Data Mart Implementation (P01)

DECISION SUPPORT SYSTEMS, 2022-23

**Nuno Mendes (2727), Rosário Silva (21138), Tiago Azevedo (21153)**

# Introduction

O objetivo deste projeto será a implementação de uma data mart para tornar mais eficiente a análise de vendas, assim como a gestão do stock na empresa World Wide Importers (WWI). O data mart será influenciado pela base de dados operacional fornecida pelo docente, permitindo uma análise profunda dos dados, assim como relatórios personalizados sobre vendas, stock, entre outros…

Os processos de negócio que são suportados pela base de dados incluem:

* Gestão de pedidos de clientes;
* Manutenção de Stock;
* Interação com fornecedores;
* Registo de Transações.

Para atender aos objetivos do projeto proposto, serão realizadas 3 etapas:

1. **Data Profiling:** Identificar todas as tabelas relevantes da base de dados, avaliar a qualidade e a consistência dos dados em cada tabela e verificar a presença de dados ausentes/duplicados/incorretos.
2. **Dimensional Modeling:** Determinar os principais objetivos de análise e respetivos relatórios para o data mart, desenvolvimento da DW matrix listando todas as tabelas de factos, assim como dimensões e respetivos atributos e, por fim, criar um modelo ER e mapas de descrição de dados para obter uma representação visual do esquema do data mart.
3. **Extract, Transform, Load:** Desenvolver processos ETL utilizando o Pentaho Data Integration para extrair dados, transformá-los de acordo com as necessidades do data mart e carregá-los nas tabelas de factos e dimensões, documentar essas transformações e, por fim, programar e executar os trabalhos ETL para manter o data mart atualizado.

# Data sources

<<This section describes the **data sources’ structure and content**. Add the relational model of the database and include a content overview of the operational database. You can use a **data profiling** tool to get a quick overview of the data quality. You can use a table (see next) to list the main objects/events.>>

Esta etapa envolve analisar a estrutura e o conteúdo das fontes de dados e avaliar a qualidade dos dados. De seguida poderá ser visualizado o diagrama de Entidade-Relação que representa a base de dados a ser utilizada para o projeto.

Figure 1 - ER Diagram of WWI database

Uma imagem com diagrama

Descrição gerada automaticamente

Table 1: Summary of WWI database contents

|  |  |  |
| --- | --- | --- |
| **Event / object** | **Table** | **Nr. Records** |
|  | *BuyingGroups* | 2 |
|  | *Cities* | 37940 |
|  | *Colors* | 36 |
|  | *Coutries* | 190 |
|  | *CustomerCategories* | 8 |
| Customer | *Customers* | 663 |
| Customer Transaction | *CustumerTransactions* | 97147 |
| Delivery Method | *DeliveryMethods* | 10 |
| Invoice Line | *InvoiceLines* | 228265 |
| Invoice | *Invoices* | 70510 |
| Order Line | *OrderLines* | 231412 |
| Order | *Orders* | 73595 |
|  | *PackageTypes* | 14 |
| Payment Method | *PaymentMethods* | 4 |
| Person | *People* | 1111 |
|  | *SpecialDeals* | 2 |
|  | *StateProvinces* | 53 |
|  | *StockGroups* | 10 |
| Item Stock | *StockItems* | 227 |
| Transaction Type | *TransactionTypes* | 13 |

Nem todas as tabelas têm um evento/objeto associado porque algumas tabelas, como por exemplo “Countries” / ”Cities” / ”Colors” podem não estar associadas a um evento específico, no entanto, podem ser usadas como dimensões de suporte para outras tabelas que contêm eventos/objetos.

# Dimensional modelling

<<Identify the **objectives** and the **questions** that the system should be able to answer, e. g.: Stores with the highest sales growth? Which products topped sales in the last month?

Based on the questions you want to analyze, identify in the **data warehouse matrix** the business processes and dimensions to be used for analysis. Then, for each process, identify the metrics and the corresponding measures to include in the fact table.>>

Table 2: Data Warehouse Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DIMENSIONS**  **BUSINESS PROCESSES** | Dimension 1 | Dimension 2 | Dimension 3 |  |  |  |  |  |  |  |  |
| Process 1 | X |  | X |  |  |  |  |  |  |  |  |
| Process 2 |  | X | X |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

# 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**).>>

# 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.>>

# 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 DSS23\_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** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |