Data Mart Implementation (P01)

DECISION SUPPORT SYSTEMS, 2021-22

**João Fernandes (18825), Carlos Martins (18836)**

<< The goal of this project is to implement a data mart based on a transactional data source. The team may use a different data source, but this must be validated by the teacher.

This document has a suitable structure for the project, however chapters can be deleted, changed, or added. Keep formatting consistent throughout the document. **Comments**, such as this one, placed between "<<" and ">>", **should be removed**. >>

# Introduction

<<Describe the scope and general objectives of the project and summarize in 2 or 3 paragraphs the business processes supported by the database >>

The scope of this project is to implement a data mart based on an operation database. In this case the database in use is **Adventure Works 2021 database**. This database is from a fictional USA company, that is both a manufacturer and seller of bicycles/accessories.

The company has two main channels: **Internet Sales** and **Retail Sales**.Since we were given the option of choosing one of these two, we chose **Retail Sales**.

# Data sources

**Person:** This table contains information about the registered people. Each person has information needed, for example: sales, stores or products. The fields in this table are: Address, entity, contact, region, email, password, phone and state.

**Sales:** This table has all the information for the sale of a product. We have: credit card, currency, customer, order detail, the person in charge of the sale, reasons, territory, the shopping cart, the store where the product was bought and special offers.

Table 1: Summary of Person and Sales database contents

|  |  |  |
| --- | --- | --- |
| **Event / object** | **Table** | **Nr. Records** |
| Person | *Person.Address*  *Person.AddressType*  *Person.BusinessEntity*  *Person.BusinessEntityAddress*  *Person.BusinessEntityContact*  *Person.ContactType*  *Person.CountryRegion*  *Person.EmailAddress*  *Person.Password*  *Person.Person*  *Person.Phone*  *Person.PhoneNumberType*  *Person.StateProvince* | 19614  6  20777  19614  909  20  238  19972  19972  19972  19972  3  181 |
| Sales | *Sales.CountryRegionCurrency*  *Sales.CreditCard*  *Sales.Currency*  *Sales.CurrencyRate*  *Sales.Customer*  *Sales.PersonCreditCard*  *Sales.SalesOrderDetail*  *Sales.SalesOrderHeader*  *Sales.SalesOrderHeaderSalesReason*  *Sales.SalesPerson*  *Sales.SalesPersonQuotaHistory*  *Sales.SalesReason*  *Sales.SalesTaxRate*  *Sales.SalesTerritory*  *Sales.SalesTerritorHistory*  *Sales.ShoppingCartItem*  *Sales.SpecialOffer*  *Sales.SpecialOfferProduct*  *Sales.Store* | 109  19118  105  13532  19820  10118  121317  31465  27647  17, 282  163  10  29  10  17  3  16  538  701 |

Through the use of the "*Open-Source* Data Quality and Profiling" tool, we conducted an in-depth study of the tables that we considered relevant for the execution of the Data Mart. With this tool we used the content summarization operation (Summary Data) to obtain the following results:

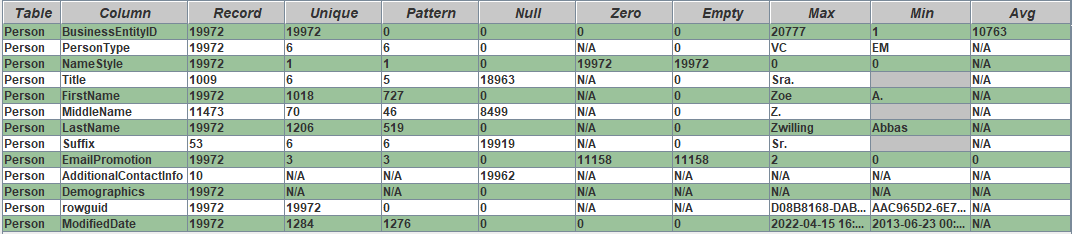
By analyzing the table, 19972 person were registered, and about 42,6% didn’t fill the middle name field, as we can see in the following figure.

Figure 1 - Summary of table Person content.

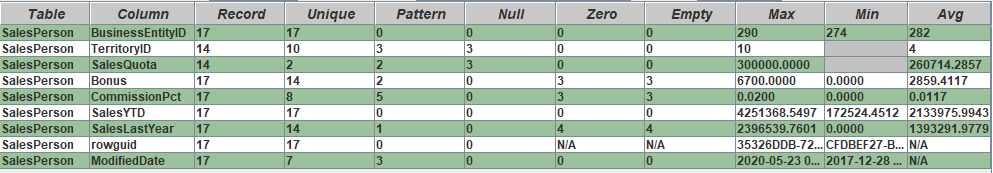
In this next table, only 17 persons are responsible of a sale, this information helps to know how many people are in charge of a sale.

Figure 2 - Summary of table SalesPerson content.

As we can see, a customer is associated with the purchase of a sale, the SalePerson is associated with the sale of that same purchase, and the person is the total of people aggregated into these two fields. This table helps you understand where and in which store the customer bought the product.

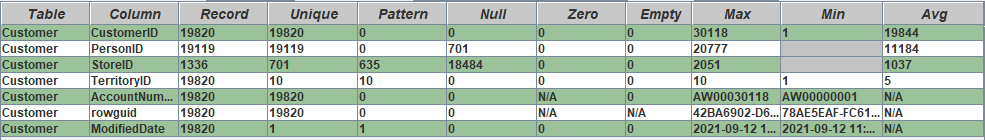


Figure 3 - Summary of table SalesCustomer content.

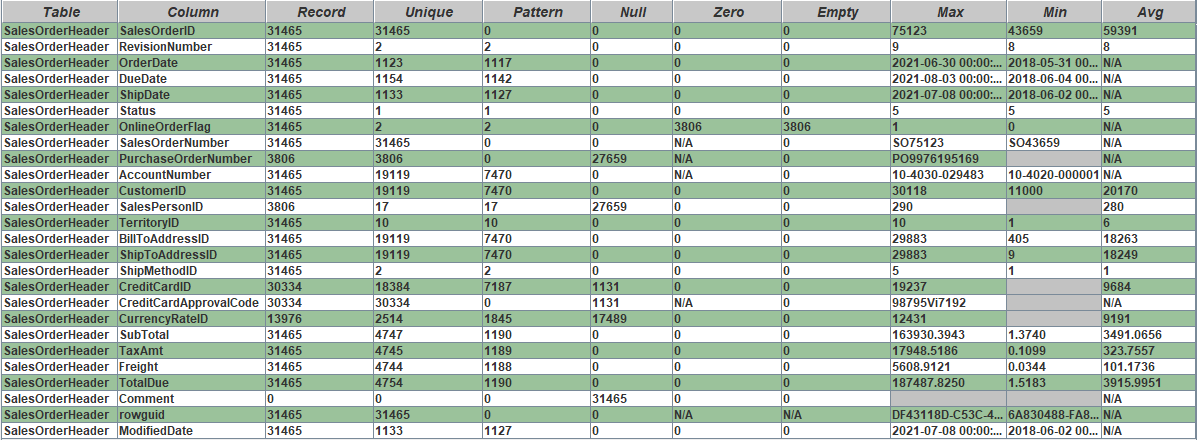
The table Sales Order Header, has a total of 31465 orders, and we can get the information of the client who placed the order, and also see all the information needed to deliver the order and also the price of the product.

Figure 4 - Summary of table SalesOrderHeader content.

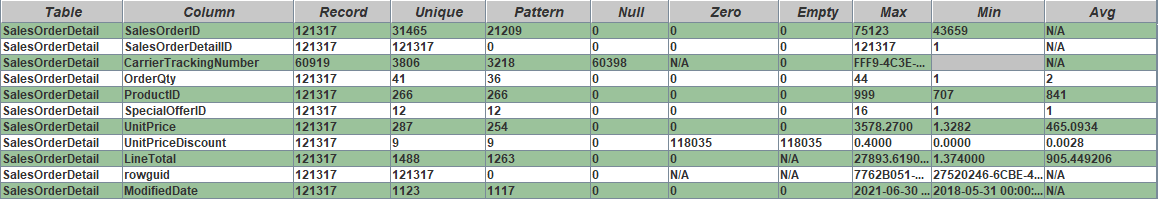
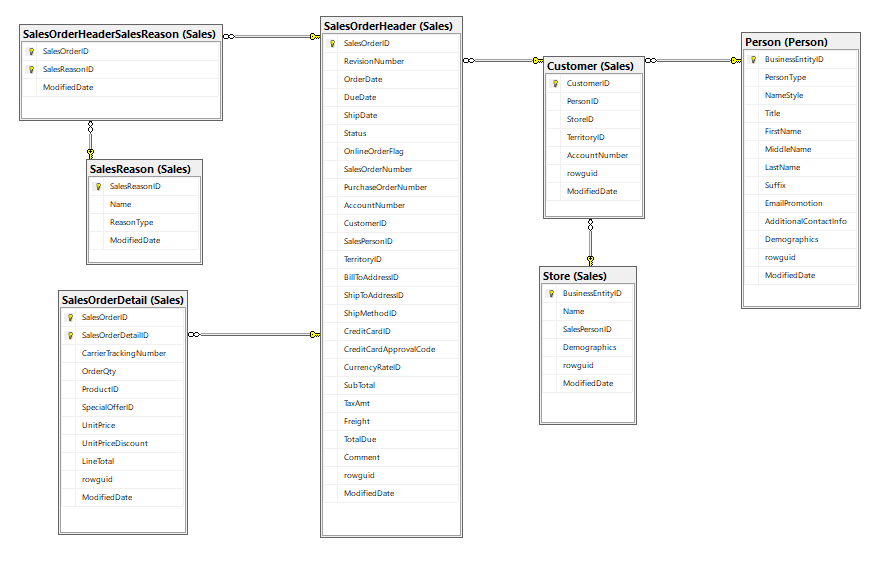
In this table we can see the details needed for the order, the price, the id, the order quantity of the product. There were 121317 orders.

Figure 5 - Summary of table SalesOrderDetail content.



# Dimensional modelling

This database stores multiple information about various business processes like **purchases** and **sales**. Given these processes we defined some questions that the system should be able to answer:

* **Q1** - What was the most profitable month?
* **Q2** – Which store stole the most in a certain location?
* **Q3** - Who is the best salesperson for each product?
* **Q4** - Which store had the biggest sales growth?
* **Q5** - Which products were the most sold per month?
* **Q6** – Who was the client who spend the most money per month?

Table 2: Data Warehouse Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DIMENSIONS**  **BUSINESS PROCESSES** | Customer | Product | Date | Location | Store | Person | Salesperson |  |  |  |  |
| Most profitable month in a year |  |  | X |  | X |  | X |  |  |  |  |
| Store with the most sales in a certain location |  |  |  | X | X |  |  |  |  |  |  |
| Best salesperson for each product |  | X |  |  |  | X | X |  |  |  |  |
| Store with the biggest sales growth |  |  |  | X | X |  |  |  |  |  |  |
| Most sold products per month |  | X | X | X | X |  |  |  |  |  |  |
| Client who spent the most money | X |  | X |  | X | X |  |  |  |  |  |

1. **Fact table(s) and its granularity**

This database is about a fictional USA company, that is both a manufacturer and seller of bicycles and accessories so this is an indicator that the fact tables should be of transactional type. There are two fact tables. The first one is the junction of **SalesOrderHeader** with **SalesOrderDetail** and this FT contains all the information of a sales process. The other one is the junction of **PurchaseOrderHeader** and **PurchaseOrderDetail** and this contains all the information of a purchase process. Both fact tables are of transactional type and each line represents a sales or purchase order respectively.

1. **Dimensions and attributes relevant to the analysis.**

We can use the following dimensions in the analysis:

* Customer:
* Product: Basic product information
* Date: Sales and purchases dates (year, month, day, week)
* Location: Data of certain locations (How much was sold in a certain location)
* Store: Information of a store
* Person: Basic information of business entities. With it we can obtain distributions of various kinds.
* Salesperson: To obtain who was responsible for certain sales.

1. **Measures of the fact table**

Considering the above questions, the system should provide the following measures:

* Number of days to delivey: DueDate – OrderDate
* Total amount of orders -> Count distinct rows by SalesOrderID
* Price without tax: SubTotal + Freight - TaxAmt
* Total order price: SubTotal + Freight + Tax -> TotalDue
* Shipping price -> Freight
* Amount paid by a product: OrderQty \* UnitPrice -> LineTotal

Many interesting metrics can be obtained from these measures, such as:

* Shipping price percentage of total cost: Freight \* 100 / Total
* Tax amount percentage of total cost: TaxAmt \* 100 / Total
* Number of online orders: Number of rows where OnlineOrderFlag == 1
* The number of orders where Card was decline Number of rows where CreditCardApprovalCode == Null

# Design of the dimensional data model

<<Definition of the fact tables (FT), describing the granularity and justifying the measures to be included. In the case of derived measures, it is necessary to present here how they are obtained.

Develop the relational model using a database modelling tool and include the ER diagram in this report. Rationalize the options that have been taken (dimensions, outriggers, DD, etc.). For each table (TF or Dim.), you should complete a data description map (see Appendix A).>>

1. **Description Maps**

* *FactTable*:
* *Dim\_Customer:* It allows us to analyse all customers, obtaining personal information from them;
* *Dim\_Store*: It allows us to see all the stores in the system, as well as their address, contact;
* *Dim\_Location*: Here we can see the location of the sales;
* *Dim\_Product*: It allows us to check all the products. We can see all the details of each product.
* *Dim\_SalesReason:* Here we can see what type of reason and the name of the reason;
* *Dim\_Date*: Date on which changes were made, either to products, customers, etc.

# Data mart implementation

<<Describe the ETL process and highlight the most relevant aspects. Include the graphical representation of the integration transformations and jobs. At the end of this section, write the summary of the data mart content, e.g., number of records loaded into each table.>>

**ETL**

ETL (Extraction, Transformation, and Loading) are software tools whose function is to extract data from various source systems, transform this data according to business rules, and finally load the data, usually into a Data Mart and/or Data Warehouse.

In this project, one of the main objectives is the creation of a Data Mart where the ETL process is performed to work on the data.

**ETL implementation**

In the ETL process, we used an application for building transformations and jobs, which helps handling the data, called PDI Client (Spoon) - Pentaho Data Integration.

We start this process by creating a connection to the database recommended by the teacher, AW2021, and we duplicate it by giving it the name AW2021DM. Next, we perform a share of both connections that we created, in order to make the necessary transformations.

Finally, using the same tool, we create a transformation for each dimension of our relational model.

**Process**

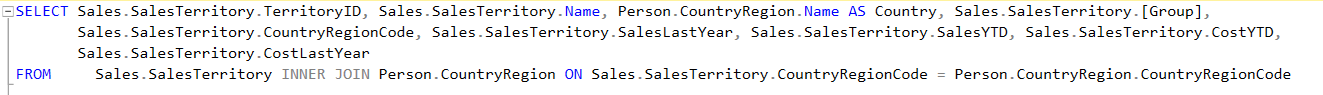
For each transformation was not used any database lookup or select values, because inner joins were performed in the tables as we can see in the following figure.

Figure 6 - Inner joins

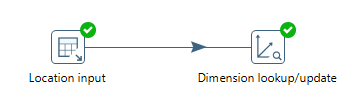


Figure 7 - Dimension Location created in spoon

The transformation **load\_location** was created, in which two tables are linked, the SalesTerritory and CountryRegion.

The customer’s dimension was also created, where two tables are stored, the person and the customer. Here in this dimension, we have all the customer information, from the name and the account number

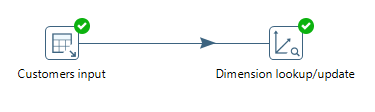


Figure 8 - Dimension Customer created in spoon

We repeat the process for the rest of the dimensions. All the detailed product information can be found in this dimension. We only use the product table.

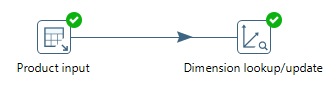


Figure 9 - Dimension Product created in spoon

A mini-dimension of the fact table has been created, as this specifies the reason for a sale, and it is important that it remains outside the fact table to minimise the information.

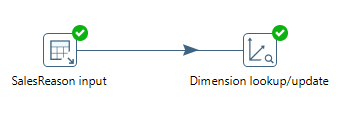


Figure 10 - Mini-Dimension created in spoon

In this dimension, the stores table was used only, where the name, the bussiness ID and the sales person, have the necessary information for the stores.

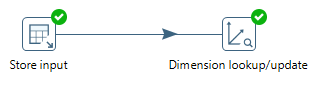


Figure 11 - Store Dimension created in spoon

And lastly, the Date Dimension was created using a file provided by the teacher.

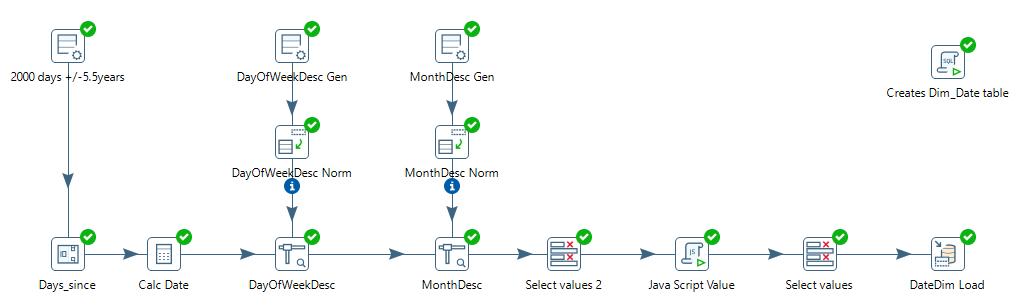


Figure 12 -Data Dimension created in spoon

# Conclusion

<<In the closing chapter, elaborate a critical review of the work done, pointing out its strengths and weaknesses. In addition, if applicable, please list possible future tasks or new options to deepen the work done.>>

# Bibliography

<< In this section, you must present, in APA format, the list of bibliographic sources consulted during the execution of the work and that were relevant for its execution.>>

# Appendix A – Data description maps

<< Display a map/table for each table included in the data mart. See document DSS22\_ETL\_Process >>

Table 3: Data description map of Dim\_A

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Type of table** | **Nr. Records** | | **Description** | | | | |
| Dim\_A | Dimension | ?? | | ?? | | | | |
|  |  |  |  |  |  |  |  |  |
| **Target (Data mart)** | | | | **Source (OLTP)** | | | | |
| **Column** | **Description** | **Data type** | **SCD** | **Table** | **Column** | **Data type** | **ETL rules** | **Example of values** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |