Implementação de Data Mart (P01)

SISTEMAS DE APOIO À DECISÃO, 2021-22

João Ricardo 18845, João Rodrigues 19431, Carlos Santos 19432

Introduction

A realização deste trabalho consiste na implementação de um Data Mart. A base de dados para a elaboração deste trabalho foi fornecida pelo docente da UC, começaremos por fazer uma análise da base de dados e também uma análise mais profunda das tabelas que consideramos ser mais relevantes, consoante os nossos objetivos.

Data sources

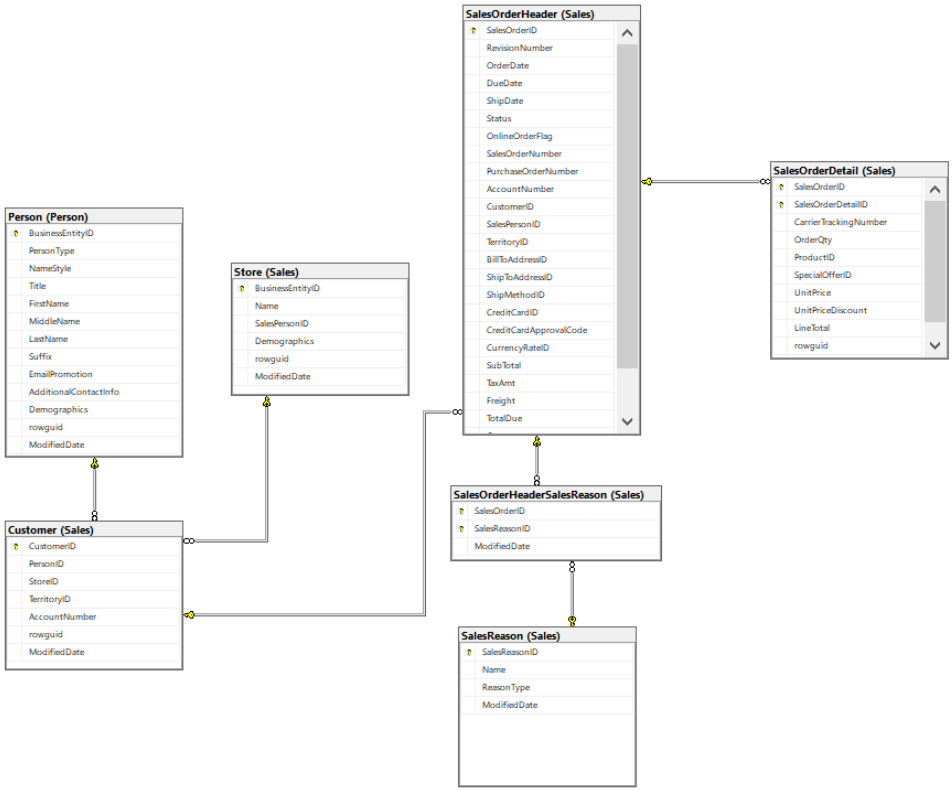
O *Data sourcing* é o processo de análise detalhada de dados de uma base de dados ou de um ficheiro existente. Esta análise permite a compreensão da estrutura, qualidade e conteúdo dos dados e as respetivas relações entre eles. Neste projeto, vamos começar por abordar um pouco do conteúdo e estrutura da base de dados *AW2021*.

Análise do esquema de dados:

Utilizamos um critério de divisão de duas categorias vendas de produtos (Sales) e as vendas, de forma a tornar a leitura dos dados mais clara. Criamos então três categorias (vendas, lojas e clientes) dentro da nossa proposta de modelo relacional. Elegemos apenas estas categorias no modelo pois consideramos serem as mais relevantes para os nossos objetivos.

Evento/Objeto	**Tabela**	**Nr. Registos**
Vendas	SalesOrderDetail	121317
	SalesOrderHeader	31465
	SalesOrderHeaderSalesReason	27647
	SalesReason	10
Lojas	Store	701
Clientes	Customer	19820
	Person	19972

Esta base de dados contém algumas tabelas com informações relativas:



Estudo do contúdo de dados

O primeiro passo para a elaboração deste projeto foi o estudo do conteúdo da base de dados. Dessa forma, vamos em seguida enumerar e explicar as tabelas que constituem a AW2021.

- Person – Registo de uma pessoa. Esta tabela contém informações acerca de uma pessoa.
- Costumer – Faz o registo de um consumidor.
- Store – Registo de uma determinada loja.
- SalesOrderDetail – Detalhes sobre uma encomenda.
- SalesReason – Razão de uma encomenda.
- SalesOrderHeaderSalesReason – Informação relativa das de todas as vendas.
- SalesOrderHeader - Informação relativa à venda de um determinado produto.

Síntense do conteúdo:

Através da utilização da ferramenta "Open Source Data Quality and Profiling" realizamos um estudo aprofundado das tabelas que consideramos relevantes para a execução do Data Mart. Com esta ferramenta utilizamos a operação de sumarização (*Summary Data*) de conteúdo para obter os seguintes resultados:

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
Person	BusinessEntityID	19972	19972	0	0	0	0	20777	1	10763
Person	PersonType	19972	6	6	0	N/A	0	VC	EM	N/A
Person	NameStyle	19972	1	1	0	19972	19972	0	0	N/A
Person	Title	1009	6	5	18963	N/A	0	Sra.		N/A
Person	FirstName	19972	1018	727	0	N/A	0	Zoe	A.	N/A
Person	MiddleName	11473	70	46	8499	N/A	0	Z.		N/A
Person	LastName	19972	1206	519	0	N/A	0	Zwilling	Abbas	N/A
Person	Suffix	53	6	6	19919	N/A	0	Sr.		N/A
Person	EmailPromotion	19972	3	3	0	11158	11158	2	0	0
Person	AdditionalContactInfo	10	N/A	N/A	19962	N/A	N/A	N/A	N/A	N/A
Person	Demographics	19972	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A
Person	rowguid	19972	19972	0	0	N/A	N/A	D08B8168-DAB4-4504-...	AAC965D2-6E72-441E-...	N/A
Person	ModifiedDate	19972	1284	1276	0	0	0	2022-04-15 16:33:33.123	2013-06-23 00:00:00.0	N/A

Com esta tabela conseguimos perceber que o número médio de registos de clientes é de 19972.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
Customer	CustomerID	19820	19820	0	0	0	0	30118	1	19844
Customer	PersonID	19119	19119	0	701	0	0	20777		11184
Customer	StoreID	1336	701	635	18484	0	0	2051		1037
Customer	TerritoryID	19820	10	10	0	0	0	10	1	5
Customer	AccountNumber	19820	19820	0	0	N/A	0	AW00030118	AW00000001	N/A
Customer	rowguid	19820	19820	0	0	N/A	N/A	42BA6902-D6E3-41...	78AE5EAF-FC61-4C...	N/A
Customer	ModifiedDate	19820	1	1	0	0	0	2021-09-12 11:15:0...	2021-09-12 11:15:0...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de clientes é de 19820.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
Store	BusinessEntityID	701	701	0	0	0	0	2051	292	1035
Store	Name	701	699	2	0	N/A	0	Yellow Bicycle Com...	A Bicycle Ass...	N/A
Store	SalesPersonID	701	13	13	0	0	0	290	275	281
Store	Demographics	701	N/A	N/A	0	N/A	N/A	N/A	N/A	N/A
Store	rowguid	701	701	0	0	N/A	N/A	A05E0815-2D28-45...	276C70D2-C4...	N/A
Store	ModifiedDate	701	1	1	0	0	0	2021-09-12 11:15:0...	2021-09-12 1...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de lojas é de 701.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
SalesReason	SalesReasonID	10	10	0	0	0	0	10	1	5
SalesReason	Name	10	10	0	0	N/A	0	Television Advertis...	Demo Event	N/A
SalesReason	ReasonType	10	3	2	0	N/A	0	Promotion	Marketing	N/A
SalesReason	ModifiedDate	10	1	1	0	0	0	2015-04-30 00:00:0...	2015-04-30 0...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de razões é de 10.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
SalesOrderHeaderSalesReason	SalesOrderID	27647	23012	4482	0	0	0	75123	43697	60458
SalesOrderHeaderSalesReason	SalesReasonID	27647	7	7	0	0	0	10	1	2
SalesOrderHeaderSalesReason	ModifiedDate	27647	1108	1073	0	0	0	2021-06-30 00:00:0...	2018-05-31 0...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de clientes é de 27647.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
SalesOrderHeader	SalesOrderID	31465	31465	0	0	0	0	75123	43659	59391
SalesOrderHeader	RevisionNumber	31465	2	2	0	0	0	9	8	8
SalesOrderHeader	OrderDate	31465	1123	1117	0	0	0	2021-06-30 00:00:0...	2018-05-31 0...	N/A
SalesOrderHeader	DueDate	31465	1154	1142	0	0	0	2021-08-03 00:00:0...	2018-06-04 0...	N/A
SalesOrderHeader	ShipDate	31465	1133	1127	0	0	0	2021-07-08 00:00:0...	2018-06-02 0...	N/A
SalesOrderHeader	Status	31465	1	1	0	0	0	5	5	5
SalesOrderHeader	OnlineOrderFlag	31465	2	2	0	3806	3806	1	0	N/A
SalesOrderHeader	SalesOrderNumber	31465	31465	0	0	N/A	0	SO75123	SO43659	N/A
SalesOrderHeader	PurchaseOrderNu...	3806	3806	0	27659	N/A	0	PO9976195169		N/A
SalesOrderHeader	AccountNumber	31465	19119	7470	0	N/A	0	10-4030-029483	10-4020-000...	N/A
SalesOrderHeader	CustomerID	31465	19119	7470	0	0	0	30118	11000	20170
SalesOrderHeader	SalesPersonID	3806	17	17	27659	0	0	290		280
SalesOrderHeader	TerritoryID	31465	10	10	0	0	0	10	1	6
SalesOrderHeader	BillToAddressID	31465	19119	7470	0	0	0	29883	405	18263
SalesOrderHeader	ShipToAddressID	31465	19119	7470	0	0	0	29883	9	18249
SalesOrderHeader	ShipMethodID	31465	2	2	0	0	0	5	1	1
SalesOrderHeader	CreditCardID	30334	18384	7187	1131	0	0	19237		9684
SalesOrderHeader	CreditCardApproval...	30334	30334	0	1131	N/A	0	98795Vi7192		N/A
SalesOrderHeader	CurrencyRateID	13976	2514	1845	17489	0	0	12431		9191
SalesOrderHeader	SubTotal	31465	4747	1190	0	0	0	163930.3943	1.3740	3491.0656
SalesOrderHeader	TaxAmt	31465	4745	1189	0	0	0	17948.5186	0.1099	323.7557
SalesOrderHeader	Freight	31465	4744	1188	0	0	0	5608.9121	0.0344	101.1736
SalesOrderHeader	TotalDue	31465	4754	1190	0	0	0	187487.8250	1.5183	3915.9951
SalesOrderHeader	Comment	0	0	0	31465	0	0			N/A
SalesOrderHeader	rowguid	31465	31465	0	0	N/A	N/A	DF43118D-C53C-47...	6A830488-FA...	N/A
SalesOrderHeader	ModifiedDate	31465	1133	1127	0	0	0	2021-07-08 00:00:0...	2018-06-02 0...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de vendas é de 31465.

Table	Column	Record	Unique	Pattern	Null	Zero	Empty	Max	Min	Avg
SalesOrderDetail	SalesOrderID	121317	31465	21209	0	0	0	75123	43659	N/A
SalesOrderDetail	SalesOrderDetailID	121317	121317	0	0	0	0	121317	1	N/A
SalesOrderDetail	CarrierTrackingNu...	60919	3806	3218	60398	N/A	0	FFF9-4C3E-A9		N/A
SalesOrderDetail	OrderQty	121317	41	36	0	0	0	44	1	2
SalesOrderDetail	ProductID	121317	266	266	0	0	0	999	707	841
SalesOrderDetail	SpecialOfferID	121317	12	12	0	0	0	16	1	1
SalesOrderDetail	UnitPrice	121317	287	254	0	0	0	3578.2700	1.3282	465.0934
SalesOrderDetail	UnitPriceDiscount	121317	9	9	0	118035	118035	0.4000	0.0000	0.0028
SalesOrderDetail	LineTotal	121317	1488	1263	0	0	N/A	27893.619000	1.374000	905.449206
SalesOrderDetail	rowguid	121317	121317	0	0	N/A	N/A	7762B051-2A02-44...	27520246-6C...	N/A

Com esta tabela conseguimos perceber que o número médio de registos de produtos é de 121317.

Dimensional modelling

Identificar objetivos:

- Qual a média de vendas por mês?
- Qual o cliente que mais compras efetuou online?
- Qual o número de vendas realizadas por motivo?

Matriz do armazém de dados

Mostraremos então na tabela abaixo os processos de negócio e as respetivas dimensões que utilizaremos.

Table 1: Data Warehouse Matrix

DIMENSÕES / PROCESSOS DE NEGÓCIOS	Clientes	Motivos	Data
Média de vendas por mês	X		X
Cliente que mais compras efetuou online	X		
Número de vendas realizadas por motivo		X	

Identificar as métricas/medidas

Métricas:

- Média de vendas por mês;
- Cliente que mais compras efetuou online;
- Número de vendas realizado por motivo.

As dimensões que iremos criar são:

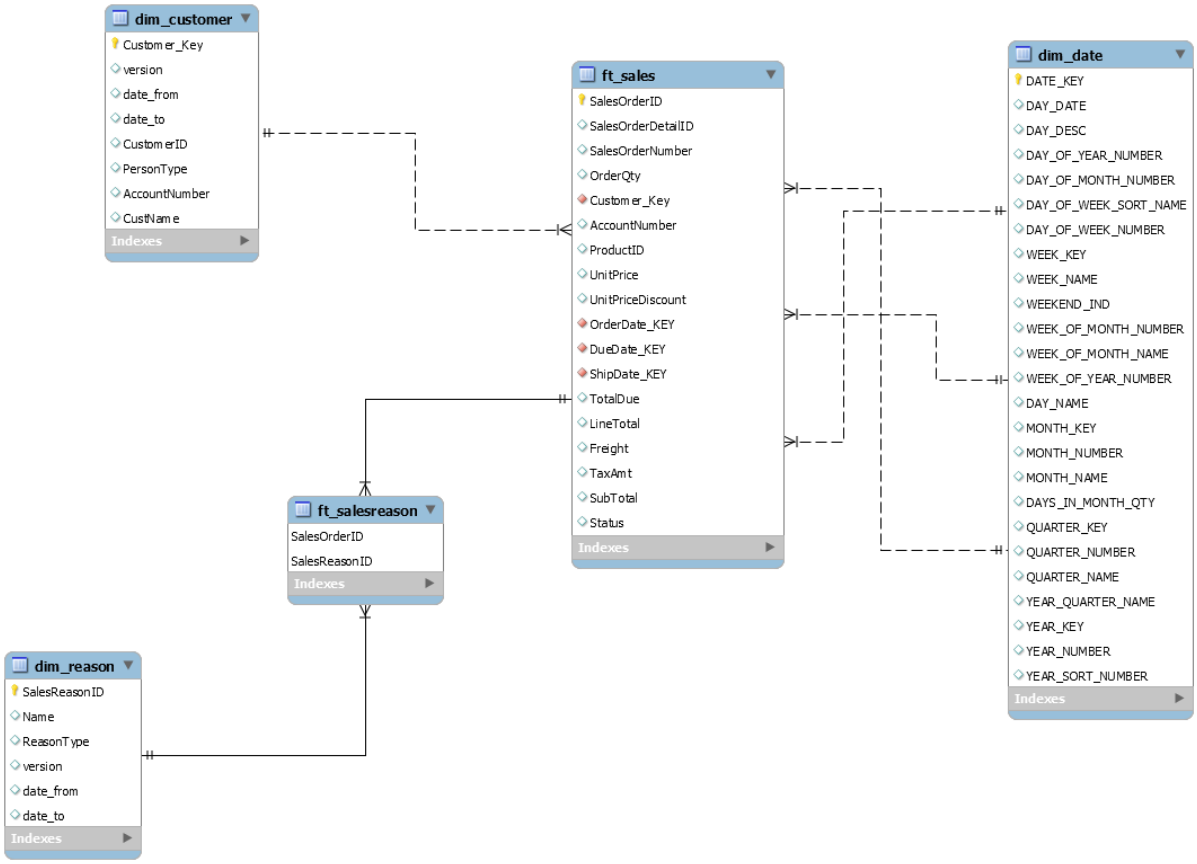
- Dim_Clientes : Permite-nos ver a informação pessoal de cada cliente
- Dim_Motivos : Permite-nos ver o motivo pelo qual foi realizada uma venda.
- Dim_Data : Permite-nos ver em que momento foi realizada uma venda

Design of the dimensional data model

A tabela de factos é a tabela dominante de um Data Warehouse e tem como objetivo estar conectada às demais tabelas, de dimensões. Na tabela de factos são armazenadas as chaves que ligam aos dados das dimensões. Neste caso, a nossa tabela de factos é SalesOrderHeader.
A granularidade da tabela de factos é de um, porque cada linha da tabela corresponde a uma venda, e trata-se de uma tabela transacional.

Medidas	**Classificação**
Média de vendas por mês	Derivada(média = num_vendas/num_mês)
Cliente que mais compras efetuou online	Derivada(count(CustomerId))
Número de vendas realizadas por motivo	Derivada(count(SalesOrderId))

Modelo ER



Mapas de descrição

- Dim_Clientes;
- Dim_Motivos;
- Dim_ListaMotivos;
- Dim_Data.

Data mart implementation

ETL (EXTRACT, TRANSFORM, LOAD)

O ETL (Extração, Transformação e Carregamento) são ferramentas de software cuja função é a extração de dados de diversos sistemas fonte, transformação desses dados conforme as regras de negócios e por fim o carregamento dos dados geralmente para um Data mart e/ou Data Warehouse.

Neste projeto, um dos principais objetivos é a criação de um *Data Mart* onde é realizado o processo de ETL para trabalhar os dados.

Implementação do processo ETL

No processo de ETL utilizamos uma aplicação de construção de transformações, que auxiliam o manuseamento dos dados, denominada PDI *Client* (*Spoon*) – *Pentaho Data Integration*.

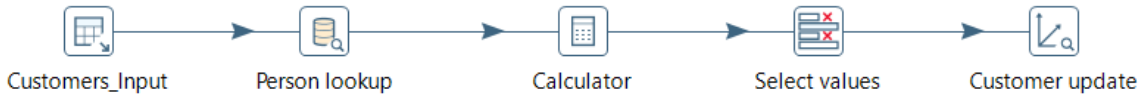
Iniciamos este processo recorrendo à criação de uma conexão com a base de dados recomendada pelo docente, *Sakila*, e efetuamos a duplicação da mesma atribuindo o nome *SakilaDM*. De seguida, realizamos um share de ambas as conexões que criamos, de forma a fazer as transformações necessárias.

Por último, através da mesma ferramenta, criamos uma transformação para cada dimensão do nosso modelo relacional.

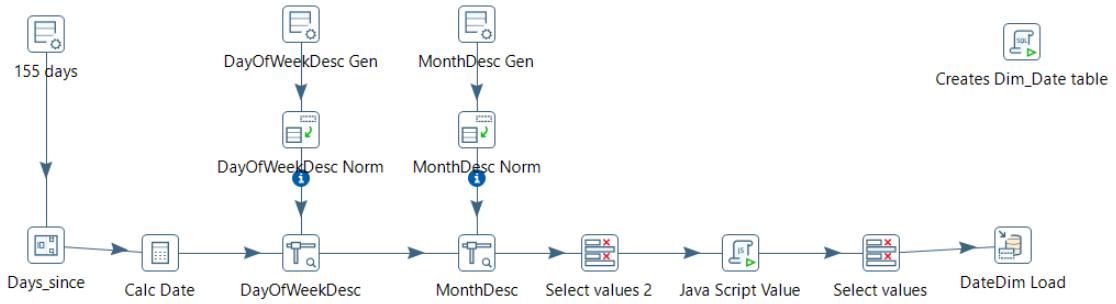
Documentação do processo

A primeira transformação efetuada foi o *costumer Dimension*, em que realizamos a criação de uma dimensão, onde é feita a ligação entre duas tabelas *Person* e *Customer*

através de um *Database Lookup*. Em seguida efetuamos um *Select Values* de forma a remover os parâmetros que não seriam mais necessários.

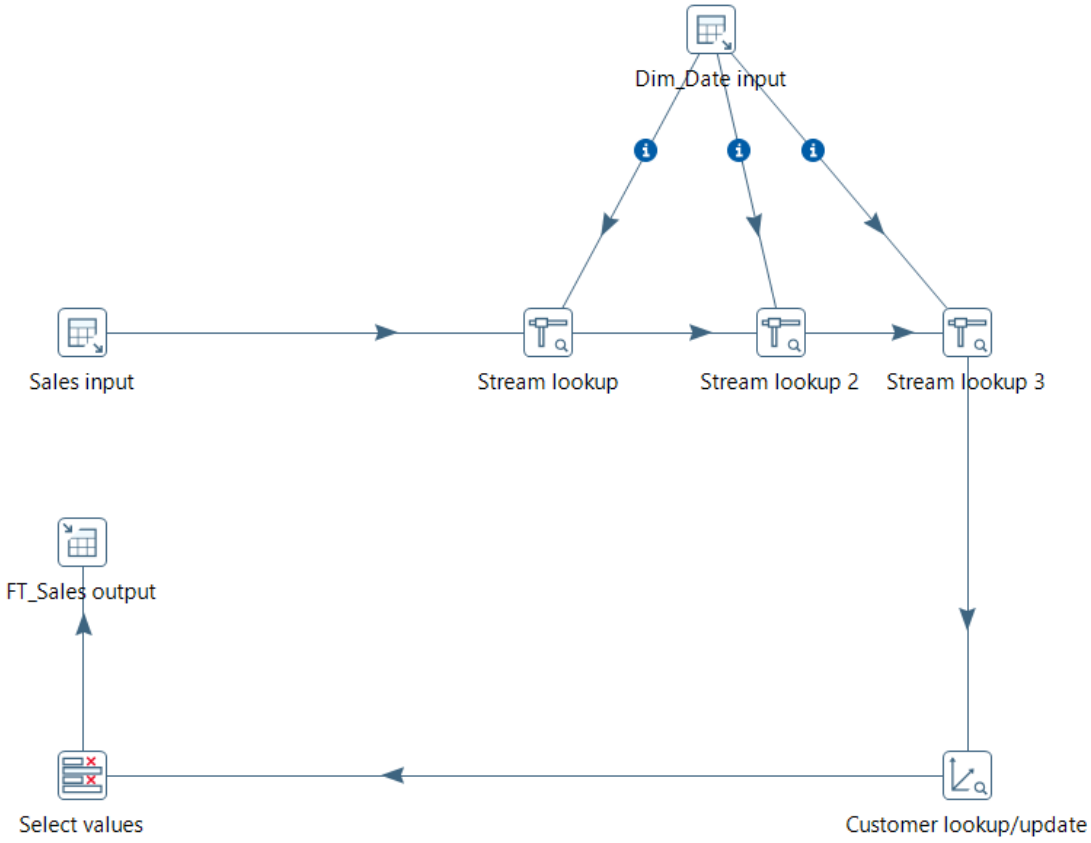


A segunda transformação efetuada foi a Date Dimension onde se recorreu a um ficheiro disponibilizado pelo docente.



A outra transformação efetuada foi a reason *Dimension*, em que realizamos a criação de uma dimensão. Em seguida efetuamos um *Select Values* de forma a remover os parâmetros que não seriam mais necessários.

Para a tabela de factos no table input fez-se um inner join da tabela *SalesOrderHeader* com a tabela SalesOrderDetail. De seguida separamos as datas das horas para comparar as datas com as da dimensão *Date* para retornar as *keys*. Por fim fizemos um *update* á tabela.

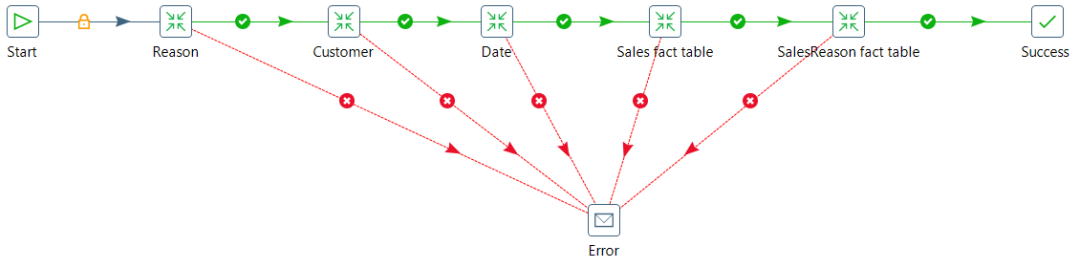


Para a tabela de factos SalesReason apenas fizemos um *update* á tabela.

De forma a agilizar todo o processo de carregamento das dimensões e tabela de factos, procedemos à criação de um *Job* para fazer este processo. Numa primeira fase, foi criado o Job para fazer o carregamento sequencial de todas as tabelas de dimensão e a

tabela de factos (*Sales e SalesReason*). Este *Job* tem como função carregar todo o processo ETL, desde as transformações das tabelas de dimensão à transformação da tabela de factos.

Os *Jobs* simplificam o processo, tornando-o mais compreensível e eficiente. Durante este processo caso exista um erro, esse mesmo problema é comunicado via e-mail.



Conclusion

Com a realização deste trabalho conseguimos obter uma maior aprendizagem no que diz respeito à unidade curricular Sistemas de Apoio à Decisão, bem como a importância de um Data Mart, sendo este bastante útil para aprofundar informação dos dados e juntar os mesmos.

Ficamos a ter mais conhecimentos relativamente ao *SQL Server*, *Pentaho* Data Integration (*Spoon*) e também nos foi possível conhecer uma nova ferramenta, nomeadamente, *Open Source Data Quality and Profiling*, que nos permitiu analisar os dados mais aprofundadamente.

Tivemos algumas dificuldades na implementação do *Data Mart* pois era algo totalmente novo para nós e nunca anteriormente implementado.

Contudo, achamos que a realização deste trabalho será futuramente importante e que foi bem sucedida.

Bibliography

- Power Points fornecidos pelo Docente

Appendix A – Data description maps

Target (Data mart)				Source (OLTP)				
Column	**Description**	**Data type**	**SCD**	**Table**	**Column**	**Data type**	**ETL rules**	**Example of values**
SalesOrderNumber	Identificador da venda	String	0	ft_sales	SalesOrderNumber	String		SO43697
OrderQty	Quantidade de um determinado produto	Int	1	ft_sales	OrderQty	Int		1
AccountNumber	Numero da conta	String	0	ft_sales	AccountNumber	String		10-4030-021768
UnitPrice	Preço por unidade	Decimal	1	ft_sales	UnitPrice	Decimal		3578.27
UnitPriceDiscount	Preço de desconto	Decimal	1	ft_sales	UnitPriceDiscount	Decimal		0.0
TotalDue	Total da compra	Decimal	1	ft_sales	TotalDue	Decimal		3953.9884
LineTotal	Preço por cada linha	Decimal	1	ft_sales	LineTotal	Decimal		3578.27
Freight	Preço de transporte	Decimal	1	ft_sales	Freight	Decimal		89.4568
TaxAmt	Taxa	Decimal	1	ft_sales	TaxAmt	Decimal		286.2616
Status	Estado da venda	Int	1	ft_sales	Status	Int		5

Nome	Tipo de tabela	Nº Registos	Descrição
dim_customer	Dimensão	18485	Permite-nos analisar o cliente.
dim_reason	Dimensão	11	Permite-nos saber a/as razão/ões de uma venda.
dim_date	Dimensão	155	Momento em que o produto é comprado (ano, mês, dia).
ft_sales	Tabela de factos	60398	Permite-nos analisar uma venda.
ft_salesreason	Tabela de factos	27647	Permite-nos ver analisar as razões de uma determinada venda