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

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

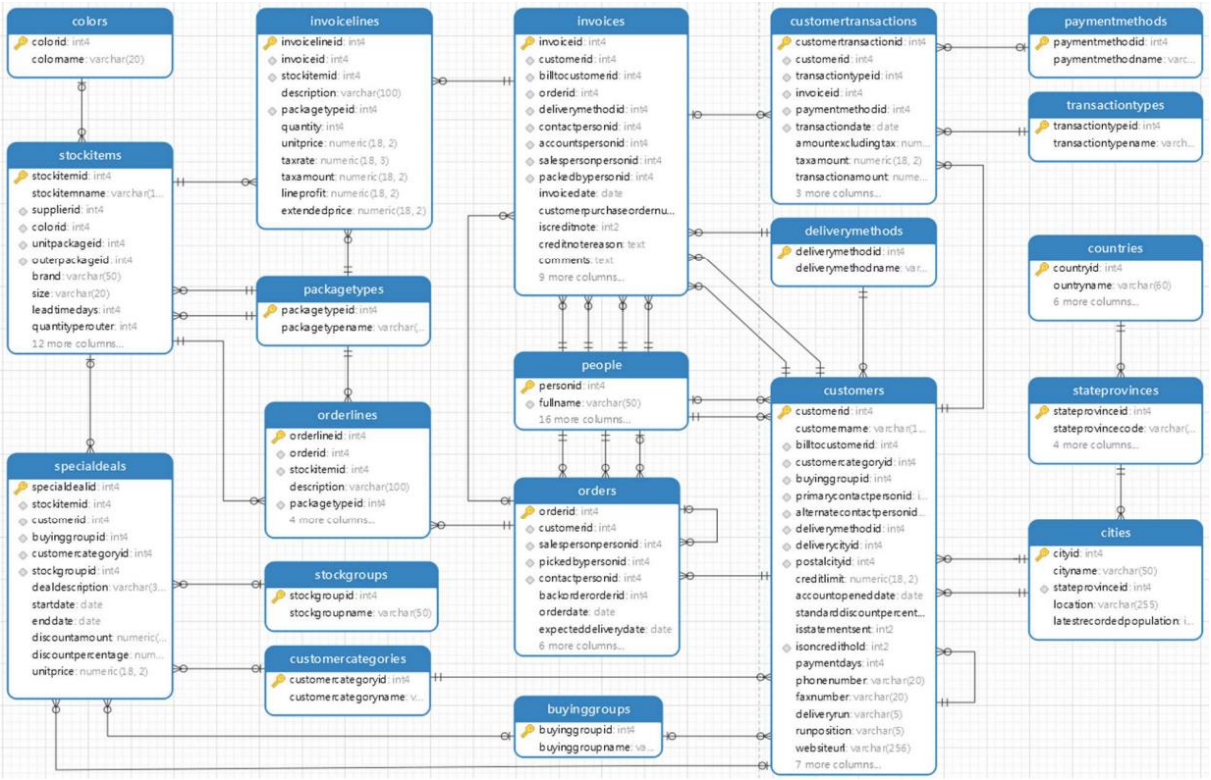

Table 1: Summary of WWI database contents

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

Table 2: Data Warehouse Matrix

DIMENSIONS / **BUSINESS PROCESSES**	Customer	Product	Time	Delivery Method	
Customer Purchase Analysis	X	X	X	-	
Product Sales Growth	-	X	X	-	
Delivery Method Analysis	-	-	X	X	
Backorder Analysis	-	X	X	-	
Order to Delivery Time	-	X	X	X	

1. **Customer Purchase Analysis:**
    - **Métrica:** Valor total das compras
    - **Medida:** Soma do valor dos pedidos de cada cliente no período especificado
2. **Product Sales Growth:**
    - **Métrica:** Crescimento das vendas (%)
    - **Medida:** ((Vendas do produto no período atual - Vendas do produto no período anterior) / Vendas do produto no período anterior) * 100
3. **Delivery Method Analysis:**
    - **Métrica:** Número de entregas por método
    - **Medida:** Contagem de entregas realizadas usando cada método de entrega no período especificado

4. **Backorder Analysis:**
    - **Métrica:** Número de backorders por produto
    - **Medida:** Contagem de backorders de cada produto no período especificado
5. **Order to Delivery Time:**
    - **Métrica:** Tempo médio entre pedido e entrega
    - **Medida:** Média do tempo decorrido entre a data do pedido e a data de entrega para cada produto no período especificado

Design of the dimensional data model

Neste projeto, a granularidade da tabela de factos será a junção de "Cliente", "Produto", "Data" e "Método de Entrega".

Tabela de factos (TF):

- **Chave primária:** FactID (um identificador único para cada registo na tabela de factos)
- **Chaves estrangeiras:** CustomerID, ProductID, DateID, DeliveryMethodID (referenciando as tabelas de dimensões apropriadas)
- **Medidas:** Valor total das compras, Crescimento das vendas, Número de entregas por método, Número de backorders, Tempo médio entre pedido e entrega.

1. **TF #1: Customer_Purchase_Facts**
    - FactID (Chave primária)
    - CustomerID (Chave estrangeira, referenciando Dim_Customer)
    - ProductID (Chave estrangeira, referenciando Dim_Product)
    - DateID (Chave estrangeira, referenciando Dim_Date)
    - TotalPurchaseValue
2. **TF #2: Product_Sales_Growth_Facts**
    - FactID (Chave primária)
    - ProductID (Chave estrangeira, referenciando Dim_Product)
    - DateID (Chave estrangeira, referenciando Dim_Date)
    - SalesGrowth
3. **TF #3: Delivery_Method_Facts**
    - FactID (Chave primária)
    - DeliveryMethodID (Chave estrangeira, referenciando Dim_Delivery_Method)
    - DateID (Chave estrangeira, referenciando Dim_Date)
    - NumberOfDeliveries

4. **TF #4: Backorder_Facts**
    - FactID (Chave primária)
    - ProductID (Chave estrangeira, referenciando Dim_Product)
    - DateID (Chave estrangeira, referenciando Dim_Date)
    - NumberOfBackorders
5. **TF #5: Order_Delivery_Time_Facts**
    - FactID (Chave primária)
    - ProductID (Chave estrangeira, referenciando Dim_Product)
    - DateID (Chave estrangeira, referenciando Dim_Date)
    - DeliveryMethodID (Chave estrangeira, referenciando Dim_Delivery_Method)
    - DeliveryTime

Cada TF e Dimensões terão um mapa de descrição que poderá ser consultado no Apêndice A, no final do documento.

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

Durante o desenvolvimento do projeto, foram executadas várias etapas, incluindo a análise do esquema de dados, a modelagem dimensional e a implementação do processo ETL. O trabalho realizado demonstrou pontos fortes, como a capacidade de identificar as dimensões e factos relevantes, bem como a organização adequada das tabelas e dos relacionamentos.

No entanto, o projeto também apresentou algumas limitações. A qualidade dos dados nem sempre foi ideal, o que exigiu esforços adicionais no processo de ETL. Além disso, nem todos os eventos e objetos foram associados a tabelas, o que pode ter impacto na eficiência da solução proposta.

Para trabalhos futuros, seria interessante aprofundar a análise dos dados disponíveis, explorando outras dimensões e métricas que podem ser relevantes para a WWI.

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

Table 3: Data description map of Customer_Purchase_Facts

Name	Type of table	Nr. Records		Description				
Customer_Purchase_Facts	Fact	??		??				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
FactID	Fact Identifier	INTEGER	-	-	-	INTEGER	-	1, 2, 3, 4, 5, …
CustomerID	Customer Identifier	INTEGER	-	Customers	CustomerID	INTEGER	-	1, 2, 3, 4, 5, …
ProductID	Product Identifier	INTEGER	-	Products	ProductID	INTEGER	-	1, 2, 3, 4, 5, …
DateID	Date Identifier	INTEGER	-	Dates	DateID	INTEGER	-	1, 2, 3, 4, 5, …
TotalPurchaseValue	Total Purchase Value for Order	DECIMAL	-	Orders	TotalPurchaseValue	DECIMAL	-	150.00, 234.56, 89.99, ...

Table 4: Data description map of Product_Sales_Growth_Facts

Name	Type of table	Nr. Records		Description				
Product_Sales_Growth_Facts	Fact	??		Facts about product sales growth				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
FactID	Fact Identifier	INTEGER	-	-	-	-	-	1, 2, 3, 4, 5, …
ProductID	Product Identifier	INTEGER	-	Products	ProductID	INTEGER	-	1, 2, 3, 4, 5, …
DateID	Date Identifier	INTEGER	-	Dates	DateID	INTEGER	-	1, 2, 3, 4, 5, …
SalesGrowth	Sales Growth Percentage	DECIMAL	-	-	-	-	Calculated	12.5, 3.2, -2.8, ...

Table 5: Data description map of Delivery_Method_Facts

Name	Type of table	Nr. Records		Description				
Delivery_Method_Facts	Fact	??		Facts about delivery methods				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
FactID	Fact Identifier	INTEGER	-	-	-	-	-	1, 2, 3, 4, 5, …
DeliveryMethodID	Delivery Method Identifier	INTEGER	-	Delivery Methods	DeliveryMethodID	INTEGER	-	1, 2, 3, 4, 5, …
DateID	Date Identifier	INTEGER	-	Dates	DateID	INTEGER	-	1, 2, 3, 4, 5, …
NumberOfDeliveries	Number of Deliveries per Method	INTEGER	-	Orders	-	-	Calculated	150, 234, 89, ...

Table 6: Data description map of Backorder_Facts

Name	Type of table	Nr. Records		Description				
Backorder_Facts	Fact	??		Facts about backordered products				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
FactID	Fact Identifier	INTEGER	-	-	-	-	-	1, 2, 3, 4, 5, …
ProductID	Product Identifier	INTEGER	-	Products	INTEGER	-	-	1, 2, 3, 4, 5, …
DateID	Date Identifier	INTEGER	-	Dates	INTEGER	-	-	1, 2, 3, 4, 5, …
NumberOfBackorders	Number of Backordered Items	INTEGER	-	Orders	-	-	Calculated	15, 32, 8, ...

Table 7: Data description map of Order_Delivery_Time_Facts

Name	Type of table	Nr. Records		Description				
Order_Delivery_Time_Facts	Fact	??		Facts about order delivery times				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
FactID	Fact Identifier	INTEGER	-	-	-	-	-	1, 2, 3, 4, 5, …
OrderID	Order Identifier	INTEGER	-	Orders	OrderID	INTEGER	-	1, 2, 3, 4, 5, …
DateID	Date Identifier	INTEGER	-	Dates	DateID	INTEGER	-	1, 2, 3, 4, 5, …
DeliveryMethodID	Delivery Method Identifier	INTEGER	-	DeliveryMethods	DeliveryMethodID	INTEGER	-	1, 2, 3, 4, 5, …
DeliveryTime	Delivery Time (in days)	INTEGER	-	-	-	-	Calculated	2, 5, 7, ...

Table 8: Data description map of Dim_Customer

Name	Type of table	Nr. Records		Description				
Dim_Customer	Dimension	??		Customer dimension				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
CustomerID	Customer Identifier	INTEGER	-	Customers	CustomerID	INTEGER	-	1, 2, 3, 4, 5, …
CustomerName	Customer Name	VARCHAR	1	Customers	CustomerName	VARCHAR	-	Tiago Azevedo, Rosário Silva, Nuno Mendes, …
CustomerCategoryID	Customer Category ID	INTEGER	2	Customers	CustomerCategoryID	INTEGER	-	1, 2, 3, 4, 5, …
City	Customer City	VARCHAR	2	Cities	City	VARCHAR	-	Barcelos, Famalicão, …
Country	Customer Country	VARCHAR	2	Countries	Country	VARCHAR	-	Portugal, Espanha, …

Table 9: Data description map of Dim_Product

Name	Type of table	Nr. Records		Description				
Dim_Product	Dimension	??		Product dimension				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
ProductID	Product identifier	INTEGER	-	Products	ProductID	INTEGER	-	1, 2, 3, 4, 5, …
ProductName	Product Name	VARCHAR	1	Products	ProductName	VARCHAR	-	Produto XPTO
ProductCategoryID	Product Category ID	INTEGER	2	Products	ProductCategoryID	INTEGER	-	1, 2, 3, 4, 5, …
Color	Product Color	VARCHAR	1	Colors	ColorName	VARCHAR	-	Branco, Cinza, …

Table 10: Data description map of Dim_Date

Name	Type of table	Nr. Records		Description				
Dim_Date	Dimension	??		Date dimension				
Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
DateID	Date Identifier	INTEGER	-	-	-	-	-	1, 2, 3, 4, 5, …
Date	Date (YYYY-MM-DD)	DATE	-	-	-	-	Generated	2023-04-01, 2023-03-05, ...
DayOfWeek	Day of the Week	VARCHAR	-	-	-	-	Generated	Segunda, Quinta, …
Month	Month	VARCHAR	-	-	-	-	Generated	Março, Abril, …
Year	Year	INTEGER	-	-	-	-	Generated	2023, 2022, …

Table 11: Data description map of Dim_Delivery_Method

Name	**Type of table**	**Nr. Records**	**Description**
Dim_Delivery_Method	Dimension	??	Delivery method dimension

Target (Data mart)				**Source (OLTP)**				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
DeliveryMethodID	Delivery Method Identifier	INTEGER	-	DeliveryMethods	DeliveryMethodID	INTEGER	-	1, 2, 3, 4, 5, …
DeliveryMethodName	Delivery Method Name	VARCHAR	1	DeliveryMethods	DeliveryMethodName	VARCHAR	-	Correio, Pickup, …
DeliveryMethodType	Delivery Method Type	VARCHAR	1	DeliveryMethods	DeliveryMethodType	VARCHAR	-	Registado, Internacional, …