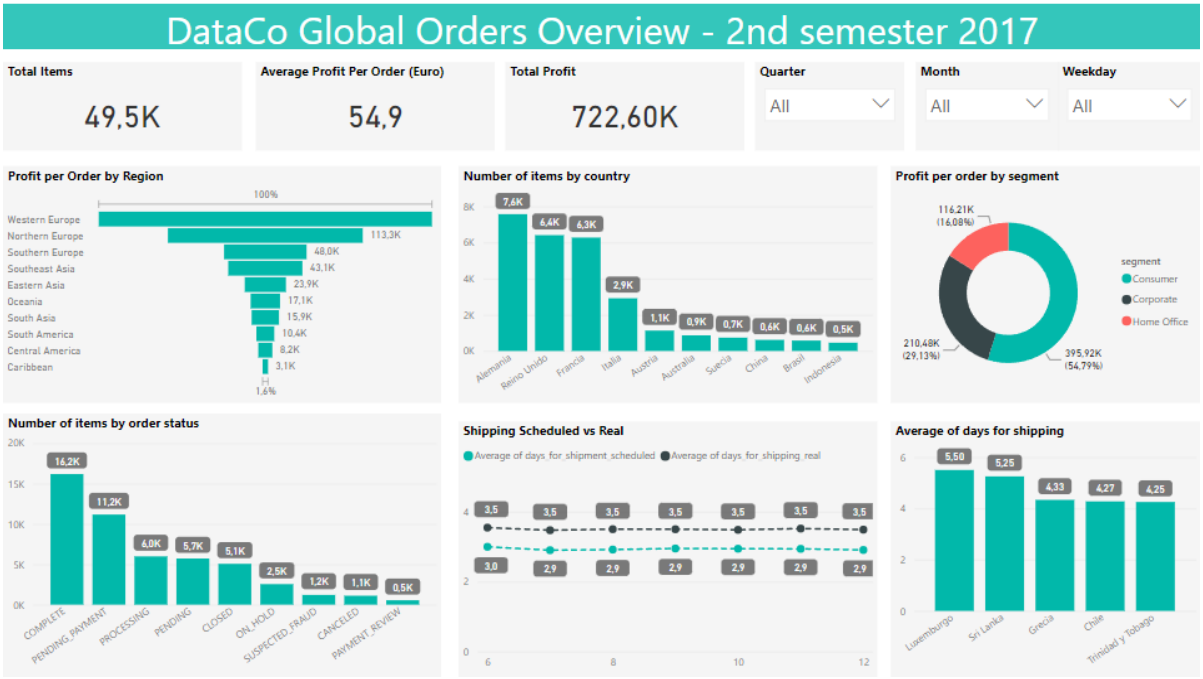


Figure 11 - Orders Overview dashboard from Power BI using data from Data Warehouse

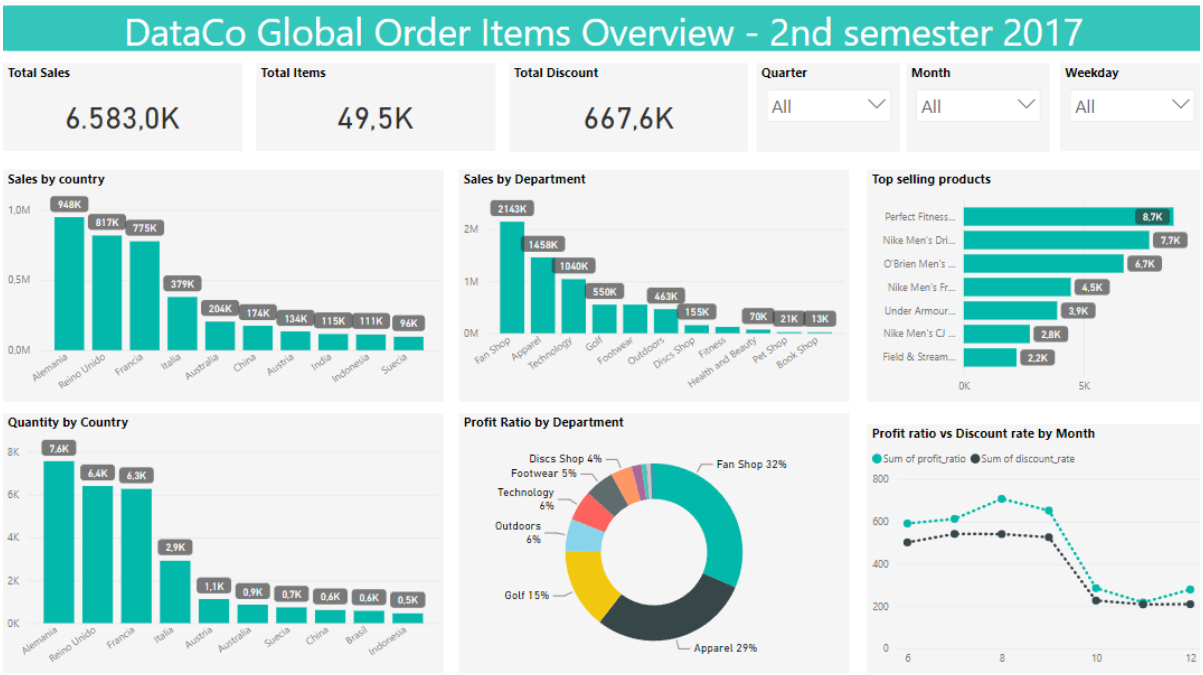


Figure 12 - Order Items Overview dashboard from Power BI using data from Data Warehouse

It was interesting to see that in contrast to what the authors expected, sales in the month of December was not the highest. It was also interesting to see that the real shipping always takes more than the schedules shipping, which imposes that improvements in the prediction of the number of days of shipping is required in DataCo. Several other interesting and important insight in respect to C-level managements can be driven from the developed dashboards.

Overall, it is worth reminding that the transition from relational model to dimensional model, allows much easier scaling, meaning that the model can handle increasing amounts of data by adding more nodes to the system. Relational databases may struggle to handle large volumes of data and may require expensive hardware upgrades to handle increased capacity. Moreover, regarding the data Integration, data warehouses can integrate data from multiple sources and transform it into a format that is suitable for analysis, in a much simpler way.

Data warehouses are optimized for complex analytical queries, making them faster and more efficient for data analysis than relational databases. Additionally, data warehouses use specialized indexing techniques and compression algorithms to optimize performance. Another important advantage of data warehousing is historical analysis. As mentioned, data warehouses can handle large amounts of historical data in a more efficient way, making them well-suited for trend analysis and other types of historical reporting. Relational databases may not be able to handle this type of analysis as effectively due to their focus on transactional processing.

4. Conclusions and Future works

In this project a dimensional model was designed and implemented for supply chain data. The designed pipelines for each of the tables can be used for future loading of new data. A job file can be created in PDI to automatize periodical loads of data from CSV files to the servers. The ETL transformation loaded data to PostgreSQL servers, and posteriorly using PowerBI connection to warehouse, interesting dashboards were made to visualize the data. Moreover, several SQL queries were made to gain different insights about the data.

Comparing the dimensional model to relational model, the DM gave the advantage to have oriented analysis of denormalized data in a systematic and efficient querying system, without losing any important information.

In future works it would be interesting to use the system to feed models and make forecasting and prediction systems to help with the management of the supply chain.

Appendix 1 – Relational Model

The relational model of DataCo supply chain data is presented in Figure 13.

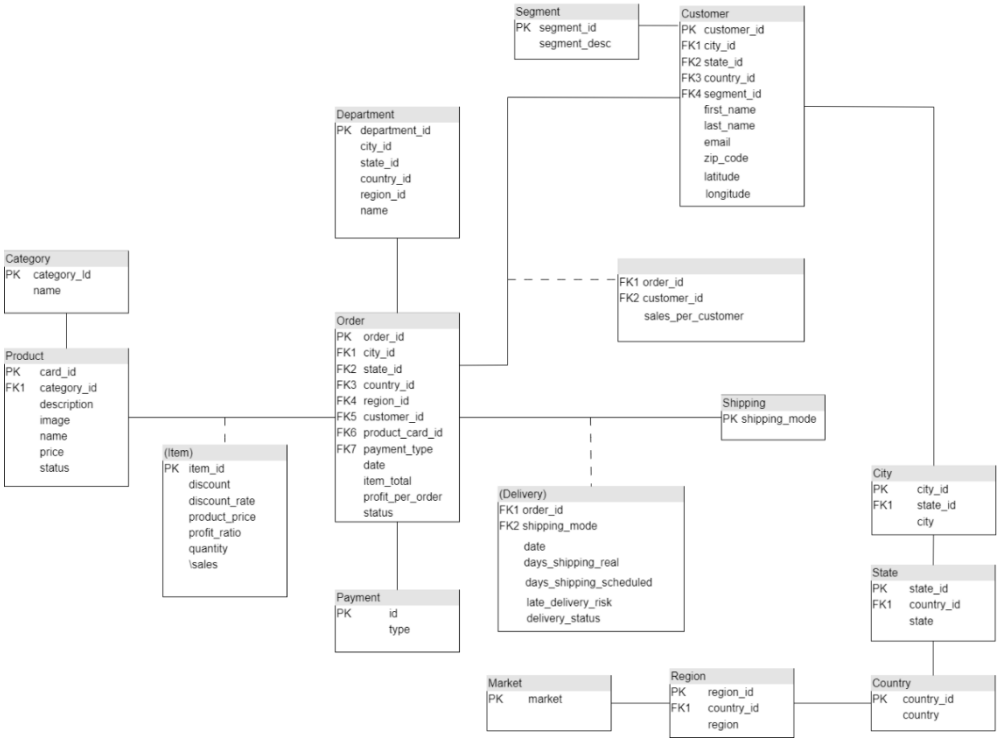


Figure 13. Relational model of supply chain data designed by the authors. The model was designed based on the data and metadate provided in the data source.

Appendix 2 – Query results

Order item table

Table 3 – Query 17 output limited to the top10 results of higher profit value

	quarter character varying	department character varying	department_profit double precision
1	All quarters	All departments	723619.4273238993
2	3	All departments	367520.55465210066
3	4	All departments	245577.78250769968
4	All quarters	Fan Shop	238726.85190660154
5	3	Fan Shop	182755.70515830055
6	All quarters	Apparel	172065.8596032004
7	All quarters	Technology	113359.05231389991
8	4	Technology	113359.05231389991
9	2	All departments	110521.09016410023
10	3	Apparel	84163.66641299984

Table 4 – Query 16 output limited to the top10 results of higher profit margin, from pdAdmin4

	category_name character varying (20)	profit_margin double precision
1	Fitness Accessories	0.13800230233841068
2	Boxing & MMA	0.11915990354887371
3	Strength Training	0.11004831350569078
4	As Seen on TV!	0.10865072858637219
5	Golf Shoes	0.10701338650306745
6	Golf Apparel	0.10609313705339268
7	Hunting & Shooting	0.1060479459993993
8	Lacrosse	0.10488568632435769
9	Women's Golf Clubs	0.10442326874270867
10	Baby	0.10402336633533811

Table 5 – Query 18 output limited to the top10 results ordered by category and month, from pdAdmin4

	month character varying	category character varying	discount_variation_percentage double precision	revenue_variation_percentage double precision	profit_variation_percentage double precision
1	All months	All categories	1.0414870763589297	1.2459499224920159	-16.903338580066873
2	9	All categories	5.401061152827418	5.101066492930887	-5.029103589924409
3	8	All categories	1.0219463124828343	0.6533272412422809	14.279760742570996
4	7	All categories	-0.39943371072827993	-0.379959492837855	-4.898756851492191
5	6	All categories	0	0	0
6	12	All categories	-26.93457974049029	-27.312795441856185	-5.897564178754091
7	11	All categories	-54.81765760862773	-56.26697170552586	-52.6033316813684
8	10	All categories	53.01868872434955	53.79735480520853	53.094959267414296
9	All months	As Seen on TV!	-17.9103178058337	-3.880855751792673	-209.4874852700689
10	9	As Seen on TV!	34.176424668227945	4.027744094291206	51.220586838117676

Order table

Table 6 - Query 19 top10 results ordered by shipment days (query 19)

	country character varying	average_shipping_days_scheduled numeric
1	Chipre	4
2	Luxemburgo	4
3	Ecuador	4
4	Sri Lanka	4
5	Grecia	3.3333333333333335
6	Jamaica	3.3333333333333335
7	Noruega	3.3157894736842106
8	Suecia	3.3144654088050314
9	Irlanda	3.25
10	Argentina	3.1875

Table 7 - Query 20 top 10 results ordered by shipping days

	country character varying	average_shipping_days numeric
1	Luxemburgo	5.5
2	Sri Lanka	5.25
3	Grecia	4.333333333333333
4	Chile	4.272727272727272
5	Trinidad y Tobago	4.25
6	Irlanda	4.05
7	Bolivia	4
8	Honduras	4
9	Chipre	4
10	Noruega	3.8421052631578947

Table 8 - Query 21 output ordered by shipping days

	payment_type character varying	avg_discount_per_order double precision
1	DEBIT	52.3385350581162
2	All payment types	50.75002887258559
3	PAYMENT	50.306628648208644
4	TRANSFER	50.01950925299894
5	CASH	48.02502100981772

Table 9 - Query 22 output limited to the top10 results ordered by number of occurrences, from pdAdmin4

	country character varying	payment_type character varying	number_of_occurrences bigint
1	All countries	All payment_types	8898
2	Alemania	All payment_types	1609
3	Reino Unido	All payment_types	1415
4	Francia	All payment_types	1311
5	Australia	All payment_types	865
6	Italia	All payment_types	652
7	Alemania	DEBIT	641
8	China	All payment_types	614
9	Reino Unido	DEBIT	519
10	Francia	DEBIT	510

Table 10 - Query 23 output limited to the top10 results ordered by completed orders and percentage

	market character varying	region character varying	country character varying	total_orders bigint	completed_orders bigint	percentage_completed_orders numeric
1	All markets	All regions	All countries	8898	2953	33.18723308608676
2	Europe	All regions	All countries	5645	1887	33.42781222320638
3	Europe	Western Europe	All countries	3193	1101	34.48167867209521
4	Pacific Asia	All regions	All countries	2964	975	32.89473684210526
5	Europe	Northern Europe	All countries	1731	564	32.582322357019066
6	Europe	Western Europe	Alemania	1609	554	34.43132380360473
7	Europe	Northern Europe	Reino Unido	1415	460	32.50883392226148
8	Europe	Western Europe	Francia	1311	452	34.477498093058735
9	Pacific Asia	Southeast Asia	All countries	975	307	31.487179487179485
10	Pacific Asia	Oceania	All countries	865	302	34.91329479768786

Table 11 - Query 24 output limited to the top10 results ordered by total profit, from pdAdmin4

	country character varying	city character varying	total_profit double precision
1	All countries	All cities	476901.7410199996
2	Francia	All cities	92861.96006000017
3	Reino Unido	All cities	91545.8104200001
4	Alemania	All cities	81480.67029000013
5	Italia	All cities	39390.40001
6	China	All cities	24324.090059999995
7	Indonesia	All cities	18479.19995999999
8	Reino Unido	London	17111.009929999997
9	Australia	All cities	17078.180049999995
10	Austria	All cities	16297.740099999992

Aggregation sales table

Table 12 - Total sales by market (query 01)

	market_id bigint	market character varying (12)	sales numeric
1	1	Europe	5083655.15
2	3	Pacific Asia	917859.72
3	2	LATAM	581534.42

Table 13 - Total sales by product: top 10 (query 02)

	product_id bigint	name character varying (45)	sales numeric
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	879956.02
2	1351	Dell Laptop	663000
3	365	Perfect Fitness Perfect Rip Deck	524552.57
4	957	Diamondback Women's Serene Classic Comfor...	504866.36
5	191	Nike Men's Free 5.0+ Running Shoe	446155.38
6	1073	Pelican Sunstream 100 Kayak	388980.57
7	502	Nike Men's Dri-FIT Victory Golf Polo	386450
8	403	Nike Men's CJ Elite 2 TD Football Cleat	357992.48
9	1014	O'Brien Men's Neoprene Life Vest	333066.72
10	1349	Web Camera	267607.69

Table 14 - Total sales by segment (query 03)

	segment_id bigint	segment character varying (11)	sales numeric
1	1	Consumer	20447518557.09
2	2	Corporate	6994677990.34
3	3	Home Office	2351210792.22

Table 15 - Total sales by quarter (query 06)

	quarter integer	sales numeric
1	3	3357485.65
2	4	2193477.15
3	2	1032086.49

Table 16 - Total sales by hour (query 04)

	hour_id integer	hour integer	sales numeric
1	6	5	286435.38
2	11	10	286021.18
3	1	0	285833.97
4	13	12	283750.03
5	12	11	282505.61
6	17	16	281179.39
7	5	4	279573.48
8	10	9	276468.61
9	21	20	275598.94
10	14	13	274699.37
11	16	15	274323.97
12	9	8	274033.42
13	2	1	274017.03
14	15	14	272852.02
15	7	6	272156.61
16	4	3	271653.6
17	24	23	270160.12
18	3	2	268247.76
19	23	22	268098.01
20	18	17	267844.01
21	8	7	267531.21
22	19	18	264952.24
23	22	21	264457.82
24	20	19	260655.51

Table 17 - Total sales by month (query 05)

	month integer	sales numeric
1	9	1143775.11
2	8	1109337.17
3	7	1104373.36
4	10	1073994.17
5	6	1032086.49
6	11	626914.38
7	12	492568.6

Table 18 - Total sale by weekday (query 07)

	week_day integer	sales numeric
1	7	1027679.88
2	1	992664.95
3	6	965534.46
4	5	947943.71
5	2	936743.16
6	3	889612.59
7	4	822870.56

Table 19 - Total sales by market by quarter (query 08)

	quarter integer	market_id bigint	market character varying (12)	sales numeric
1	2	2	LATAM	581534.42
2	2	1	Europe	450552.07
3	3	1	Europe	3357485.65
4	4	1	Europe	1275617.43
5	4	3	Pacific Asia	917859.72

Table 20 - Total sales by region by quarter (query 10)

	quarter integer	market_id bigint	region character varying (20)	sales numeric
1	2	1	Western Europe	285904.73
2	2	2	South America	265286.86
3	2	2	Central America	248428.09
4	2	1	Northern Europe	87659.19
5	2	1	Southern Europe	76988.15
6	2	2	Caribbean	67819.48
7	3	1	Western Europe	1955024.97
8	3	1	Northern Europe	710713.27
9	3	1	Southern Europe	691747.4
10	4	1	Western Europe	742288.16
11	4	3	Southeast Asia	295706.19
12	4	1	Northern Europe	276917.28
13	4	1	Southern Europe	256411.99
14	4	3	Eastern Asia	244415.27
15	4	3	Oceania	204081.13
16	4	3	South Asia	173657.11

Table 21 - Total sales by segment by quarter (query 11)

	quarter integer	segment_id bigint	sales numeric
1	2	1	507861.78
2	2	2	319067.66
3	2	3	205157.05
4	3	1	1769964.69
5	3	2	997340.73
6	3	3	590180.23
7	4	1	1159886.81
8	4	2	674108.97
9	4	3	359481.37

Table 22 - Total sales by quarter by product: top 10 (query 09)

	quarter integer	product_id bigint	name character varying (45)	sales numeric
1	2	1004	Field & Stream Sportsman 16 Gun Fire Safe	209589.53
2	2	957	Diamondback Women's Serene Classic Comf…	122991.8
3	2	365	Perfect Fitness Perfect Rip Deck	116020.66
4	2	191	Nike Men's Free 5.0+ Running Shoe	99890.01
5	2	502	Nike Men's Dri-FIT Victory Golf Polo	94450
6	2	1073	Pelican Sunstream 100 Kayak	92195.39
7	2	403	Nike Men's CJ Elite 2 TD Football Cleat	87743.26
8	2	1014	O'Brien Men's Neoprene Life Vest	77568.96
9	2	627	Under Armour Girls' Toddler Spine Surge Runni	36230.94
10	2	724	LIJA Women's Mid-Length Panel Golf Shorts	6400

Table 23 - Highest selling product sales per quarter (query 15)

	quarter integer	product_id bigint	name character varying (45)	sales numeric
1	3	1004	Field & Stream Sportsman 16 Gun Fire Safe	663566.84
2	2	1004	Field & Stream Sportsman 16 Gun Fire Safe	209589.53
3	4	1004	Field & Stream Sportsman 16 Gun Fire Safe	6799.66

Table 24 - Highest selling product sales per month (query 14)

	month integer	product_id bigint	name character varying (45)	sales numeric
1	9	1004	Field & Stream Sportsman 16 Gun Fire Safe	226388.69
2	8	1004	Field & Stream Sportsman 16 Gun Fire Safe	219189.05
3	7	1004	Field & Stream Sportsman 16 Gun Fire Safe	217989.11
4	6	1004	Field & Stream Sportsman 16 Gun Fire Safe	209589.53
5	10	1004	Field & Stream Sportsman 16 Gun Fire Safe	6799.66

5. References

[1] K. LOGISTICS, “KANE LOGISTICS JOINS ID LOGISTICS GROUP,” ID LOGISTICS, 18 09 2013. [Online]. Available: https://www.kanelogistics.com/blog/data-warehouses-and-business-intelligence-drives-logistics. [Acedido em 23 03 2023].

[2] S. a. D. A. a. V. P. Kamble, “Data mining and data warehousing for Supply Chain Management,” *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pp. 1-6, 2015.

[3] D. G. C. G. D. S. N. D. P. a. S. R. Shoumen, “Forecasting and Risk Analysis in Supply Chain Management,” *Managing Supply Chain Risk and Vulnerability*, pp. 187-203, 2009.

[4] M. Ross, W. Thornthwaite, J. Mundy e B. Becker, “Introducing the Kimball Lifecycle,” em *The Data Warehouse Lifecycle Toolkit*, 2nd ed., Wiley Publishing, 2007.

[5] F. Constante, F. Silva e A. Pereira, “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS,” *Mendeley Data*, vol. 5, 2019.