Graphical user interface

Description automatically generated with low confidence

Integração e Processamento Analítico de Informação Project

**Stage 2: Dimensional Modelling**

**Group 1**

Tiago Rodrigues nº49593

Ivo Oliveira nº50301

Martim Silva nº51304

Alexandre Sobreira nº59451

2022/2023

Table of Content

[Introduction 3](#_Toc577098410)

[1. Fix problems detected in the previous stage 3](#_Toc687529371)

[2. Declare the grain and type of the fact table 3](#_Toc890904604)

[3. Model business dimensions, including data hierarchies 4](#_Toc1495710633)

[Facts Table 4](#_Toc695975628)

[Dimension Product 5](#_Toc1637442024)

[Dimension Date – Role-Playing – Order & Ship 5](#_Toc125782272)

[Dimension Holiday (Outrigger - Connected to Dimension Date) 6](#_Toc1728793806)

[Dimension Seller 6](#_Toc284472306)

[Dimension Customer 7](#_Toc621994266)

[Dimension GDP (Outrigger - Connected to Dimension Customer) 7](#_Toc738924459)

[Dimension Order Information 8](#_Toc111288751)

[4.Identify numerical measurements in the fact table(s) - IVO [pp 62:67 Trabalho exemplo] 8](#_Toc230170514)

[5.Draw the data warehouse star diagram 8](#_Toc1009893224)

# Introduction

As we move into this next stage, we have an opportunity to build upon the work done in the previous stage by creating a multidimensional model that is well-suited for a data warehouse. The focus here will be on identifying and addressing any issues that were detected in the previous stage, as well as declaring the grain and type of the fact table(s), modelling business dimensions, and identifying numerical measurements that will be included in the fact table(s). Additionally, we will be creating a data warehouse star diagram that will provide a clear visual representation of our data model. As we undertake these tasks, our goal will be to create a comprehensive, efficient, and effective data warehouse that will serve the needs of our organization.

# 1. Fix problems detected in the previous stage

Irregularities with the original formatting of the data, when taken from each source in stage 1, were treated in this stage. Different representations for data with the same meanings were brought down to a unified standard such as the object type representations of dates for orders and holidays. All columns referencing an ID were changed to hold values in a string format, this was done also for other columns where it would make sense to have their values be strings like City, Product Name, etc. Columns with a high count of missing values that did not contribute in major ways to our business process were promptly discarded. Initially columns with Boolean values in string formats like “Yes” or “No” were to be replaced with actual Boolean True/False values, however, this will be changed to remain in string format but be more descriptive of the character of the data the value is about, as an example we change the values of a column “Returned” from “Yes” or “No” not to True or False but to “was returned” and “was not returned” to be adequately elucidative for subsequent analyses in stage 3. For this stage we also changed the average GDP per state column associations with the states themselves to allow the creation of a primary key that would describe both a state and its average GDP together.

# 2. Declare the grain and type of the fact table

To define the specifics of the star schema to be used for the building process of the data warehouse it is also necessary to clearly define what is going to constitute the contents of each of the tables that follow suit.

Under a star schema representation, our data will be separated in a fact table where a row in this table stores (and details as much as needed for posterior analyses) each recording or fact of the same type of business event that the would-be operational system would perform (the grain) and several dimension tables that are connected using unique identifying values (respective primary keys) to the fact table (holding foreign keys to these dimensional tables) and each of these holds specific information about a part of the business and for that part alone.

Our delineation of what constitutes the grain in this project as we have a single fact table, and the chosen business process of focus is on product profit will then be that of a row that represents a **purchase** of some **quantity** of a **single product** belonging of a type of **category** and **subcategory** that belongs to a **specific order** made on a **specific ordering date** that was made with a **priority** of low, medium or high degree. This purchase was shipped with a **shipping cost** at a specific monetary value amount associated with a **shipping mode** that included first class, second class, standard class or same day options with a specific **shipping date** and a **discount value** on the product ranging between 0 (inclusive) and 1 (exclusive). It was conducted by a **single named customer** belonging to a corporate or consumer **segment** to the superstore from a specific **market** done from a specific **region**, **country, state and city** which was **sold** at a specific monetary value amount and incurred in a **profit** of another specific monetary value amount unless it was a product that belonged to an order made from a **certain region** that was later **returned** to the superstore.

The type of fact table detailed here contains transactions where there is a specific moment in time where each event took place and the dimensions that play a part in defining the grain of the fact table are the Date, Seller, Order, Product and Customer dimensions.

In the world of business intelligence and data analytics, modelling business dimensions is a crucial step in developing an effective data architecture. Business dimensions are the various categories that describe an organization's operations, customers, products, and other key aspects. By modelling these dimensions, businesses can gain insights into their operations and make informed decisions based on data.

One key element of modelling business dimensions is the use of dimensional tables. These tables provide a standardized way to organize data related to a particular business dimension, such as products, customers, or time. Each dimension table typically includes a set of attributes that define the dimension, such as product name, customer ID, or date.

Another important aspect of modelling business dimensions is the creation of data hierarchies. Data hierarchies represent the relationships between different levels of data within a given dimension. For example, a product hierarchy might include categories such as product type, brand, and model. By defining these hierarchies, businesses can analyse data at different levels of granularity, providing deeper insights into their operations.

In addition to modelling business dimensions and creating data hierarchies, another important aspect of designing a data architecture is optimizing the size and performance of fact tables. In some cases, it may be necessary to split a large fact table into multiple smaller ones to improve query efficiency and reduce storage requirements.

The next part of this report will have an analysis of each of the facts and dimensional tables used for the data under study. From these tables, it will be possible to see each of the features inside each fact and dimensional table...

# 3. Model business dimensions, including data hierarchies

According to the charactestics of the data used for this project it can be said that it is Multidimensional for which there are three mains possibilities for storage: ROLAP, MOLAP, HOLAP. In this project a ROLAP system will be used, in which there is normaly a main table, called fact table. This table, the fact table, relates the business measures to the dimensions which have their own tables, called dimenssion tables. This systems will be described above.

## Facts Table

Given the information about the grain referred on the previous point, the Facts Table will have six dimensions: Dimension Product; Dimension Customer; Dimension GDP (Outrigger - Dimension Customer); Dimension Order Information; Dimension Seller; Dimension Date (Role-Playing Dimension for Order Date and Shipment Date); Dimension Holiday (Outrigger – Dimension Date). The Facts Table will also contain the following additive measures: Sales; Quantity; Discount; Profit; ShippingCost. Lastly is important to refer that a degenerate dimension was created (PurchaseKey) which in certain form represents the grain, each purchase that was executed is aligned with the type of the Facts Table (Transaction) where the grain is one row per transaction.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| Purchasekey | Unique Key - Degenerate Dimension | NUMERIC | 1 |
| ProductKey (FK) | Foreign Key | NUMERIC | 1 |
| CustomerKey (FK) | Foreign Key | NUMERIC | 1 |
| OrderKey (FK) | Foreign Key | NUMERIC | 1 |
| OrderDateKey (FK) | Foreign Key | NUMERIC | 1 |
| ShipDateKey (FK) | Foreign Key | NUMERIC | 1 |
| SellerKey (FK) | Foreign Key | NUMERIC | 1 |
| Sales | Sales in $ of a certain purchase | NUMERIC | 82,674 |
| Quantity | Quantity bought of a certain product in a certain purchase | NUMERIC | 2 |
| Discount | Discount aplied to a certain purchase | NUMERIC | 0,7 |
| Profit | Profit obtain in $ from a certain purchase | NUMERIC | -157,086 |
| ShippingCost | Cost in $ of a certain shippement of certain order | NUMERIC | 5,69 |

## Dimension Product

The product dimension table contains valuable information about the products sold by the organization, including product names, categories, and sub-categories. In addition, the primary key for this dimension is the ProductKey, providing a unique identifier for each product sold. These attributes allow businesses to analyze product sales and trends across different categories, brands, and manufacturers.

The product name, category, and sub-category are the data hierarchies of this table. The data hierarchy goes in descending order, category, sub-category, and product name. The category attribute provides high-level information about the product, while the sub-category attribute narrows down to a more specific type of product. Finally, the product name attribute provides the individual name of each product. Using this data hierarchy, businesses can better understand their product sales and identify which categories, sub-categories, and individual products are performing well or poorly.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Description** | **Data type** | **Example** |
| ProductKey (PK) | Primary Key | NUMERIC | 1 |
| ProductID | Unique ID of the Product | NUMERIC | OFF-FA-6129 |
| Product Name | Name of the product | VARCHAR | Staples |
| Category | Category of a given purchased product | VARCHAR | Office Supplies |
| SubCategory | Sub-Category of a given purchased product | VARCHAR | Binders |

## Dimension Date – Role-Playing – Order & Ship

The Dimension Date can be seen as a fundamental dimension of every Datawarehouse, enabling a longitudinal analysis of the business process and therefore uncovering possible trends in the data.

A table with multiple valid relationships between itself and another table is known as a role-playing dimension. This is most commonly seen in dimensions such as Time /Date. The Facts Table has 2 relationships to the Dimension Date on the keys OrderDateKey and ShipDateKey. One possibility to operationalize this could be to have one physical table with all dates and obtain multiple logical tables using synonyms or SQL views given that views allow attribute names to be role-specific.

Another important aspect to refer to in this Dimension is the presence of hierarchies which play an important role in future navigation in the data cube (e.g., drill-dow and roll-up operations) and also to pre-calculate aggregate values for each hierarchical level. Starting from the broader attribute, the Year, followed by the Semester, Quarter, Month, Week of the Month, Day of the Month and Day of the Week. This can be considered a hierarchy of fixed depth given all levels of the hierarchy always have values.

Lastly, to avoid further increase of this dimension size, a Foreign Key with the name of HolidayKey has been created as well as another attribute called HolidayIndicator which informs if the Date of Order or Shipment was a holiday or not. The HolidayKey will be connected to the Dimension Holiday which will be an Outrigger.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| DateKey (PK) | Primary Key | NUMERIC | 1 |
| FullDate | Full Date of a certain Order (year/month/day) | TIMESTAMP | 2012/10/02 |
| Year | Year of a certain Order | NUMERIC | 2012 |
| Semester | Semester of a certain Order | VARCHAR | First |
| Quarter | Quarter of a certain Order | VARCHAR | First |
| Month | Month of a certain Order | NUMERIC | 10 |
| WeekMonth | Week of the Month of o certain Order | NUMERIC | 1 |
| DayMonth | Day of the Month of a certain Order | NUMERIC | 1 |
| DayWeek | Day of the Week of a certain Order | VARCHAR | Sunday |
| HolidayKey (FK) | Foreign Key | NUMERIC | 1 |
| HolidayIndicator | Indication of if it was a Holiday or not | VARCHAR | It is Holiday |

### Dimension Holiday (Outrigger - Connected to Dimension Date)

The Dimension Holiday contains the United States of America Holiday Dates which are attributes that change very infrequently, and Outrigger was created resulting from snowflaking the “monster” Date Dimension. The connection from Dimension Holiday with Dimension Date was made through the HolidayKey.

This dimension contains the full date of the holiday, the name of the holiday and a fixed type hierarchy: Year of the Holiday; Month of the Holiday; Day of the Month of the Holiday; Day of the Week of the Holiday.

It is expected that this table will enable to detection of consumer behaviour patterns related to the occurrence of certain holidays if they exist.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| HolidayKey (PK) | Primary Key | NUMERIC | 1 |
| FullHolidayDate | Full Date of the Holiday (year/month/day) | TIMESTAMP | 04/07/2004 |
| HolidayName | Name of Holiday | VARCHAR | 4th of July |
| YearHoliday | Year of Holdiday | NUMERIC | 2004 |
| MonthHoliday | Month of Holdiday | NUMERIC | 7 |
| DayMonthHoliday | Day of th Month of Holdiday | NUMERIC | 4 |
| DayWeekHoliday | Day of the Week of Holdiday | VARCHAR | Sunday |

## Dimension Seller

The Dimension Seller is a simple dimension with information regarding the worker of the e-commerce chain. Besides the seller name, these dimension attributes relate primarily to location information enabling the creation of a hierarchy: Seller Market; Seller Region; Seller Country, Seller State and Seller City.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| SellerKey (PK) | Primary Key | NUMERIC | 1 |
| SellerName | Name of the Seller | VARCHAR | Marilène Rousseau |
| SellerMarket | Market of the Seller | VARCHAR | Asia Pacific |
| SellerRegion | Region of the Seller | VARCHAR | Caribbean |
| SellerCountry | Country of the Seller | VARCHAR | United States |
| SellerState | State of the Seller | VARCHAR | California |
| SellerCity | City of the Seller | VARCHAR | New York City |

## Dimension Customer

The customer dimension table contains valuable information about the customers of the organization, including customer names, segments, states, regions, and city postal codes. In addition, the primary key for this dimension is the CustomerKey, which provides a unique identifier for each customer in a given location. By analysing these attributes, businesses can gain insights into customer behaviour and preferences, as well as identify opportunities for targeted marketing and customer retention strategies.

For the customer dimension table, the hierarchies are the ones with the segment and customer name. The segment attribute allows businesses to group customers based on their demographic or behavioural characteristics, while the customer’s name attribute provides a unique identifier for each customer. By analyzing these hierarchies, businesses can gain insights into customer behaviour and preferences, identify profitable customer segments, and target their marketing efforts more effectively.

Another important hierarchy for the customer dimension table is the State, Region, City and Postal Code. This hierarchy provides businesses with valuable insights into the geographic distribution of their customer base. By analysing this hierarchy, businesses can identify trends and patterns in customer behaviour across different regions, target their marketing efforts more effectively, and optimize their supply chain and logistics operations. Moreover, businesses can identify areas with high customer concentration, prioritize them for expansion, and allocate resources accordingly.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| CustomerKey (PK) | Primary Key | NUMERIC | 1 |
| CustomerID | Unique ID to identify each customer. | VARCHAR | SV-203651406 |
| CustomerName | Name of the customer | VARCHAR | Muhammed Yedwab |
| Segment | Segment where the customer belongs | VARCHAR | Consumer |
| State | State of residence of the customer | VARCHAR | California |
| StateKey (PK) | Primary Key | NUMERIC | 1 |
| Region | Region of residence of the customer | VARCHAR | West |
| City | City of residence of the customer | VARCHAR | New York City |
| PostalCode | Postal Code of the customer | VARCHAR | 42420 |

### Dimension GDP (Outrigger - Connected to Dimension Customer)

In the data architecture for this business intelligence system, the customer dimension plays a critical role in understanding customer behaviour and preferences. To further enhance the usefulness of this dimension, an outrigger dimension was created that takes into account the Gross Domestic Product (GDP) of the region where the customer lives.

The decision to create this outrigger was based on the fact that whenever the analysis is performed on GDP, it is likely to be associated with the customer dimension. By incorporating GDP data into the customer dimension, businesses can gain deeper insights into customer behaviour and preferences concerning economic factors, such as spending power and purchasing habits.

The primary key for this sub-dimension table is the RegionKey, which provides a unique identifier for each region where the organization's customers reside. Key attributes in this table include the region name and the actual GDP of the region.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| StateKey (FK) | Foreign Key | NUMERIC | 1 |
| CustomerRegion | The region of reference for the GDP | VARCHAR | West |
| AvgStateGDP | AVG GDP value in $ per State between xxxx and xxxx | NUMERIC | 10000000 |

## Dimension Order Information

The Dimension Order Information was created to have a more isolated physical location for information regarding the Orders given the high availability in our original data sources. About the orders, this dimension can present the Order ID, which contains the attribute ReturnedIndicator which informs if a certain order was returned to the seller or not. The Dimension Order Information also contains the attributes ShipMode and OrderPriority which are self-explanatory.

The decision to create this dimension had the goal to promote an easier analysis of the Orders to better extract insights that may be valuable for the business process, for example, analyse if certain orders that contain specific products are recurrently returned or if certain products have, for example, high order priority.

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Desciption** | **Data type** | **Example** |
| OrderKey (PK) | Primary Key | NUMERIC | 1 |
| OrderID | Unique ID to identify each order. | VARCHAR | AE-2012-PO8865138-41184 |
| ReturedIndicator | Indication of if the Order was Returned or not | VARCHAR | Was Returned |
| ShipMode | Shipmment Mode of the Order | VARCHAR | Standard Class |
| OrderPriority | Level of Priority of the Order | VARCHAR | Medium |

# 4.Identify numerical measurements in the fact table(s) - IVO [pp 62:67 Trabalho exemplo]

# 5.Draw the data warehouse star diagram

The most popular data structure (or data modelling technique) used in data warehouses (DWs) is the star schema (SAMSTAR: A Semi-Automated Lexical Method for Generating Star Schemas from an Entity-Relationship Diagram), because of its logical construction of table structures, specifically to facilitate the execution of high-volume and intricate queries commonly referred to as online analytical processing (OLAP) (The Translation of Star Schema into Entity-Relationship Diagrams Michael Krippendorf and Il-Yeol Song).

In the star schema, the data is organized into a central fact table and surrounding it dimension tables (Star Schema Advantages on Data Warehouse: Using Bitmap Index and Partitioned Fact Tables), creating a structure that resembles a star shape. The central facts table typically not only has quantitative measures of the data (for example sales and/or revenue) but also stores the foreign keys which will connect the dimension tables to the centre facts table (Efficient Execution of Joins in a Star Schema). The dimension tables contain the key which is used to connect to the central fact table and also contains descriptive data that provide context to the measures in the fact table (for example time periods, geography and product categories) (An Analysis of Many-to-Many Relationships Between Fact and Dimension Tables in Dimensional Modeling). These attributes are the ones used for a more in-depth or more generalized view of the data, using methods like slice and dice, drill-down and roll-up.

An advantage that the star schema has is that it enables fast and efficient querying and analysis of large datasets (Star Schema Advantages on Data Warehouse: Using Bitmap Index and Partitioned Fact Tables). It also provides a simple and intuitive way of organizing data, making it easier to understand and explain to other people.

For the dataset chosen the star schema designed had the following information:

* Facts Table – containing the measures of the Sales, Quantity, Discount, Profit, ShippingCost and having the foreign keys for the dimension tables.
* Customer Dimension – The dimension includes attributes such as CustomerID, CustomerName, CustomerSegment, CustomerState, CustomerRegion, CustomerCity, and CustomerPostalCode.
* GDP Dimension (Outrigger) – The dimension includes attributes such as RegionKey, CustomerRegion, and RegionGDP.
* Product Dimension – The dimension includes attributes such as ProductID, ProductName, Category, and SubCategory.
* Date Dimension (Role-Playing Dimension) - The dimension includes attributes such as Semester, Quarter, Month, WeekMonth, DayMonth, DayWeek, HolidayIndicator.
* Holiday Dimension (Outrigger) - The dimension includes attributes such as FullHolidayDate, HolidayName, Year, Month, DayMonth, DayWeek.
* Order Information Dimension – The dimension includes attributes such as OrderID, ReturnedIndicator, ShipMode, OrderPriority
* Seller Dimension – The dimension includes attributes such as SellerName, SellerMarket, SellerRegion, SellerCountry, SellerState, SellerCity



