Stage 2: Dimensional Modelling

* In this stage, the elements gathered in the previous report can be improved. However, here the emphasis is on creating a multidimensional model, suitable for a data warehouse. The objectives of this stage are as follows:

# 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 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 and 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 characteristic 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 “wasReturned” and “wasNotReturned” so as to be elucidative upon subsequent analyses in stage 3.

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

In order 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 the 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.

This next part of this report will have an analysis on 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 - Alex e Jimmy [pp 45:57 Trabalho exemplo]

### Facts Table -

* PurchaseKey (Degenerate Dimension)
* ProductKey (FK)
* CustomerKey (FK)
* OrderKey (FK)
* OrderDateKey (FK)
* ShipDateKey (FK)
* SellerKey (FK)
* Sales (Measure)
* Quantity (Measure)
* Profit (Measure)
* Shipping Cost (Measure)
* Discount (Measure)

### Dimension Product – Jimmy

Add: ProductKey (PK)

The product dimension table contains valuable information about the products sold by the organization, including product name, category, and sub-category. In addition, the primary key for this dimension is the product ID, providing a unique identifier for each individual 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 from 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** |
| PK Product ID | Primary Key | Int64 | OFF-FA-6129 |
| Product Name | Name of the product | String | Staples |
| Category | Category of a given purchased product | Category | Office Supplies |
| Sub-Category | Sub-Category of a given purchased product | Category | Binders |

### Dimension Date – Role Play – Order & Ship - Alex

* DateKey (PK)
* FullDate
* Year
* Semestre
* Quarte
* Month
* WeekMonth
* DayMonth
* DayWeek
* IsHoliday

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

* + - * + DateKey (FK)
        + FullHolidayDate
        + HolidayName
        + Year
        + Month
        + DayMonth
        + DayWeek

### Dimension Seller - Alex

* + - SellerKey (PK)
    - SellerName
    - SellerMarket
    - SellerRegion
    - SellerCountry
    - SellerState
    - SellerCity

### Dimension Customer – Jimmy

Add: CustomerKey (PK) & RegionKey(PK)

The customer dimension table contains valuable information about the customers of the organization, including customer names, segments, states, regions, city postal codes. In addition, the primary key for this dimension is the customer ID, which provides a unique identifier for each individual customer. 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 one with the segment and customer name. The segment attribute allows businesses to group customers based on their demographic or behavioral characteristics, while the customer’s name attribute provides the unique identifier for each individual customer. By analyzing these hierarchies, businesses can gain insights into customer behavior 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** | **Description** | **Data type** | **Example** |
| PK Customer ID | Primary Key | Int64 | SV-203651406 |
| Customer Name | Name of the customer | String | Muhammed Yedwab |
| Segment | Segment where the customer belongs | Category | Consumer |
| State | State of residence of the customer | Category | California |
| Region | Region of residence of the customer | Category | West |
| City | City of residence of the customer | Category | New York City |
| Postal Code | Postal Code of the customer | String | 42420 |

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

The decision to create this sub-dimension was based on the fact that whenever an 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 behavior and preferences in relation to economic factors, such as spending power and purchasing habits.

The primary key for this sub-dimension table is the region ID, 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.

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

Add: Region Key (FK)

For this mini dimension

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Description** | **Data type** | **Example** |
| PK Region ID | Primary Key | Int64 | 1 |
| Region | The region of reference for the GDP | String | West |
| GDP | GDP value in $ | Float64 | 10000000 |

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

## 5.Draw the data warehouse star diagram

