



Universidade do Minho

Mestrado Integrado em Engenharia Informática

Mestrado em Engenharia Informática

Unidade Curricular de Data Warehousing

Ano Lectivo de 2018/2019

Belo Mundo

Diana Costa A78985

José Oliveira A78806

Luís Costa A74819

Sérgio Jorge A77730

Vítor Castro A77870

Janeiro de 2019

DW

Data de Recepção	
Responsável	
Avaliação	
Observações	

Belo Mundo

Diana Costa A78985

José Oliveira A78806

Luís Costa A74819

Sérgio Jorge A77730

Vítor Castro A77870

Janeiro de 2019

Resumo

Este relatório descreve as diferentes etapas do projeto de *Data Warehousing*, que surgiu da necessidade de implementar um sistema de suporte à decisão que apoiasse a agência de viagens “Belo Mundo”.

A primeira fase consiste na fundamentação e contextualização do sistema, e é onde é feita a definição da identidade do sistema, bem como definição de medidas de sucesso e identificação de recursos. Ou seja, será por isso descrito o enquadramento e o modelo de funcionamento, bem como as principais razões que levaram a esta implementação e objetivos a alcançar na sua conclusão. É feito também um plano de desenvolvimento.

De seguida, definiram-se os requisitos de controlo, de exploração e de descrição e, com estes, partiu-se para a modelação dimensional que, assente no método de 4 passos, orientou e ajudou o grupo a encontrar grão, dimensões e medidas. Definiu-se, assim, um modelo dimensional concetual e um modelo dimensional lógico. Por fim, caracterizaram-se as fontes de informação.

Finalmente, procedeu-se à especificação e implementação da área de retenção e do *Data Warehouse* propriamente dito, com recurso ao *MySQL* e ao *Pentaho*, seguidos de obtenção e discussão de resultados a partir da ferramenta *PowerBI*.

Área de Aplicação: Desenvolvimento de um sistema de Data Warehouse;

Palavras-chave: Data Warehouse, Fontes de Informação, Pentaho, Vendas.

Índice

Resumo	i
Índice	ii
Índice de Figuras	iv
2.1. Definição da Identidade do Projeto	5
2.2. Identificação de Recursos	5
2.3. Estabelecimento do Plano de Desenvolvimento	6
2.4. Definição de Medidas de Sucesso	7
3.1. Apresentação do Método de Aquisição de Requisitos	8
3.2. Requisitos de Descrição	8
3.3. Requisitos de Exploração	9
3.4. Requisitos de Controlo e Acesso	9
3.5. Revisão dos Requisitos com os Utilizadores	9
4.1. Apresentação da Metodologia de Desenvolvimento	10
4.2. Esquematização da Matriz de Decisão	10
4.3. Definição e Caracterização dos Data Marts e Grãos	11
4.4. Definição e Caracterização das Dimensões	12
4.5. Definição e Caracterização das Tabelas de Factos	13
4.6. Esquematização do Esquema Dimensional	16
4.7. Revisão do Esquema Dimensional Desenvolvido	16
5.1. Identificação e Descrição das Fontes de Informação do Sistema	17
5.2. Análise dos Dados das Fontes	18
5.2.1 Associação entre atributos e entidades	18
5.2.2 Domínio dos Atributos	20
5.3. Desenvolvimento do Esquema de Mapeamento de Dados	22
5.4. Revisão do Esquema de Mapeamento de Dados	23
6.1. Apresentação e Descrição Sumária das Estruturas do Data Warehouse a Povoar e das Fontes de Dados Envolvidas	24
6.2. Apresentação e Descrição do Modelo Lógico	26
6.3. Esquematização do Esquema Concetual do Sistema de Povoamento em BPMN	32

6.4. Descrição Detalhada do Sistema de Povoamento	37
6.5. Descrição e Caracterização de todos os Elementos de Dados	44
7.1. Escolha das Plataformas Computacionais	47
7.2. Implementação dos Esquemas Físicos dos Sistemas de Dados	48
7.2.1 O Sistema de Dados do Data Warehouse	48
7.2.2 O Sistema de Dados da Área de Retenção	49
7.3. Implementação do Sistema de Povoamento	49
7.4. Análise da Execução do Sistema de Povoamento	50

Índice de Figuras

Figura 1 - Esquema ilustrativo das fontes e respectivas proveniências dos dados	24
Figura 2 – Esquematização do processo geral de ETL	32
Figura 3 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados MySQL	33
Figura 4 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados MongoDB.	33
Figura 5 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados Excel	34
Figura 6 - Processo de extração e carregamento da dimensão Calendário	34
Figura 7 – Processo de limpeza para os dados dos clientes do MySQL	35
Figura 8 – Processo de conformidade para os dados dos clientes	35
Figura 9 – Processo de conciliação para os dados dos clientes	36
Figura 10 - Processo de conciliação para os dados das viagens	36
Figura 11 - Processo de conciliação para os dados das vendas	36
Figura 12 - Extração da fonte MySQL	37
Figura 13 - Extração da fonte MongoDB	38
Figura 14 - Extração da fonte Excel	39
Figura 15 - Processo de geração de datas	39
Figura 16 - Processo de conformidade	40
Figura 17 - Processo de conciliação para as dimensões	40
Figura 18 - Processo de conciliação para a tabela de factos - MySQL	41
Figura 19 - Processo de conciliação para a tabela de factos - MongoDB	41
Figura 20 - Processo de conciliação para a tabela de factos - Excel	42
Figura 21 - Limpeza da área de retenção	42
Figura 22 - Armazenamento das últimas vendas na área de retenção	43
Figura 23 - Tabelas de cleanup da fonte MySQL	44
Figura 24 - Tabelas de cleanup da fonte MongoDB	44
Figura 25 - Tabelas de cleanup e conformidade da fonte Excel	45
Figura 26 - Tabelas de quarentena	45
Figura 27 - Tabelas de conciliação	46

Figura 28 - Modelo Dimensional	48
Figura 29 - Média do número de vendas por mês	51
Figura 30 - Número de vendas por país destino	51
Figura 31 - Número de vendas por local	51
Figura 32 - Número de vendas por cliente	51

1. Introdução

1.1. Contextualização do Sistema

Em pleno século XXI, com a crescente globalização cultural e económica, que se iniciou na década de 1980, tem vindo a ser promovida a integração cultural e social das próprias comunidades. Com isso, aumenta o número de pessoas a querer conhecer mais e melhor as cidades e países que os rodeiam, levando a um considerável aumento de viagens turísticas.

Naturalmente, diversos indivíduos identificaram uma oportunidade de negócio, a partir da necessidade de fazer viagens de força fácil e cómoda. Assim, mais e mais agências de viagem se têm tentado impôr nesse que é um mercado altamente competitivo e diversificado. É o caso da agência “Belo Mundo”, que neste momento se encontra em fase de expansão e, tendo já um volume de negócio considerado sólido, na sua área de ação, adquiriu duas agências concorrentes.

Objetivamente, o volume de dados crescerá de forma extraordinária, uma vez que será necessário reunir todos os dados das três empresas e partilhar os mesmos. Também, segundo a administração, seria importante visualizar de forma rápida e a qualquer momento do horário de trabalho, informações acerca das viagens vendidas pela empresa, quais os tipos de cliente, os países, a duração, o custo, entre outras características. Esta é uma capacidade muito importante para definir estratégias de ação nos mercados altamente competitivos atuais, que permite ter perceção das áreas de influência da empresa, melhorar onde já se encontra estabelecida, bem como recuperar aquelas áreas em que não está a atingir as metas propostas. Assim, a empresa terá mais hipótese de se tornar maior e mais importante no seu ramo.

Quando a equipa de desenvolvimento de software foi contactada, após algumas reuniões e um estudo de negócio, ficou acordado que a melhor solução seria a construção de um sistema de *Data Warehouse*. Este oferece uma solução a preço viável e escalável, cumprindo todos os requisitos inicialmente identificados, bem como os que surgiram no decorrer das reuniões.

1.2. Fundamentação do Sistema

O objetivo inicial, que despoletou o contacto com a *software house*, foi o da necessidade de integrar os dados das agências compradas com a agência sede. Tendo sido compradas as diversas agências, e querendo-se que a gestão seja centralizada, é imperativo juntar todas as bases de dados usadas. Isto permite ter mais informações sobre os clientes, bem como as suas preferências. Todavia, este é um processo difícil e dispendioso, mas que se torna imperativo na fase de expansão da empresa.

Posteriormente, após diversas reuniões com os administradores da agência, foi possível identificar necessidades relativas à obtenção de informações das viagens vendidas. Ora, havendo a necessidade desta consulta ser feita em horário de expediente e a qualquer hora, por parte dos administradores, e dessa mesma ser expedita, é impossível contar com o sistema atual. De facto, o sistema em que a agência sede tem suporte, utiliza uma base de dados que não corresponde às necessidades, pois é orientada a pequenas transações de viagens.

Naturalmente, devido à união das três agências, há alguns problemas relativos à unificação dos dados. Mais uma vez, um *Data Warehouse* poderá facilitar o aumento de qualidade dos dados face ao contexto, uma vez que permite controlar os mesmos na importação.

Assim, conseguir-se-á uma maior fidelidade dos dados, por estes serem atualizados em tempo específico, não sobrecarregando os sistemas de bases de dados originalmente concebidos para a agência, que continuarão a funcionar em pleno e do mesmo modo que até agora. Com isto, há também uma redução nos custos de expansão, uma vez que não é necessário modificar o sistema já utilizado.

1.3. Viabilidade da Implementação do Sistema

De modo a analisar a viabilidade do *Data Warehouse* a instalar, coube à *software house* adicionar, sobre a análise inicial da empresa contratante, um estudo sobre a adequação da proposta acordada em reunião, relativamente às circunstâncias do mercado.

Uma agência de viagens de dimensão respeitável tem, tal como muitos outros negócios da atualidade, uma página online disponível, para além do balcão de atendimento físico. Assim, identifica-se uma necessidade de disponibilidade total obrigatória dos recursos informáticos durante o horário compreendido entre as 8h até às 24h. Com ordem à manutenção desta disponibilidade e, ao mesmo tempo, a capacidade de reunir e relacionar as diversas informações sobre o negócio, várias grandes empresas têm construído *Data Warehouses*. De facto, a popularidade desta solução já evidencia a sua praticidade.

As empresas concorrentes, com vista ao fortalecimento do seu negócio, têm vindo a tomar cada vez mais iniciativas com mais investimento a nível informático e a “Belo Mundo” não pode ficar atrás. Não sendo nenhuma inovação para o cliente a nível direto, é um recurso que permitirá estabelecer um *profiling* dos clientes e das características que estes mais apreciam numa viagem. Assim, será possível criar melhores acordos nos destinos mais desejados, reduzindo custos e aumentando margens, bem como melhorar as condições oferecidas nas viagens atuais, para além de perceber novos possíveis destinos de interesse. Com isto, a agência terá uma oferta mais agradável e mais utilizadores procurarão a mesma, sendo que o retorno será obtido pelo aumento de vendas associado às melhores decisões tomadas.

Conscientemente, foi feito um estudo de mercado, do qual surgiu esta necessidade. Os clientes expressaram que nem sempre as ofertas eram adequadas para eles, vendo-se forçados a procurar outras agências concorrentes. Esta desadequação partia não só do preço das viagens, mas também de fatores como duração, aeroporto de partida, pensão oferecida, atividades incluídas, datas, entre outros. Foi confirmada a receptividade de fazer mais viagens e com maior frequência, desde que estas se adaptassem aos interesses pessoais dos entrevistados.

Uma avaliação final da proposta por parte do departamento estratégico e financeiro da empresa concluiu que o orçamento estipulado se adequa às possibilidades e à expectativa de crescimento do ano. Mais ainda, tendo em conta a capacidade de armazenamento de informação e relacionamento do mesmo, este empreendimento é dado como uma mais valia futura para a organização e crescimento sustentado do negócio.

1.4. Estrutura do Relatório

Este relatório está estruturado seguindo uma linha de raciocínio sequencial e de acordo com os passos que devem ser executados em qualquer implementação de um *Data Warehouse*.

Assim, inicialmente, faz-se uma contextualização, onde se detalha a área de atuação deste projeto, assim como a sua fundamentação. Analisa-se também a viabilidade da implementação do sistema, necessária antes de qualquer projeto.

Como consequência, é necessário identificar com clareza a identidade do sistema, os recursos com que o grupo pode contar, e dividir o trabalho a ser realizado pelos elementos do grupo.

De forma a saber que objetivos e restrições o sistema deveria respeitar, foram elaboradas várias entrevistas entre a equipa e o cliente, com vista a levantamento de requisitos. Estes requisitos dividem-se em vários tipos, que sugerem diferentes necessidades do projeto.

Depois de recolhidas todas as informações essenciais à elaboração do *Data Warehouse* requerido, construiu-se a matriz de decisão, e iniciou-se o método dos 4 passos de Kimball. Assim, inicialmente, identifica-se a área de suporte à decisão a implementar, que consiste na escolha do *data mart* a desenvolver. Segue-se a escolha do grão, que indica, implicitamente, as dimensões e as medidas a serem construídas. Esta secção termina com a apresentação do modelo dimensional.

A próxima etapa do data design é a análise das fontes de dados, onde se estudam os atributos a serem extraídos, e elabora um pré mapeamento de dados, uma vez que já se dispõe de um modelo dimensional, devidamente caracterizado.

Seguidamente, descrevem-se, através de diagramas BPMN, todos os passos a seguir em cada processo: extração, transformação e carregamento. Estes esquemas permitem uma abstração útil que permite domar a complexidade real do sistema de povoamento.

Por fim, elabora-se uma secção de implementação, onde se mostra de concretamente os sistemas de povoamento e estruturas que o suportam (área de retenção, *Data Warehouse*, entre outros).

2. Planeamento e Gestão do Projeto

2.1. Definição da Identidade do Projeto

O projeto que é pretendido realizar traduz-se na realização de um Data Warehouse que consiste num repositório de dados organizados por assunto, integrados, variando no tempo e não voláteis que suportam os processos de tomada de decisão dos gestores.

O objetivo deste projeto é então criar uma ferramenta que seja capaz de unificar os dados de três fontes de informação diferentes e assistir as tomadas de decisões dos seus gestores para aumentar o rendimento da empresa Belo Mundo.

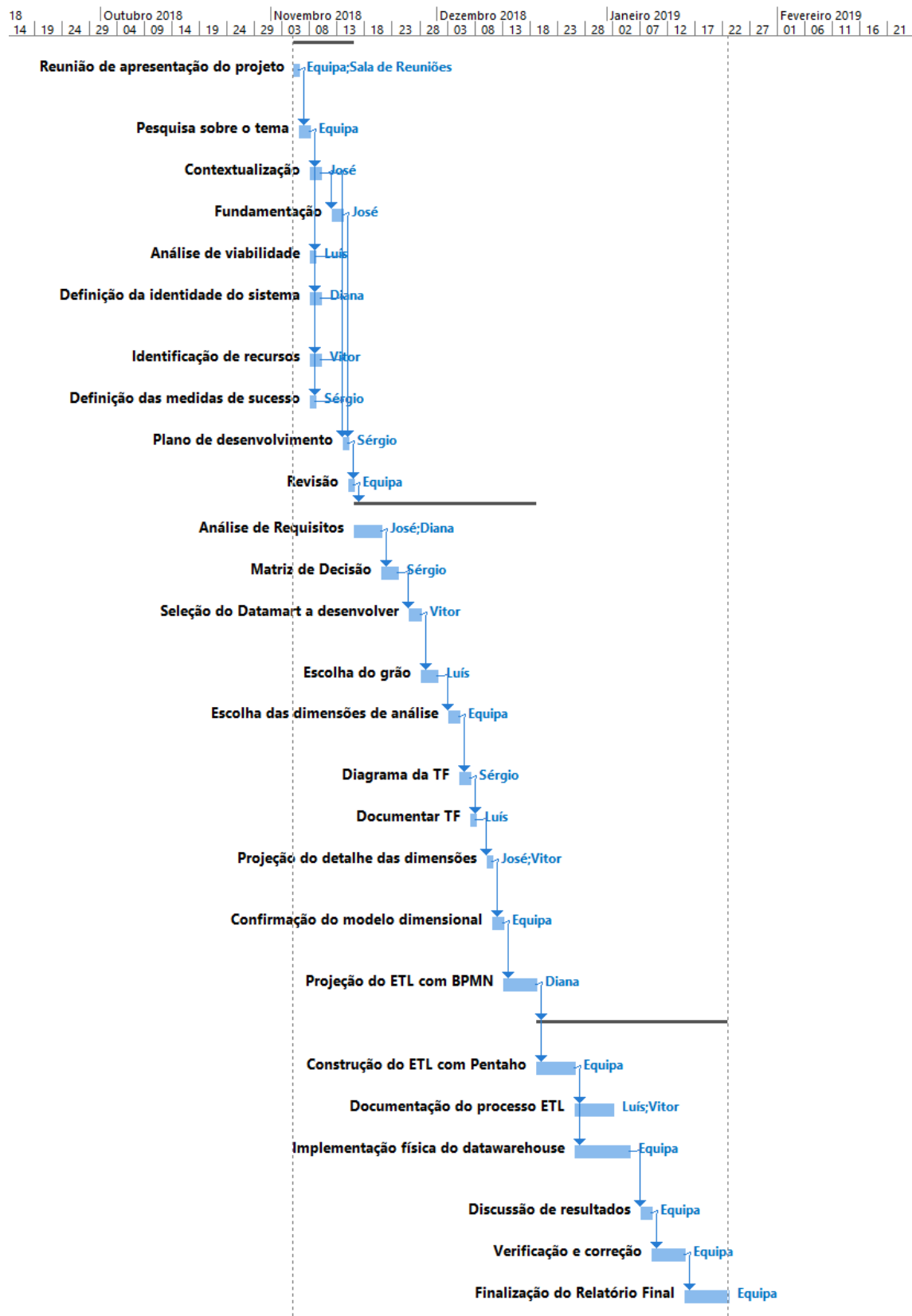
2.2. Identificação de Recursos

Para um bom desenvolvimento do projeto e para que não existam complicações de rendimento são necessários os seguintes recursos:

- Uma equipa de cinco pessoas capazes de realizar os objetivos pretendidos.
- Cinco máquinas, uma por cada elemento, que permitam usar ferramentas como o Edraw, MySQL Workbench, MongoDB, Compass, Robot3D, Microsoft Office Excel, Bizagi, Pentaho Data Integration, Microsoft Power BI e Microsoft Office Word.
- Um espaço em que a equipa possa trabalhar e discutir as suas ideias.
- Ajudas de custo em termos de deslocamento para o contacto com o cliente.

2.3. Estabelecimento do Plano de Desenvolvimento

Segue-se o plano de desenvolvimento, elaborado na ferramenta *Microsoft Project*.



2.4. Definição de Medidas de Sucesso

Para garantir a eficiência e rentabilidade na execução do projeto, é necessário estabelecer um conjunto de medidas no sentido de ser possível determinar o que é pretendido obter até o final do prazo estabelecido:

- O projeto deve estar terminado dentro do tempo previsto;
- O projeto deve ser concluído dentro do orçamento previsto;
- Ter ocorrido um desenvolvimento natural do projeto sem interrupções ou prejuízos nas atividades normais da equipa;
- Utilização eficiente dos recursos disponibilizados (ferramentas, equipamentos e pessoas) sem desperdícios;
- Ter atingido a qualidade e o desempenho desejados;
- Ter sido concluído com o mínimo possível de alterações do pretendido.
- Ter sido aceite sem restrições pelo contratante e/ou cliente.
- O projeto ser útil para o cliente;

3. Levantamento e Análise de Requisitos

3.1. Apresentação do Método de Aquisição de Requisitos

Para efetuar o levantamento de requisitos, a equipa de desenvolvimento procurou estabelecer com o cliente uma via de comunicação transparente, para tal, foram realizadas várias reuniões e entrevistas, onde foram discutidos os principais objetivos e restrições que o sistema a desenvolver deve respeitar.

3.2. Requisitos de Descrição

- Para a análise de cada venda deverão ser consideradas e registadas informações provenientes do cliente, das viagens compradas e da data da venda;
- A data deve especificar o dia (do mês e da semana), o mês, o ano, o trimestre e a época do ano associada.
- Todas as vendas devem estar registadas e identificadas inequivocamente, bem como o respetivo valor e número de viagens;
- Cada cliente é identificado com um número único e deve ser registado o seu nome, número de identificação fiscal, profissão e a localidade onde reside.
- A informação relativa a uma determinada viagem deve incluir um número identificativo, a data de realização, o preço do voo, o preço do hotel, o número de dias que durou a viagem, o país de destino e de chegada, o nome do hotel e o regime de pensão.

3.3. Requisitos de Exploração

- Identificar os períodos do ano com mais e menos fluxo de vendas;
- Averiguar os destinos com mais afluência;
- Averiguar os hotéis mais requisitados;
- Identificar a quantidade de viagens por faixa monetária;
- Relacionar o preço do hotel, do voo e o total com a quantidade de vendas;
- Identificar as épocas do ano em que se verificam maior valor de vendas;
- Analisar o tempo de duração mais comum das viagens realizadas;
- Identificar o regime de pensão mais e menos requisitado;
- Identificar os clientes que comprem mais viagens;
- Estabelecer relações entre as características dos clientes e as viagens que estes realizam;

3.4. Requisitos de Controlo e Acesso

- A equipa responsável pela gestão comercial da agência de viagens tem permissões de consulta de toda a informação presente no sistema;
- O administrador do sistema tem permissões ilimitadas;

3.5. Revisão dos Requisitos com os Utilizadores

Após o estabelecimento de todos os requisitos foram realizadas reuniões com o cliente com o intuito de validar os mesmos, tendo sido aprovadas todas as medidas consideradas para implementação deste sistema de *Data Warehousing*.

4. Modelação Dimensional do Sistema

4.1. Apresentação da Metodologia de Desenvolvimento

A metodologia adotada para o desenvolvimento do modelo dimensional foi a de quatro passos, de Kimball. Nesta, procura-se definir o processo de negócio e as suas envolventes, caracterizando todas as suas especificidades. Seguidamente, declara-se o grão, com a consciência da necessidade de manutenção da informação sobre o negócio analisado. Posteriormente, é feita a definição das dimensões e facto(s) associados ao processo.

Neste processo, é necessário entender que as necessidades requeridas são tão possíveis quanto os dados existentes nas fontes. Por isto, é necessário um forte diálogo e explanação do que é possível fazer com o que será, posteriormente, um *Data Warehouse*.

4.2. Esquematização da Matriz de Decisão

Caracterização do <i>Data Mart</i> “Comercial”	
Identificação: Comercial	
Descrição geral: Informação para suporte à tomada de decisão na área de venda de viagens, relativa à agência “Belo Mundo”, providenciando elementos de dados selecionados acerca das viagens vendidas e suas características, em todas as lojas, com motivação à gestão e controlo das mesmas, para ações comerciais e análise de vendas.	
Estrutura base	
Tabela de Factos >>>	TF-Vendas
<<< Dimensões	
Calendário	√
Viagens	√
Clientes	√
Número de Dimensões	3
Tipo	Transaccional
Periodicidade	Diária
Descrição	Transações comerciais de viagens.
Utilidade estratégica	Avaliação das características mais apreciadas nas viagens.

	Incentivar a venda de viagens. Identificar alterações na compra de viagens pelas suas características. Definição de ações promocionais. <i>Profiling</i> de clientes. Melhoria de base de negociação com cadeias hoteleiras, de transportes e de atividades. Identificação de áreas em sub-rendimento e sobre-rendimento face ao esperado.
Utilizadores	Administradores de operações.
Observações:	
Nada a assinalar.	

4.3. Definição e Caracterização dos Data Marts e Grãos

Após a análise do negócio, foi possível estabelecer o seguinte grão, que o define:

“A venda de uma viagem, em determinada quantidade e numa determinada data, requisitadas por um cliente, perfazendo um custo total.”

De facto, o motivo para a construção deste *Data Warehouse*, prende-se com o aumento de vendas e respetivo lucro. Por este motivo, e não tendo sido identificada outra necessidade que não diretamente relacionada com Vendas, esta é a única tabela de factos a impor. A dimensão Calendário consta para poder fazer relacionamentos com a data de uma venda. A dimensão

Clientes permite relacionar os clientes com as respetivas vendas, e a Viagens com as diferentes vendas feitas. Cada viagem é unitária, como se de um bilhete se tratasse. Estas dimensões serão atualizadas diariamente, sendo que a que sofrerá mais alterações, por adição de entradas, será a Viagens. Naturalmente, a dimensão Calendário permanecerá inalterada até ao momento em que se esgote a sua validade. A dimensão Clientes, crescerá de acordo com o número de novos clientes diários ou alterações a um cliente existente.

4.4. Definição e Caracterização das Dimensões

Dimensões do <i>Data Mart</i> “Comercial”			
Nr	Identificação	Descrição	Esquema (Tipo)
1	Cliente	Identificação e caracterização dos clientes das diversas agências.	Dim-Cliente (com Variação, com Criação de novos registo na tabela base)
2	Viagem	Identificação e caracterização de cada uma das viagens vendidas.	Dim-Viagem (Normal)
3	Calendário	A dimensão temporal. Acolhe todos os atributos que sustentem análises temporais, como a data, dia, dia da semana, dia do ano, mês, mês do ano, trimestre, ano e época. Regista a data de uma venda.	Dim-Calendário (Normal)

4.4.1 Dimensão Clientes

Esta dimensão guarda os dados dos clientes que já compraram uma viagem, tendo como atributos: skCliente, nome, NIF, profissão e local. Destes, há o interesse e possibilidade em agrupar por:

- Profissão;
- Local.

A dimensão tem variação e, por isso, importa notar quais os atributos e se têm histórico ou não.

- Profissão (com histórico);
- Local (com histórico).

4.4.2 Dimensão Viagens

Esta dimensão guarda os dados relativos à viagem vendida, tendo como atributos: skCliente, preço do voo, preço do hotel, preço total, número de dias, época, país de partida, país de destino, nome do hotel e pensão. Destes, há o interesse e possibilidade em agrupar por:

- Número de Dias;
- Época;
- País de Partida;
- País de Destino;
- Pensão.

4.4.3 Dimensão Calendário

Esta dimensão guarda a data em que cada uma das viagens foi vendida, tendo como atributos: id da data, data (“dia-mês-ano”), dia da semana (segunda, terça,...), dia da data (de 1 a 31), mês da data (de 1 a 12), ano da data, dia do ano (“dia-ano”), mês do ano (“mês-ano”), trimestre (“trim-ano”) e época. Destes, há o interesse e possibilidade em agrupar por:

- Época;
- Dia do Ano;
- Mês do Ano;
- Mês, que pode ser agrupado por *roll-up* com Trimestre e, por sua vez, Ano;
- Semana e, por *roll-up*, Dia da semana.

4.5. Definição e Caracterização das Tabelas de Factos

Os processos do negócio são representados por um modelo dimensional que consiste numa tabela de factos, onde as medições numéricas factuais associadas às dimensões serão registadas. Para além disso, as tabelas de dimensões irão conter todas as informações verdadeiras no momento de carregamento das mesmas. A tabela de factos deve ser simples, por forma a simplificar a sua leitura e também as operações sobre a mesma.

A tabela de factos representada a seguir refere-se a uma venda de uma viagem. Nessa venda será guardada informação relativa ao número de viagens compradas na mesma, bem como o valor total das mesmas.

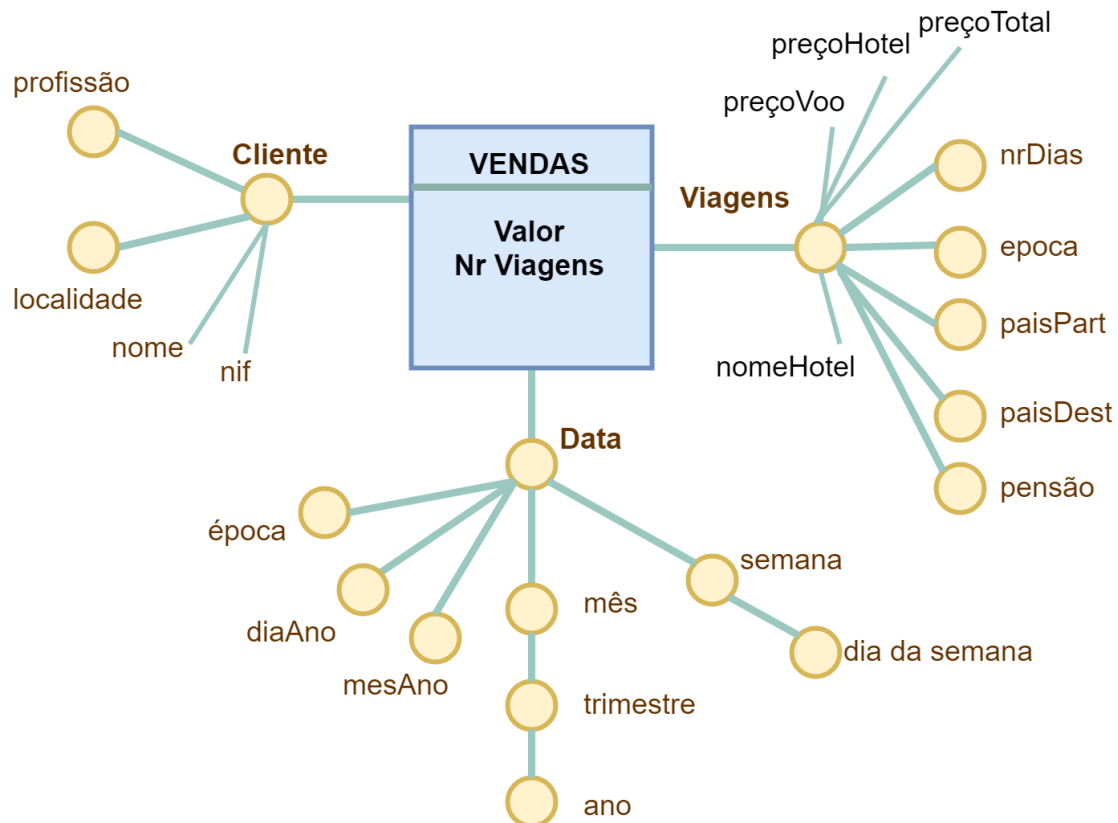
Foi definido o seguinte grão, como explicitado anteriormente, que dá lugar à tabela de factos abaixo:

“A venda de uma viagem, em determinada quantidade e numa determinada data, requisitadas por um cliente, perfazendo um custo total.”

Caracterização da tabela de factos				
Identificação		TF-Vendas		
Descrição		Tabela que acolhe os vários registos de cada uma das vendas a cada um dos clientes.		
Data <i>mart</i>		Comercial		
Tipo		Transacional		
Utilidade Estratégica		Avaliação das características mais apreciadas nas viagens. Incentivar a venda de viagens. Identificar alterações na compra de viagens pelas suas características. Definição de ações promocionais. <i>Profiling</i> de clientes. Melhoria de base de negociação com cadeias hoteleiras, de transportes e de atividades. Identificação de áreas em sub-rendimento e sobre-rendimento face ao esperado.		
Povoamento		Realizado diariamente entre as 00:00h até às 8:00h.		
Dimensão Inicial		65 KB		
Crescimento		25% ano.		
Período de dados		Desde o início do ano de 2015. As restantes informações ficarão em arquivos.		
Atributos				
Dimensões				
Nr	Identificação	Chave	Descrição	Exemplo
1	skCliente	S	Chave de substituição para um cliente do conjunto de agências.	1
2	skViagem	S	Chave de substituição para uma viagem do	1

			conjunto de agências.	
3	idData	S	Código da data referente à data de uma venda.	1
Medidas				
Nr	Identificação	Domínio	Descrição	Exemplo
1	nrViagens	Inteiro	Número de viagens vendidas numa venda.	2
2	valor	Inteiro	Valor total da venda efetuada.	2400
Perfis de Utilização				
Administrador da base de dados.				
Observações				
-				

4.6. Esquematização do Esquema Dimensional



4.7. Revisão do Esquema Dimensional Desenvolvido

Depois de bastante diálogo com o cliente, e após concluído aquele que viria a ser o modelo dimensional definitivo, o mesmo foi apresentado.

Após uma reunião muito descritiva sobre os assuntos relacionados com o modelo dimensional, e o que levou à tomada das diversas decisões que envolveram a sua construção, o mesmo foi aprovado.

O cliente observou que as necessidades de obtenção de informação relativas às viagens vendidas, bem como os seus atributos, a determinados clientes, se cumpriu. Foi, assim, dado início ao processo de ETL.

5. Caracterização das Fontes de Informação

5.1. Identificação e Descrição das Fontes de Informação do Sistema

As fontes de dados são algo que, naturalmente, acabou por surgir. São estas que compõem o Data Warehouse, sendo que, é a partir dessa informação e desses dados que se fazem todas as análises necessárias.

Além disso, uma vez que está em causa o balanço e a exploração de dados de três empresas, verificou-se a necessidade de trabalhar com três bases de dados.

Foi também possível verificar, junto do cliente, que todas elas são iguais, no sentido em que, representam a mesma informação. Ou seja, há na realidade uma empresa no mercado português especializada em montagem de Bases de Dados, à qual estas firmas recorreram, no passado, e por isso, a diferença entre as fontes reside, única e exclusivamente, no motor de dados. Estava-se, então, perante uma base de dados relacional (MySQL), uma base de dados orientada por documentos e uma base de dados em Microsoft Excel.

5.2. Análise dos Dados das Fontes

5.2.1 Associação entre atributos e entidades

Nome da entidade	Atributo	Descrição	Tipo	Nulo	Multivalor	Derivado	Composto
Cliente	ID	Identificador do cliente	Inteiro	Não	Não	Não	Não
	Nome	Nome do cliente	Caracteres	Não	Não	Não	Não
	Contacto	Nº telemóvel do cliente	Inteiro	Não	Não	Não	Não
	Email	Email do cliente	Caracteres	Não	Não	Não	Não
	NIF	Número de Identificação Fiscal do cliente	Inteiro	Não	Não	Não	Não
	Hobbies	Hobbies do cliente	Caracteres	Não	Sim	Não	Não
	Profissão	Profissão do cliente	Caracteres	Não	Não	Não	Não
	Endereço	Morada do cliente	Caracteres	Não	Não	Não	Sim
Venda	Número	Identificador da venda	Inteiro	Não	Não	Não	Não
	Data	Data da venda	DATETIME	Não	Não	Não	Não
	Valor	Preço da venda	Inteiro	Não	Não	Sim	Não
	NrViagens	Número de viagens da venda	Inteiro	Não	Não	Não	Não
Viagem	Id	Identificador da viagem	Caracteres	Não	Não	Não	Não
	PreçoVoo	Preço do voo	Inteiro	Não	Não	Não	Não
	PreçoHotel	Preço do hotel	Inteiro	Não	Não	Não	Não
	PreçoTotal	Preço total	Inteiro	Não	Não	Sim	Não

	NrDias	Número de dias da viagem	Inteiro	Não	Não	Não	Não
	Categoria	Categoria da viagem	Caracteres	Não	Sim	Não	Não
	Época	Época da viagem	Caracter	Não	Não	Não	Não
	Data	Data da viagem	DATE	Não	Não	Não	Não
	PaísPart	País de partida	Caracteres	Não	Não	Não	Não
	PaísDest	País de chegada	Caracteres	Não	Não	Não	Não
	Pensão	Pensão da viagem	Caracteres	Não	Não	Não	Não
Hotel	Código	Identificador do hotel	Inteiro	Não	Não	Não	Não
	Nome	Nome do hotel	Caracteres	Não	Não	Não	Não
	Contacto	Contacto do hotel	Inteiro	Não	Não	Não	Não
	Email	Email do hotel	Caracteres	Não	Não	Não	Não
	Endereço	Morada do hotel	Caracteres	Não	Não	Não	Sim
	Estrelas	Número de estrelas do hotel	Inteiro	Não	Não	Não	Não
	Pensão	Pensões oferecidas pelo hotel	Caracteres	Não	Sim	Não	Não
Rota	Id	Identificador da rota	Caracteres	Não	Não	Não	Não
	Companhia	Companhia aérea responsável	Caracteres	Não	Não	Não	Não
	Duração	Duração da viagem	Inteiro	Não	Não	Não	Não
	AeroportoPart	Aeroporto de partida	Caracteres	Não	Não	Não	Não
	AeroportoDest	Aeroporto de destino	Caracteres	Não	Não	Não	Não

5.2.2 Domínio dos Atributos

- **Entidade Cliente**
 - **ID:** Inteiro, que identifica o cliente no sistema;
 - **Nome:** String, que contém o nome do cliente;
 - **Contacto:** Inteiro, onde está o contacto telefónico do cliente;
 - **Email:** String, que contém o endereço eletrónico do cliente;
 - **NIF:** Inteiro, que corresponde à identificação fiscal do cliente;
 - **Hobbies:** String que é um atributo multivalorado onde estão guardados os passatempos do cliente;
 - **Profissão:** String, que contém a profissão do cliente
 - **Endereço:** String, onde está guardada a morada do cliente;
- **Entidade Venda**
 - **Número:** Inteiro, identifica a venda na loja;
 - **Data:** DATETIME, onde está a data e a hora da venda;
 - **Valor:** Inteiro, que contém o valor total da venda
 - **NrViagens:** Inteiro, onde está o número de viagens vendidas na venda;
- **Entidade Viagem**
 - **ID:** String, identifica a viagem;
 - **PrecoVoo:** Inteiro, que contém o preço do voo ida/volta;
 - **PrecoHotel:** Inteiro, que contém o preço do hotel;
 - **PrecoTotal:** Inteiro, que é um atributo derivado dos dois anteriores e que guarda o valor total da viagem;
 - **NrDias:** Inteiro, onde está o número de dias da viagem;
 - **Categoria:**String, que identifica a categoria da viagem. Ex: cultura, negócios, etc;
 - **Época:** Caracter, que identifica a época do ano em que a viagem se realizará;
 - **Data:** DATE, que contem a data da viagem;
 - **PaísPart:** String, que contém o país de partida. Corresponderá sempre ou, na maioria das vezes, a Portugal;
 - **PaísDest:** String, que identifica o país de destino;
 - **Pensão:** String, que identifica a pensão do hotel. Poderá ser sem alimentação (SA), pensão completa (PC), meia pensão (MP) ou pequeno almoço (PA);

- **Entidade Hotel**

- **Código:** Inteiro, que identifica o hotel no sistema;
- **Nome:** String, que contém o nome do hotel;
- **Contacto:** Inteiro, onde está o contacto telefónico do hotel;
- **Email:** String, onde está o endereço eletrónico do hotel;
- **Endereço:** String, que contém a morada do hotel;
- **Estrelas:** Inteiro, que contém o número de estrelas do hotel;
- **Pensão:** String, que é um atributo multivalorado e que identifica as pensões oferecidas pelo hotel. Poderá ser sem alimentação (SA), pensão completa (PC), meia pensão (MP) ou pequeno almoço (PA);

- **Entidade Rota**

- **ID:** String, que identifica a rota/voo no sistema;
- **Companhia:** String, que identifica a companhia aérea responsável pelo voo;
- **Duração:** Inteiro, que guarda a duração do voo ida/volta;
- **AeroportoPart:** String, que contém o aeroporto de partida (Aeroporto da Portela ou Aeroporto Sá Carneiro);
- **AeroportoDest:** String, que contém o aeroporto de destino;

5.3. Desenvolvimento do Esquema de Mapeamento de Dados

Depois de analisados todos os atributos de todas as fontes de dados, descreve-se agora, genericamente, como serão mapeados os valores entre as fontes e o destino (*data warehouse*). Este estudo é uma espécie de pré *source-to-target data map*. Assim, divide-se a análise em três partes, como se vê se seguida:

- MySQL
 - O modelo final guarda informações relativas ao cliente, tais como nome, NIF, profissão e local. Esta informação é proveniente das tabelas originais “Cliente” e “LocalCliente”, onde a extração é direta, com o auxílio de uma operação de *join* para o caso do local;
 - No Data Warehouse também existirá informação correspondente às viagens, com atributos como preço do voo, preço do hotel, número de dias da viagem, época, país de destino, nome do hotel, pensão, entre outros. Para tal, a informação será diretamente extraída das tabelas originais “Viagem” e “Hotel”, com a ajuda de operações de *join*;
 - Por fim, para a tabela de factos, todos os dados (chaves e medidas como valor da venda ou número de viagens adquiridas) são de extração direta das tabelas originais “Venda” e Viagem, mais uma vez com auxílio de *joins*.
- MongoDB
 - Para a dimensão do cliente, os dados são diretos da coleção “Clientes”, não precisando de nenhuma operação ou transformação;
 - No que toca à dimensão das viagens, os dados são provenientes das coleções originais “Viagens” e “Hoteis”, apenas se efetuando algumas comparações de ids, de forma a unir informação de diferentes coleções;
 - Já para a tabela de facto, referente às vendas, os dados vieram da tabela original “Cliente”, sendo a passagem direta.
- Excel
 - Para a dimensão correspondente aos clientes, viagens ou tabela de factos de vendas, a informação advém toda da mesma folha de cálculo, e é direta, não precisando de nenhuma transformação.

5.4. Revisão do Esquema de Mapeamento de Dados

Após a realização do esquema de mapeamento de dados, este foi revisto pela equipa de desenvolvimento. Nenhum problema foi detetado e, por isso, foi aceite.

6. Modelação do Sistema de Povoamento

Apresentam-se, nesta secção, os detalhes associados a todo o sistema de povoamento do DW, desde a simples extração dos dados das três fontes de informação, passando pela transformação dos mesmos, até às decisões tomadas no carregamento dos dados para o sistema de apoio à decisão.

6.1. Apresentação e Descrição Sumária das Estruturas do Data Warehouse a Povoar e das Fontes de Dados Envolvidas

Tal como descrito na secção 1.1, a agência de viagens Belo Mundo, ao comprar duas outras agências concorrentes, passou a ter que lidar com dados provenientes de três fontes, já que todas continuaram a existir separadamente. Assim, cada uma tem clientes, viagens, hotéis, rotas e vendas que é preciso analisar.

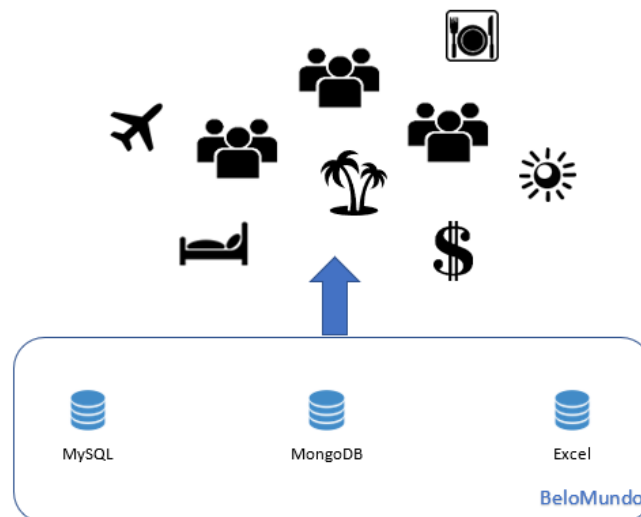


Figura 1 - Esquema ilustrativo das fontes e respetivas proveniências dos dados

Para tal, e segundo o esquema dimensional já apresentado na secção 4.6, optou-se por analisar vendas por **clientes**, **viagens** e **data**. Estas necessidades de análise constituem as tabelas que serão necessárias povoar.

No que toca ao cliente, será extraída, tratada e povoada informação referente ao seu nome, NIF (critério de comparação), profissão e local (morada). Informação como a profissão e o local permitem sugerir viagens por *profiling*, e fornecer ofertas personalizadas a todos os clientes da Belo Mundo.

Já para as viagens, a informação constituinte do povoamento inclui o preço do voo, do hotel e total, o que é útil para analisar tendências de custos. Constam também dados relativos ao número de dias da viagem, época, pensão, países de origem e destino, e nome do hotel. Mais uma vez, é possível analisar tendências e fazer *profiling* por cliente, sugerindo melhores viagens.

Por fim, para conseguir fornecer dados numa perspectiva histórica, tem-se uma estrutura que serve de calendário. Nesta existem dados como a data completa, o dia da semana, mês do ano, dia do ano, ano, trimestre, entre outros, de forma a analisar informação conforme um período de tempo.

6.2. Apresentação e Descrição do Modelo Lógico

MySQL

Tabela Origem	Atributo Origem	Tipo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
Cliente	Id	INT	DIRETO	DIM-Clientes	idCliente	INT
Cliente	Nome	VARCHAR	DIRETO	DIM-Clientes	nome	VARCHAR
Cliente	NIF	INT	DIRETO	DIM-Clientes	nif	INT
Cliente	Profissão	VARCHAR	DIRETO	DIM-Clientes	profissao	VARCHAR
LocalCliente	Designacao	VARCHAR	JOIN Cliente ON LocalCliente.id = Cliente.LocalCliente_Id	DIM-Clientes	local	VARCHAR
Viagem	Id	INT	DIRETO	DIM-Viagens	idViagens	INT
Viagem	Preco_Voo	INT	DIRETO	DIM-Viagens	precoVoo	INT
Viagem	Preco_Hotel	INT	DIRETO	DIM-Viagens	precoHotel	INT
Viagem	Preco_Total	INT	DIRETO	DIM-Viagens	precoTotal	INT
Viagem	Nr_Dias	INT	DIRETO	DIM-Viagens	nrDias	INT
Viagem	Epoca	CHAR	DIRETO	DIM-Viagens	epoca	CHAR
Viagem	Pais_Part	VARCHAR	DIRETO	DIM-Viagens	paisPart	VARCHAR
Viagem	Pais_Dest	VARCHAR	DIRETO	DIM-Viagens	paisDest	VARCHAR
Viagem	Pensao	CHAR(2)	DIRETO	DIM-Viagens	pensao	CHAR(2)

Tabela Origem	Atributo Origem	Tipo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
Hotel	Nome	VARCHAR	JOIN VIAGEM ON Hotel.Codi go = Viagem.H otel_Codig o	DIM-Viagens	nomeHotel	VARCHAR
Venda	Numero	INT	DIRETO	TF-Vendas	idVenda	INT
Venda	Valor	INT	DIRETO	TF-Vendas	valor	INT
Venda	Nr_Viagens	INT	DIRETO	TF-Vendas	nrViagens	INT
Venda	Cliente_Id	INT	DIRETO	TF-Vendas	DIM- Clientes_idC liente	INT
Viagem	Id	INT	JOIN VENDA ON Viagem.V enda_Nu mero = Venda.Nu mero	TF-Vendas	DIM- Viagens_idV iagens	INT

MongoDB

Coleção Origem	Atributo Origem	Tipo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
Clientes	Id	INT	DIRETO	DIM-Clientes	idCliente	INT
Clientes	Nome	VARCHAR	DIRETO	DIM-Clientes	nome	VARCHAR
Clientes	NIF	INT	DIRETO	DIM-Clientes	nif	INT
Clientes	Profissão	VARCHAR	DIRETO	DIM-Clientes	profissao	VARCHAR
Clientes	Local	VARCHAR	DIRETO	DIM-Clientes	local	VARCHAR
Viagens	Id	INT	DIRETO	DIM-Viagens	idViagens	INT
Viagens	Preco_Voo	INT	DIRETO	DIM-Viagens	precoVoo	INT
Viagens	Preco_Hotel	INT	DIRETO	DIM-Viagens	precoHotel	INT
Viagens	Preco_Total	INT	DIRETO	DIM-Viagens	precoTotal	INT
Viagens	Nr_Dias	INT	DIRETO	DIM-Viagens	nrDias	INT
Viagens	Epoca	VARCHAR	DIRETO	DIM-Viagens	epoca	CHAR
Viagens	Pais_Part	VARCHAR	DIRETO	DIM-Viagens	paisPart	VARCHAR
Viagens	Pais_Dest	VARCHAR	DIRETO	DIM-Viagens	paisDest	VARCHAR
Viagens	Pensao	VARCHAR	DIRETO	DIM-Viagens	pensao	CHAR(2)
Hoteis	Nome	VARCHAR	DIRETO	DIM-Viagens	nomeHotel	VARCHAR
Clientes	Vendas -> Numero	INT	DIRETO	TF-Vendas	idVenda	INT
Clientes	Vendas -> Valor	INT	DIRETO	TF-Vendas	valor	INT

Coleção Origem	Atributo Origem	Tipo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
Clientes	Vendas -> NrViagens	INT	DIRETO	TF-Vendas	nrViagens	INT
Clientes	Id	INT	DIRETO	TF-Vendas	DIM- Clientes_i dCliente	INT
Clientes	Vendas -> Viagem -> Id	INT	DIRETO	TF-Vendas	DIM- Viagens_i dViagens	INT

Excel

Atributo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
Id	DIRETO	DIM-Clientes	idCliente	INT
Nome	DIRETO	DIM-Clientes	nome	VARCHAR
NIF	DIRETO	DIM-Clientes	nif	INT
Profissão	DIRETO	DIM-Clientes	profissao	VARCHAR
Designacao	DIRETO	DIM-Clientes	local	VARCHAR
Id	DIRETO	DIM-Viagens	idViagens	INT
Preco_Voo	DIRETO	DIM-Viagens	precoVoo	INT
Preco_Hotel	DIRETO	DIM-Viagens	precoHotel	INT
Preco_Total	DIRETO	DIM-Viagens	precoTotal	INT
Nr_Dias	DIRETO	DIM-Viagens	nrDias	INT
Epoca	DIRETO	DIM-Viagens	epoca	CHAR
Pais_Part	DIRETO	DIM-Viagens	paisPart	VARCHAR
Pais_Dest	DIRETO	DIM-Viagens	paisDest	VARCHAR
Pensao	DIRETO	DIM-Viagens	pensao	CHAR(2)
Nome	DIRETO	DIM-Viagens	nomeHotel	VARCHAR
NrVenda	DIRETO	TF-Vendas	idVenda	INT
ValorVenda	DIRETO	TF-Vendas	valor	INT

Atributo Origem	Regra	Tabela Destino	Atributo Destino	Tipo Destino
NrViagensVenda	DIRETO	TF-Vendas	nrViagens	INT
idCliente	DIRETO	TF-Vendas	DIM-Clientes_idCliente	INT
idViagem	DIRETO	TF-Vendas	DIM-Viagens_idViagens	INT

6.3. Esquematização do Esquema Concetual do Sistema de Povoamento em BPMN

Apresenta-se, de seguida, toda a modelação conceptual do processo de ETL. Começa-se por mostrar um diagrama mais geral, onde estão descritas as três fases do sistema de povoamento, e depois especifica-se cada fase: extração, transformação (limpeza, conformidade, conciliação) e carregamento dos dados para o *Data Warehouse*.

Neste diagrama geral, é visível a extração dos dados para cada fonte de dados, com as consequentes transformações ao nível da área de retenção. Este esquema termina com o carregamento dos dados transformados para o esquema físico final, em *MySQL*.

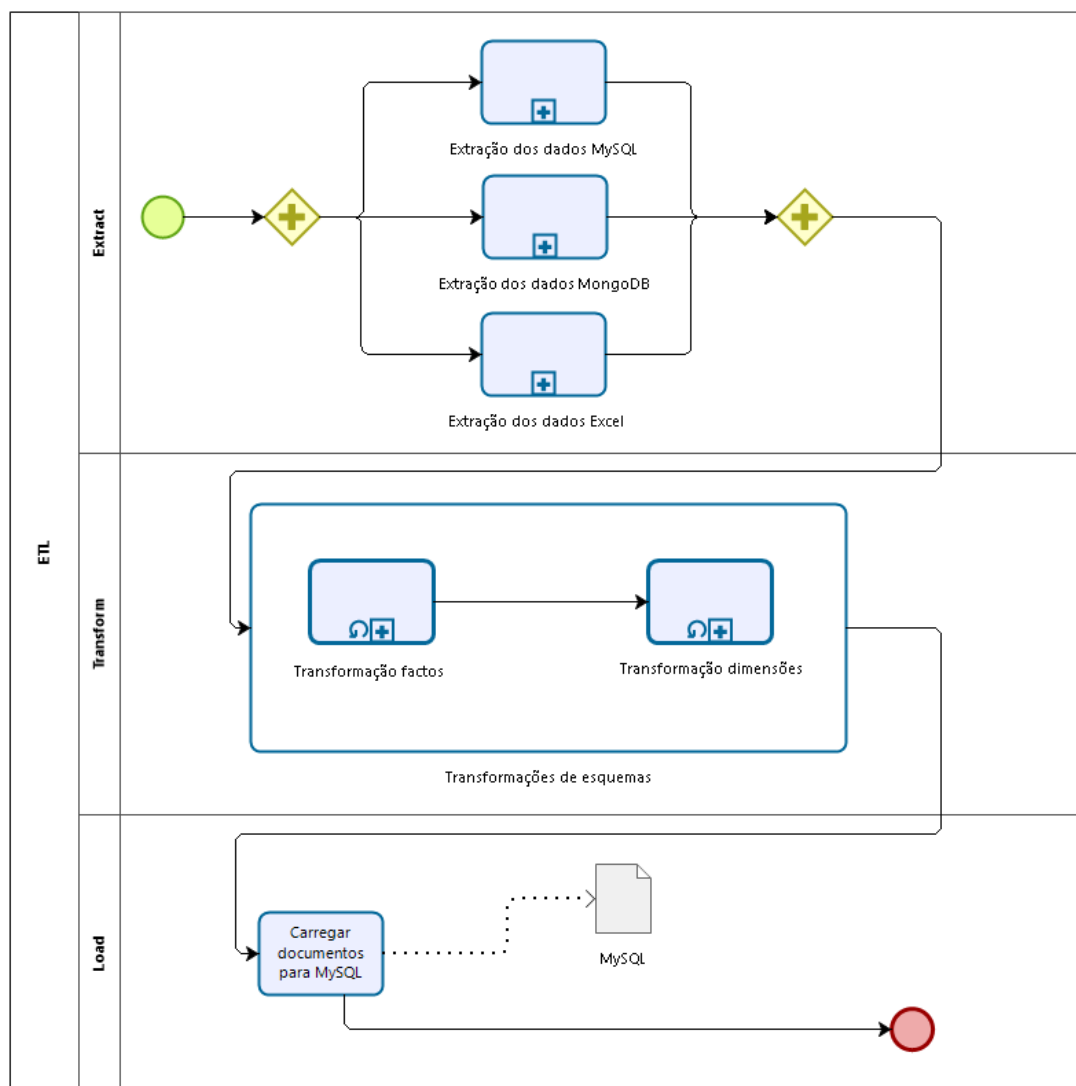


Figura 2 – Esquematização do processo geral de ETL

6.3.1 Extração

A extração é a primeira fase do processo de povoamento, e aqui ocorre a extração dos dados de cada uma das fontes - *MySQL*, *MongoDB* e *Excel* –, com o posterior carregamento dos dados para a área de retenção. Esta extração, efetuada em paralelo para as três fontes, tem como objetivo a seleção e adequação dos dados originais de forma a poderem ser mantidos em tabelas de auditoria e, mais tarde, carregados nas tabelas de dimensão e de factos. Assim, essencialmente, são feitas operações nos dados como *joins* e *sorts*, processos específicos que serão explicados mais em detalhe na secção 6.4. Para além disso, efetua-se já alguma transformação (conformidade) dos dados, de maneira que alguns já se assemelham ao seu formato final (do *Data Warehouse*).

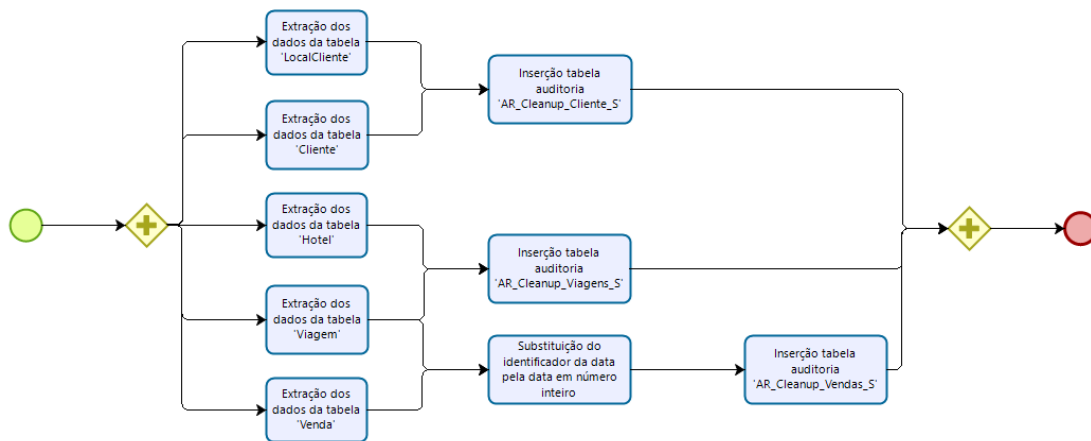


Figura 3 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados MySQL

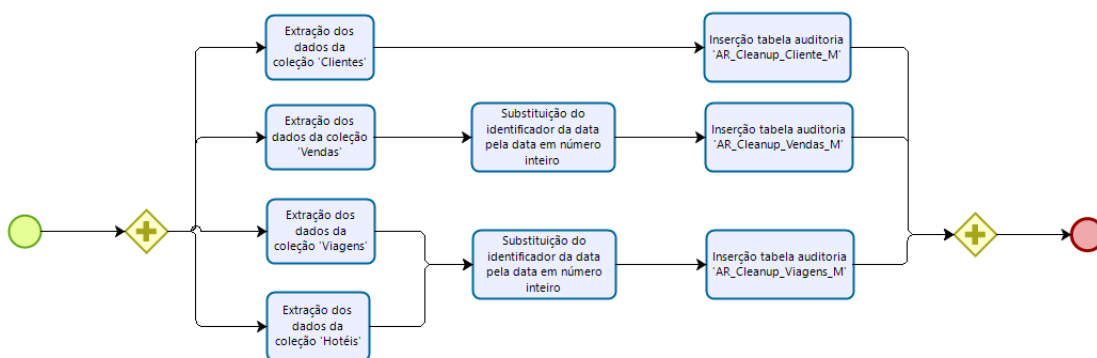


Figura 4 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados MongoDB.

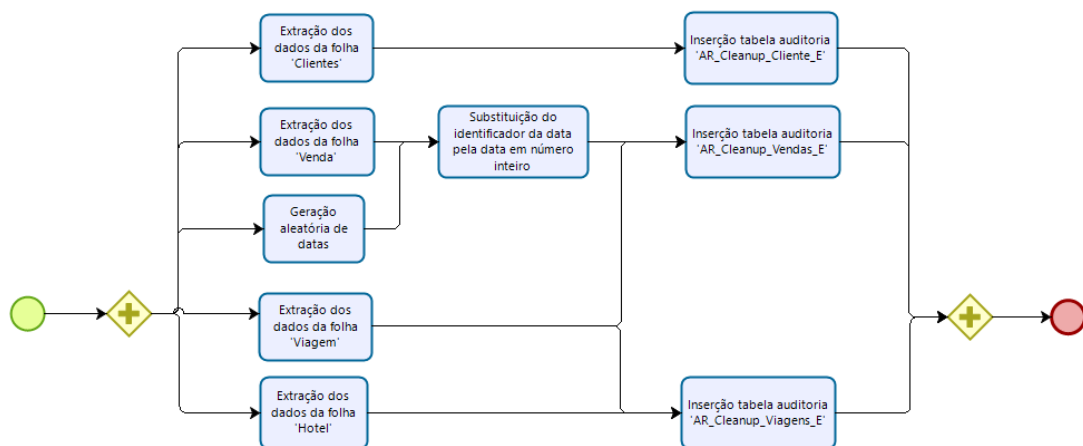


Figura 5 - Processo de extração dos clientes, viagens, hotéis e vendas da fonte de dados Excel

O processo ETL para a dimensão Calendário terá um comportamento excecional, na medida que esta dimensão é quase carregada de imediato para o *Data Warehouse*. Isto deve-se ao facto de ser uma dimensão sem variação e sem qualquer tipo de falhas, uma vez que as datas foram geradas automaticamente.

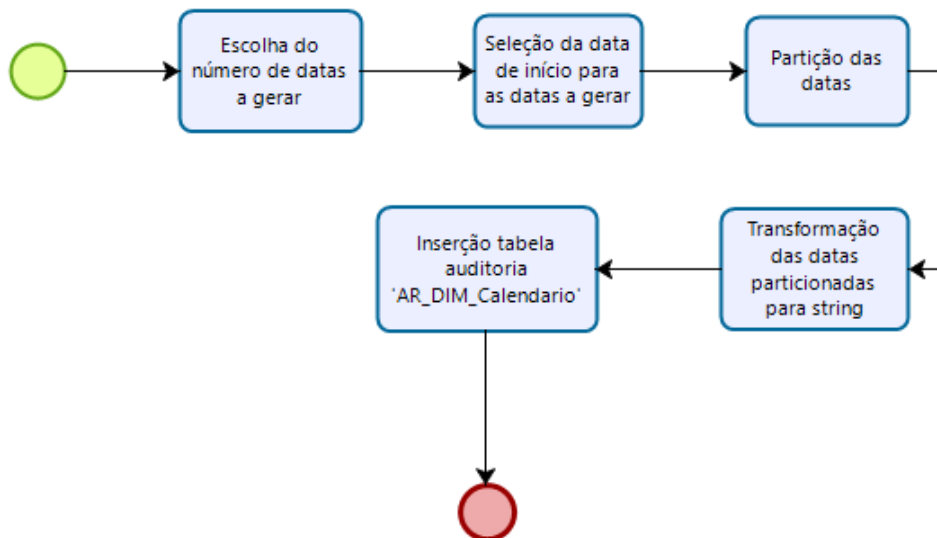


Figura 6 - Processo de extração e carregamento da dimensão Calendário

6.3.2 Transformação – Limpeza

A transformação inicia-se com o processo de limpeza. Nesta fase, é necessário analisar e “limpar” os dados extraídos das fontes de informação, de modo a eliminar dados que não exprimam informação relevante ou precisa para o sistema de apoio à decisão. Deste modo, este processo passa apenas pela verificação do local do cliente da fonte de dados MySQL, onde verificou-se que é algo típico de acontecer. Em caso de não existir (valor nulo), é atribuída a string ‘desconhecido’ ao registo em questão. (Este processo de limpeza ocorre aquando da extração, mas situa-se nesta secção de forma a separar as etapas da transformação.)

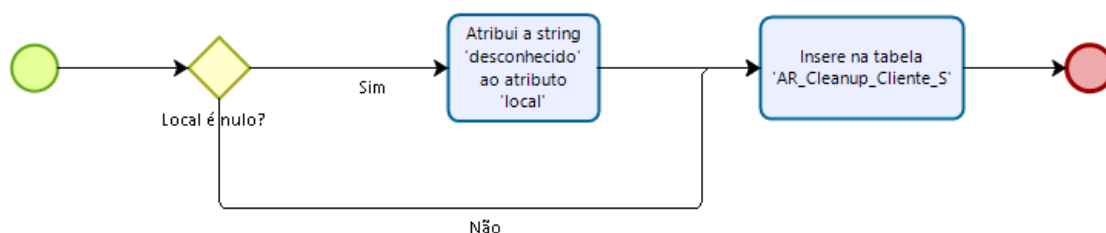


Figura 7 – Processo de limpeza para os dados dos clientes do MySQL

6.3.3 Transformação - Conformidade

Nesta etapa intermédia da transformação, a informação já se encontra limpa e extraída, sendo agora necessário transformá-la para o seu devido formato final, a ser carregado no DW. A maioria dos dados já sofre uma conformidade no processo de extração, com exceção dos dados da fonte Excel. Para tal, é necessária a concatenação do primeiro e último nome do cliente, provenientes da fase de limpeza.



Figura 8 – Processo de conformidade para os dados dos clientes

6.3.4 Transformação – Conciliação

A fase final da transformação inclui a integração dos dados das diferentes fontes entre si, preparando-os para a fase final de carregamento dos dados. Posto isto, e dado que as fontes de informação são heterogêneas, é necessário o recurso a chaves de substituição, de forma a evitar a duplicação de chaves e a consequente perda de integridade.

As chaves de substituição são geradas aquando uma operação de inserção, verificando a ocorrência de um nulo imprevisto (ou inserido um valor em branco; sem significado). Em caso afirmativo, o registo é enviado para quarentena, onde ficará para ser tratado e inserido quando ocorrer um refrescamento. Caso contrário é gerada uma chave de substituição e o registo é inserido no *Data Warehouse*, na respetiva dimensão.

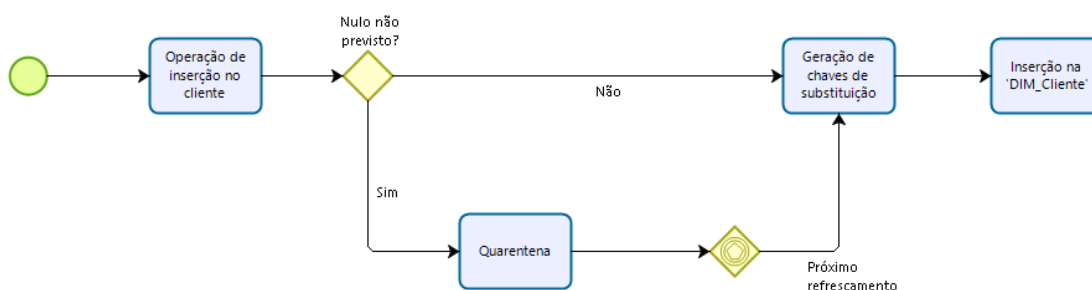


Figura 9 – Processo de conciliação para os dados dos clientes

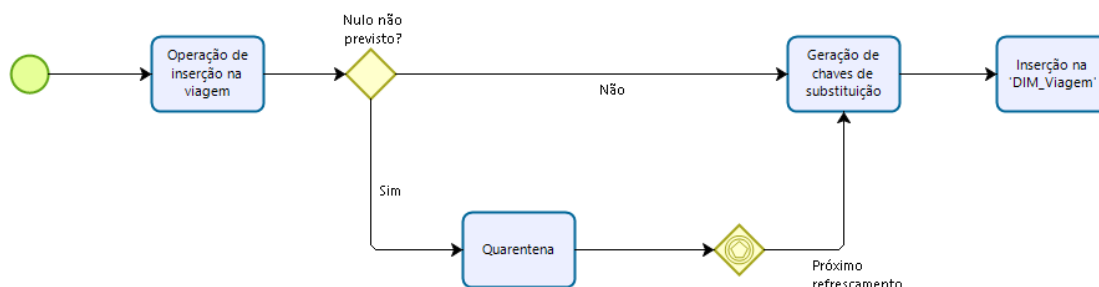


Figura 10 - Processo de conciliação para os dados das viagens

No caso dos dados referentes às vendas, a geração de chaves é direta, não precisando de passar por um processo de tratamento de nulos imprevistos.

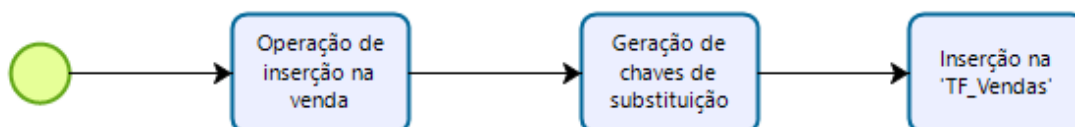


Figura 11 - Processo de conciliação para os dados das vendas

6.3.5 Carregamento

Finalmente, após a execução de todos os processos anteriores, realiza-se o carregamento definitivo dos dados presentes na área de retenção para o *Data Warehouse*. Este processo é praticamente direto, uma vez que a transformação permitiu que o tipo de dados fosse compatível com os do sistema de suporte à decisão. Ainda assim, é necessário ter em conta as dimensões com variação (tipo 2) e os respetivos *inserts*.

É importante referir que, optou-se por uma implementação de tipo 2, uma vez que se considerou que alterações nos campos profissão ou local representam mudanças no histórico, relativamente a determinando NIF, conceitualmente. Ou seja, um estudante terá potencialmente um emprego no seu futuro e por isso maior poder de compra. Quer-se analisar por *profiling* e não por cliente específico.

Desta forma, todos os diagramas BPMN para as transformações referentes ao carregamento seriam diretos e triviais, o que não justifica a sua elaboração.

6.4. Descrição Detalhada do Sistema de Povoamento

Uma vez descrito, de uma maneira leve, o sistema de povoamento através dos diagramas BPMN, cabe ao grupo mostrar agora, com detalhe, todos os passos necessários desde a extração das fontes até ao carregamento no *Data Warehouse* final. Aborda-se não apenas o sistema de povoamento inicial, mas também o que é imposto para o refrescamento.

Assim, começa-se pela fase de **carregamento**. Esta fase está dividida em três etapas, correspondentes às três fontes de dados. Quanto à extração da BD implementada em MySQL, têm-se as seguintes transformações:

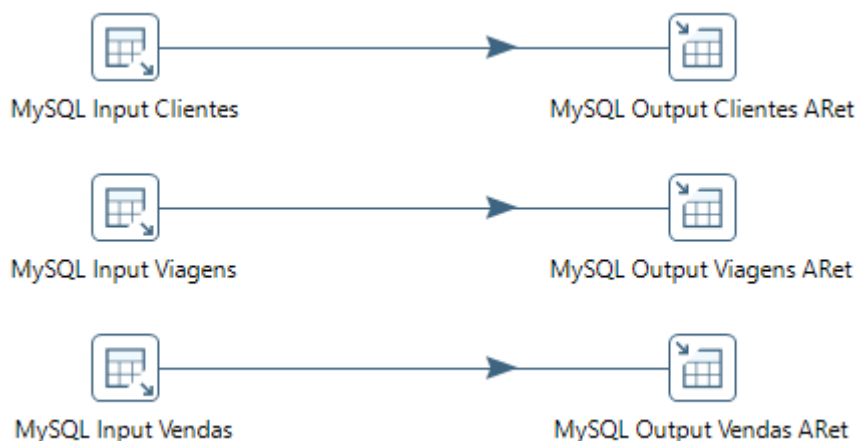


Figura 12 - Extração da fonte MySQL

Cada transformação é simples e direta. De cada tabela, ou conjunto de tabelas, da base de dados original são combinados (*joins*) e selecionados os dados de modo a inserir numa tabela de *cleanup*, que já tem os tipos e atributos iguais aos do *Data Warehouse*. Chama-se à atenção apenas para a transformação correspondente às vendas, onde já se transforma o identificador da data para um inteiro do tipo 20180205 (exemplo).

Passando para a extração da base de dados implementada em MongoDB, têm-se as seguintes transformações:

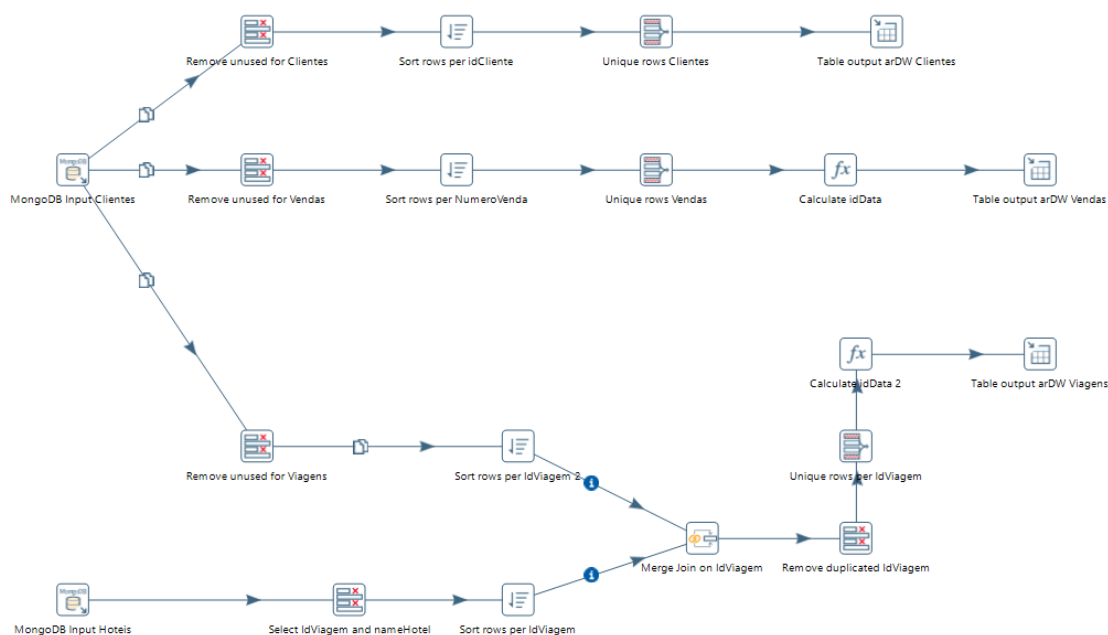


Figura 13 - Extração da fonte MongoDB

Aqui o cenário já se complica, devido à natureza das bases de dados não relacionais. Ainda assim, para a tabela de *cleanup* da área de retenção correspondente aos clientes os passos são simples: selecionam-se os atributos requeridos do cliente e removem-se os duplicados, antes de inserir na tabela da AR. Esta duplicação surge da necessidade de fazer desagregação dos documentos que têm como atributo outros documentos. Para a tabela de *cleanup* das vendas os passos mantêm-se, com o *step* adicional da transformação do id da data num número inteiro no formato já explicado para a fonte de dados MySQL. Por fim, para a tabela das viagens é necessária a junção de dados de duas fontes (fonte com dados referentes aos hotéis e fonte com dados referentes aos clientes). As transformações passam por seleccionar os campos requeridos para o DW das duas fontes, seguido de uma operação de *join*. Posto isto, segue-se uma nova seleção de campos, eliminação de duplicados, substituição do identificador da data pelo número inteiro habitual e inserção na tabela correspondente da

área de retenção. A junção acontece por junção de clientes e hotéis, uma vez que a cada hotel estão atribuídas as viagens vendidas.

No que toca à fonte de dados Excel, têm-se as seguintes transformações:

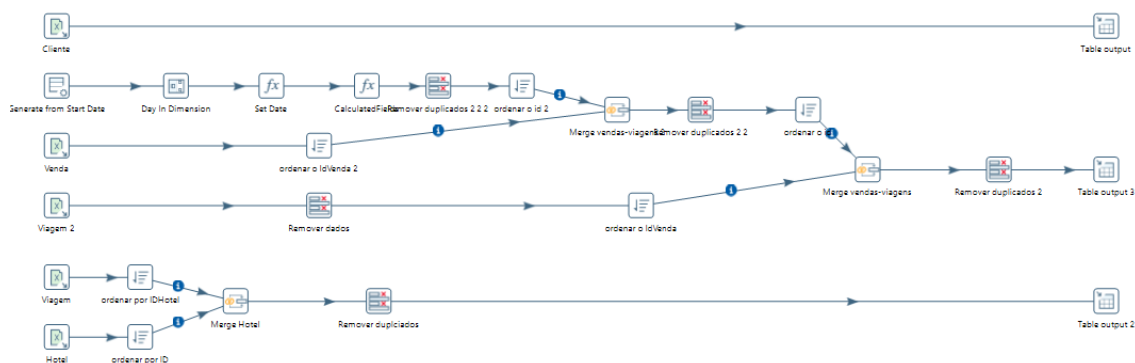


Figura 14 - Extração da fonte Excel

Mais uma vez, o cenário parece complicado. Começa-se pela extração dos clientes. Esta transformação é direta, e todos os dados são extraídos para a respetiva tabela da área de retenção. Para a transformação abaixo, que insere na tabela de *cleanup* das viagens os dados provenientes da fonte, foi necessário unir os dados relativos a hotel e viagem, passando depois por um processo de ordenação, *merge* e remoção de duplicados, graças à natureza desta fonte de dados. Por fim, para obter os dados relativos às vendas, e posterior inserção na respetiva tabela na área de retenção, usaram-se dados da venda e das viagens, onde os passos são os típicos: ordenações, remoções de duplicados, *merges*, entre outros. Chama-se só a atenção para a obtenção do id numérico das datas de forma diferente das outras fontes, que não se deveu a mais senão a uma decisão de implementação diferente.

Por fim, as datas têm um processo diferente da maioria das restantes dimensões. Assim, são geradas datas e os campos necessários à dimensão, sendo que estes dados vão para a área de retenção e não sofrem absolutamente mais nenhuma alteração, viajando quase diretamente para o *Data Warehouse* final. Isto deve-se ao facto de ser uma dimensão sem variação e sem qualquer tipo de falhas, uma vez que as datas foram geradas automaticamente.



Figura 15 - Processo de geração de datas

Terminada a fase de carregamento, consta a fase de **conformidade**. Não há muito a demonstrar, uma vez que esta apenas ocorre para o nome do cliente proveniente do Excel, que é necessário concatenar.

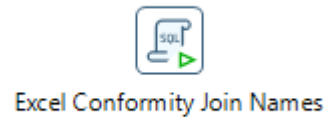


Figura 16 - Processo de conformidade

Assim, passa-se para a fase de **conciliação**. Para a inserção dos dados tratados nas dimensões, as transformações passam pela união de dados únicos, que advêm das três fontes, e posterior inserção ou *update* do valor na dimensão. Caso este processo “corra mal”, e apareça um nulo não previsto, este valor não será inserido no DW, e os dados vão para a quarentena, onde aguardarão tratamento. A junção acontece pelo NIF, Profissão e Local, uma vez que a manutenção de histórico é do tipo 2, pelo que interessa diferenciar por estes atributos. O NIF de um Cliente nunca muda, e na mudança de local ou profissão, uma nova entrada vai ser registada. No caso das viagens, é feita a comparação pelos seus atributos, que acabam por definir uma viagem de forma muito específica, de acordo com os pacotes que as agências de viagem normalmente criam.

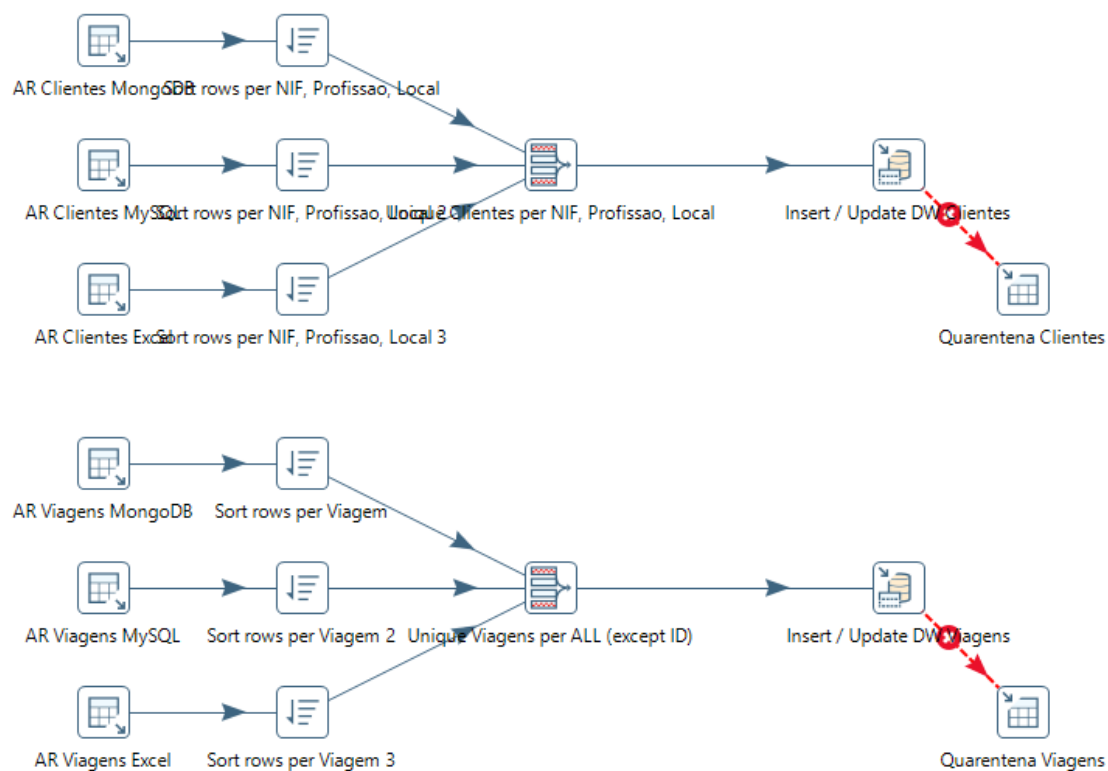


Figura 17 - Processo de conciliação para as dimensões

Para a conciliação referente à tabela de factos, o processo é mais extenso, já que exige mais cuidados. Assim, começa-se pelo processo de conciliação por parte dos dados da fonte MySQL. Antes de mais, é importante dizer que este processo é válido, quer para um povoamento inicial, quer para um refrescamento. Sendo assim, vai-se buscar informação das tabelas de cleanup da área de retenção, quer das vendas, clientes ou viagens (novas viagens a inserir; no caso do povoamento inicial são todas). Todos os passos seguintes envolvem os típicos steps de *merge*, *sort* ou remoção de duplicados, sendo que, no fim, tal como nas dimensões, algum caso não previsto é inserido na tabela de quarentena respectiva. O alvo de maior cuidado é precisamente a seleção de vendas apenas não inseridas. Para isto, guardam-se todas as vendas anteriores desta fonte numa tabela específica para isso e faz-se a diferença entre o novo carregamento e o anterior.

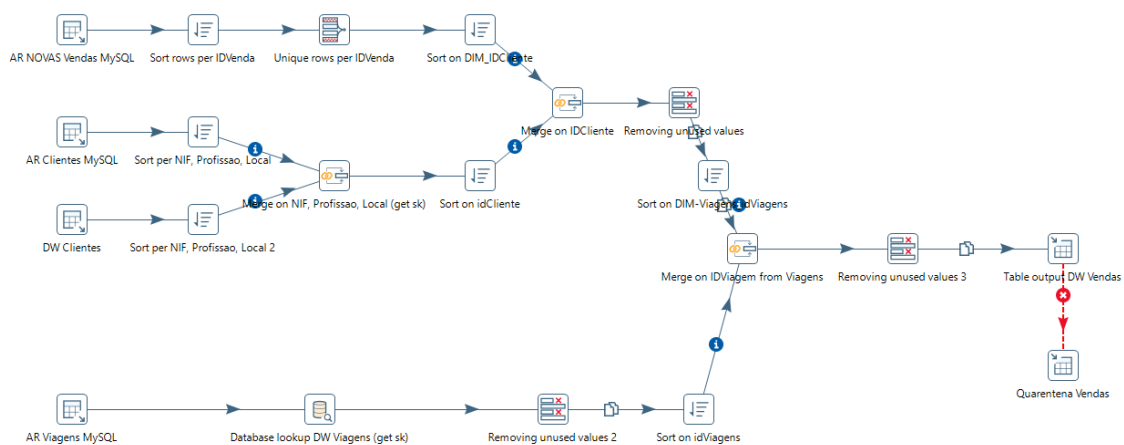


Figura 18 - Processo de conciliação para a tabela de factos - MySQL

Passando para a conciliação das vendas provenientes do MongoDB, o processo é muito idêntico ao anterior, não sendo necessária nenhuma explicação adicional.

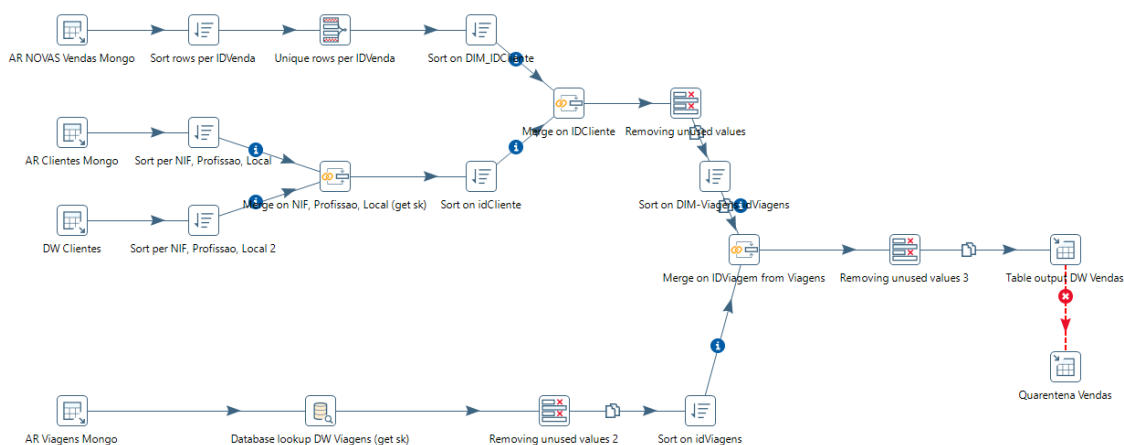


Figura 19 - Processo de conciliação para a tabela de factos - MongoDB

Para os dados fornecidos pelo Excel, a lógica mantém-se, e, mais uma vez, não se tecem quaisquer comentários adicionais.

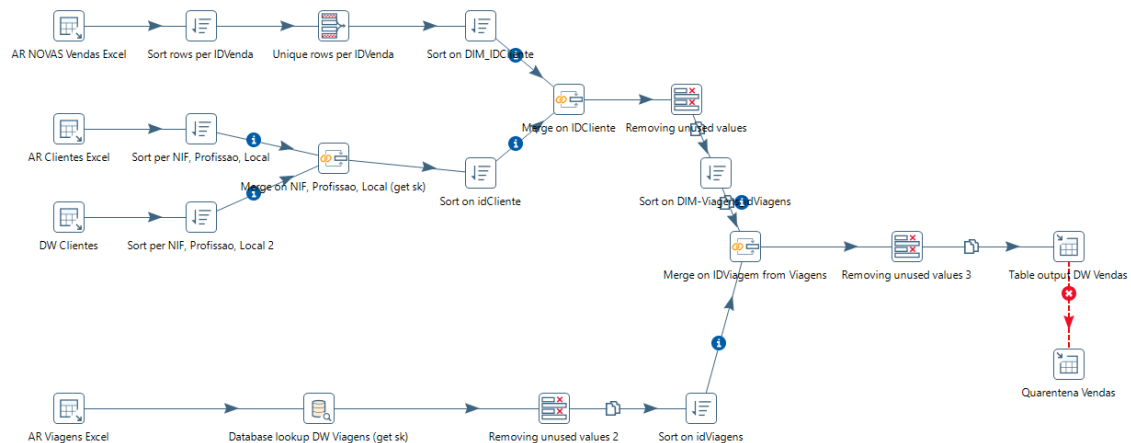


Figura 20 - Processo de conciliação para a tabela de factos - Excel

Todo o processo de ETL está agora explicado. Falta apenas mencionar dois detalhes necessários para o bom funcionamento destes processos no caso de um refrescamento. Assim, antes de mais, é executado um script para limpeza dos dados da área de retenção no fim de cada povoamento.



Figura 21 - Limpeza da área de retenção

Posteriormente, todas as vendas que foram para o Data Warehouse são mantidas em tabelas especiais na área de retenção, para ajudar no processo de conciliação e atribuição de chaves substituição. O próximo script sugere o armazenamento destes dados na AR. (Isto acontece apenas para a dimensão referente às vendas e não para os clientes e viagens devido ao tipo de tabela.)

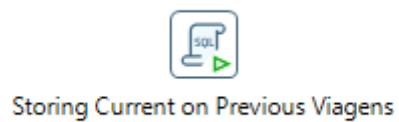
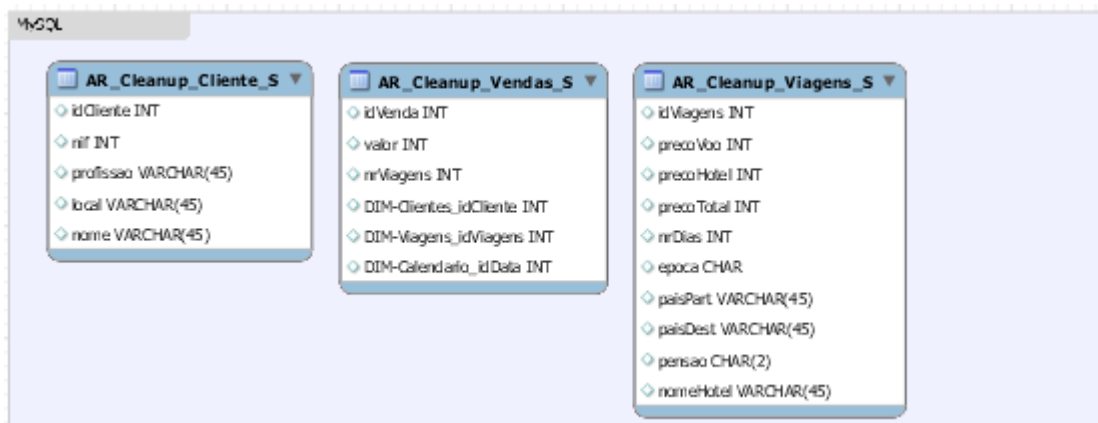


Figura 22 - Armazenamento das últimas vendas na área de retenção

6.5. Descrição e Caracterização de todos os Elementos de Dados

Para dar suporte ao ETL, a área de retenção (*staging area*) desempenha um papel essencial, sendo que uma boa organização do esquema de dados é fundamental para o suporte dos dados.

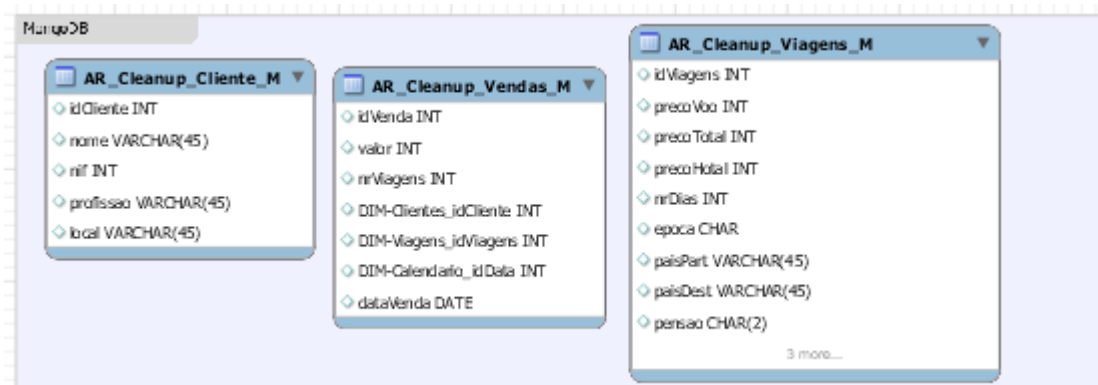
Deste modo, para armazenar os dados da limpeza e conformidade, seguem-se as três áreas distintas. A primeira é responsável por guardar temporariamente os dados limpos provenientes da fonte MySQL. A segunda tem como objetivo o armazenamento dos dados limpos que chegam da fonte MongoDB. Já a terceira zona guarda não só dados limpos do Excel, mas também os resultantes da conformidade.



MySQL

AR_Cleanup_Cliente_S	AR_Cleanup_Vendas_S	AR_Cleanup_Viagens_S
idCliente INT	idVenda INT	idViagens INT
nif INT	valor INT	precoVoo INT
profissao VARCHAR(45)	nrViagens INT	precoHotel INT
local VARCHAR(45)	DIM-Clientes_idCliente INT	precoTotal INT
nome VARCHAR(45)	DIM-Viagens_idViagens INT	nrDias INT
	DIM-Calendario_idData INT	epoca CHAR
		paisPart VARCHAR(45)
		paisDest VARCHAR(45)
		pensao CHAR(2)
		nomeHotel VARCHAR(45)

Figura 23 - Tabelas de cleanup da fonte MySQL



MongoDB

AR_Cleanup_Cliente_M	AR_Cleanup_Vendas_M	AR_Cleanup_Viagens_M
idCliente INT	idVenda INT	idViagens INT
nome VARCHAR(45)	valor INT	precoVoo INT
nif INT	nrViagens INT	precoTotal INT
profissao VARCHAR(45)	DIM-Clientes_idCliente INT	precoHotel INT
local VARCHAR(45)	DIM-Viagens_idViagens INT	nrDias INT
	DIM-Calendario_idData INT	epoca CHAR
		paisPart VARCHAR(45)
		paisDest VARCHAR(45)
		pensao CHAR(2)
		3 more...

Figura 24 - Tabelas de cleanup da fonte MongoDB

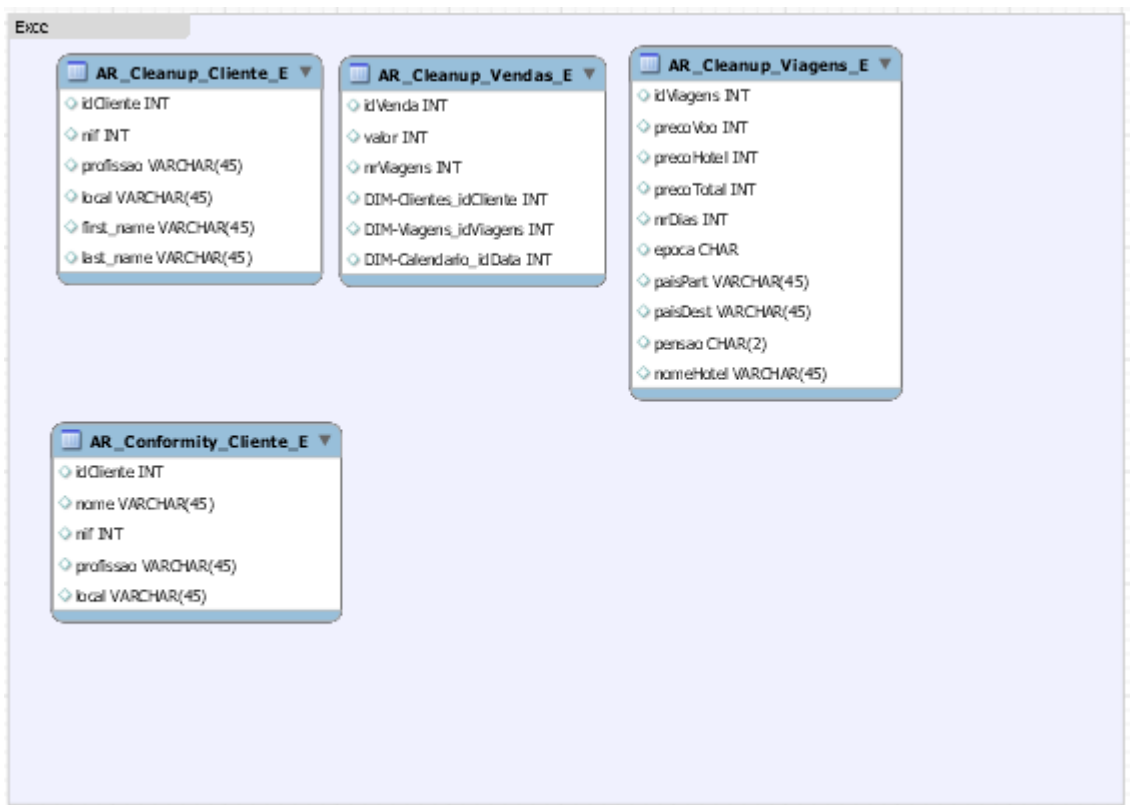


Figura 25 - Tabelas de cleanup e conformidade da fonte Excel

Para a quarentena, no caso se algum valor nulo imprevisto surgir, tem-se as quatro tabelas seguintes, iguais às dimensões do Data Warehouse. Os dados permanecem nesta tabela até ao próximo refresco, altura em que já estarão corrigidos por uma entidade externa.

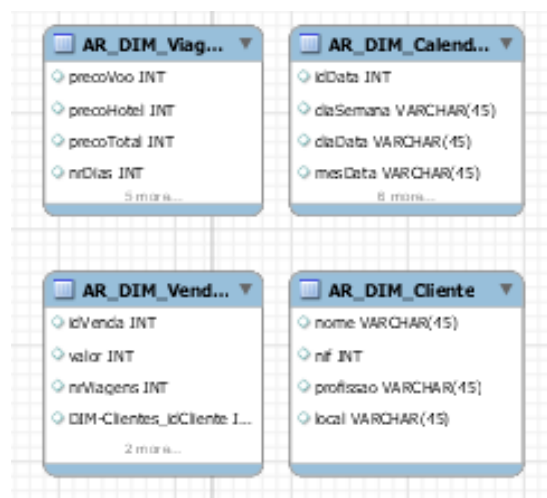


Figura 26 - Tabelas de quarentena

Por fim, restam três tabelas na área de retenção. Ora, estas estruturas auxiliam o processo de conciliação e posterior geração de *surrogate keys* no DW.

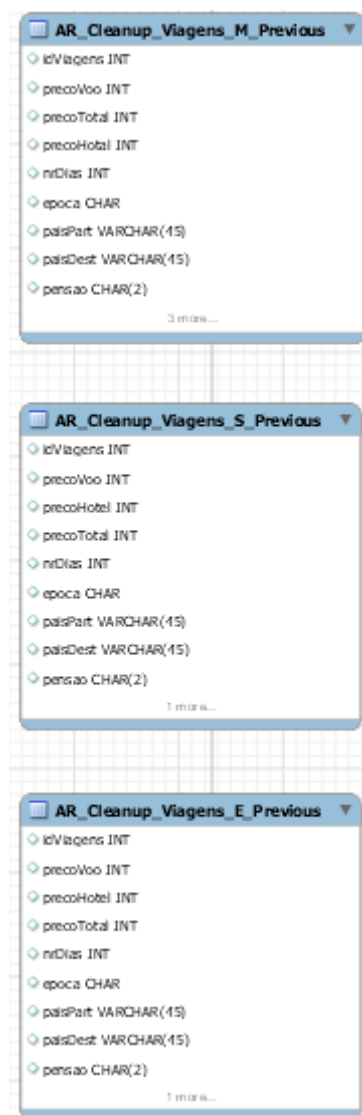


Figura 27 - Tabelas de conciliação

7. Implementação do Sistema de Data Warehousing

Tendo sido concluída a fase de análise de dados e de modelação, iniciou-se o desenvolvimento das estruturas físicas propostas. Para esta fase, teve de se decidir sobre quais as ferramentas a utilizar, de acordo com as necessidades correntes.

7.1. Escolha das Plataformas Computacionais

De forma a enumerar e explicar as plataformas e ferramentas computacionais escolhidas, vai-se seguir a linha de trabalho do grupo desde o início, onde se criaram as fontes de dados a analisar. Assim, para a primeira fonte de dados, usou-se a ferramenta *Edraw* para desenhar o modelo conceptual, e o *MySQL Workbench* para criar o esquema lógico e físico da base de dados, assim como para a povoar. Para a segunda fonte, usou-se o *MongoDB Compass* e o *Robot3D* de forma a implementar a base de dados e visualizar e manipular os dados. Por fim, usou-se o *Microsoft Office Excel* de modo a implementar esta BD não relacional.

Passando para a modelação do esquema dimensional, utilizou-se, mais uma vez, o *MySQL Workbench*. O facto de o grupo já estar ambientado com esta ferramenta cria uma tendência para o uso da mesma.

No que toca à modelação do processo de ETL, usou-se o *Bizagi* para elaborar diagramas BPMN. Assim, consegue-se uma maior abstracção, domando a complexidade para estre processo, e é mais fácil compreender os passos necessários para a implementação concreta posterior. De modo a suportar o ETL, criou-se uma área de retenção (*staging area*) no *MySQL Workbench*.

Passando para a implementação das fases do ETL, o *Pentaho Data Integration (Kettle)* foi de grande utilidade, já que este permite as conexões a todos os esquemas (fontes de dados, área de retenção e *Data Warehouse*) e transição de dados entre eles. Para além disso, abstrai o processo de transformação para um utilizador externo através de *jobs*, *transformations* e *hops*.

Finalmente, e depois de povoado o DW, foi utilizado o *Microsoft Power BI*, de modo a visualizar os dados e tirar alguma informação útil dos mesmos. Esta ferramenta de análise e visualização de dados permite avaliar os dados de uma forma interativa e visual, com a vantagem da partilha de *dashboards* entre todos os elementos do grupo.

7.2. Implementação dos Esquemas Físicos dos Sistemas de Dados

Foram implementados dois esquemas físicos durante o processo de construção do Data Warehouse. O primeiro foi construído para criar a estrutura da área de retenção e o segundo para criar a estrutura do Data Warehouse.

Para a implementação dos esquemas físicos foi utilizado a ferramenta disponibilizada pelo MySQL Workbench, forward engineer, que recebendo um modelo lógico devolve o schema correspondente e o seu script de criação.

7.2.1 O Sistema de Dados do Data Warehouse

Inicialmente, para ser possível criar o modelo físico foi realizado o seguinte modelo lógico do Data Warehouse. Este contém quatro tabelas diferentes. A tabela TF-Vendas, que representa a tabela de factos, é responsável por interligar as dimensões do modelo através de chaves estrangeiras e assim representar as transações realizadas (vendas de pacotes de viagens). As tabelas DIM-Clientes, DIM-Viagens e DIM-Calendário representam as dimensões do nosso Data Warehouse que armazenam os registos descritivos referentes aos factos.

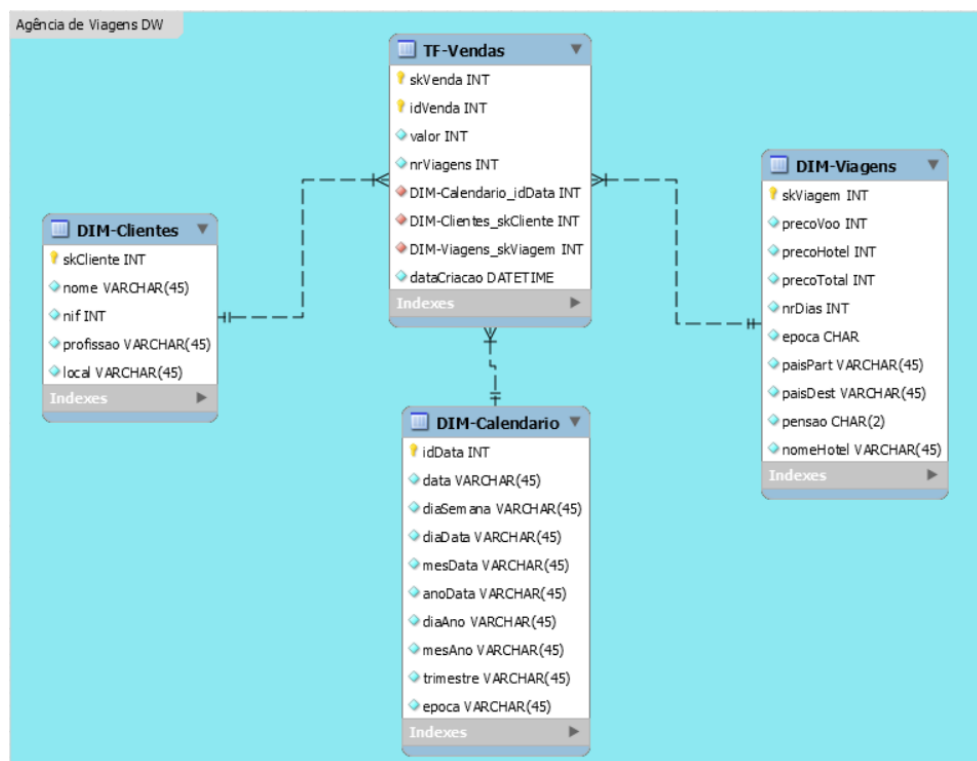


Figura 28 - Modelo Dimensional

Após a criação deste modelo foi utilizada a ferramenta forward engineer e assim estabelecido o seu modelo físico.

7.2.2 O Sistema de Dados da Área de Retenção

A área de retenção do sistema do grupo é constituída por uma base de dados desenvolvida em *MySQL*, esta possui um conjunto de tabelas específicas que suportam as operações de *transformação* dos dados provenientes do processo de *extração* das fontes.

É importante referir que nesta base de dados não foram consideradas restrições de integridade, uma vez que se trata de uma *staging area* cujo principal objetivo é auxiliar na transformação dos dados antes de serem carregados para o *Data Warehouse*.

O esquema lógico concebido está de acordo com as tabelas apresentadas na secção 6.5.

Como a base de dados foi desenvolvida em *MySQL*, o processo de transição do modelo lógico para o físico foi conseguido através da funcionalidade *forward engineering*.

7.3. Implementação do Sistema de Povoamento

Após a implementação dos esquemas físicos dos sistemas de dados do Data Warehouse e da área de retenção, é altura de criar os processos idealizados na ferramenta Pentaho.

7.4. Análise da Execução do Sistema de Povoamento

O sistema de povoamento do Data Warehouse é composto por duas etapas, nomeadamente, o povoamento inicial e as periódicas atualizações da informação (refrescamentos).

Todo o processo foi implementado com o recurso à ferramenta Pentaho (Kettle), desde a extração dos dados das fontes para área de retenção até ao carregamento final para o Data Warehouse.

Inicialmente, procedeu-se à recolha dos dados das várias fontes, tomando por base uma política estática. Este processo requer a utilização de algumas técnicas de manipulação de dados (eliminação de repetidos, junções, etc.), tendo em conta a desnormalização verificada em algumas fontes.

O processo de transformação será suportado pelas tabelas da área de retenção e executado através das funcionalidades do Kettle. Começando pela operação de limpeza, serão criadas tabelas que recebem os dados das vendas, dos clientes e das viagens que sofreram alterações no processo de verificação da existência de valores do tipo *null*, para os atributos em que era previsto tal acontecer.

Na operação de *conformidade*, será criada uma tabela com o objetivo de auxiliar a alteração do nome dos clientes provenientes da base de dados *Excel*, garantindo que se encontra de acordo com o formato pretendido, ou seja, o primeiro e último nome devem ser concatenados.

Para garantir a consistência dos dados na eventualidade da ocorrência de falhas no processo de carregamento, serão utilizadas tabelas de quarentena, que servirão para armazenar esses dados para, posteriormente, serem processados de novo.

Será criada uma tabela para agregar a informação relativa ao calendário, proveniente de uma fonte externa, conforme os atributos da dimensão tempo do Data Warehouse. Este processo será feito, igualmente, no Pentaho e apenas no povoamento inicial.

Finalmente, o processo de carregamento dos dados da área de retenção para o Data Warehouse, será feito de modo idêntico ao referido em etapas anteriores, ou seja, utilizando métodos de *merge* e *sort* é possível agregar os dados no seu destino final. Nesta etapa seguimos uma política incremental, na medida em que só serão carregados os dados novos ou os que foram alterados, por comparação às vendas anteriormente inseridas, guardadas em tabela própria na área de retenção.

8. Business Intelligence

Uma vez implementado e povoado o *Data Warehouse*, o grupo usou o *Microsoft Power BI* de modo a visualizar os dados e tirar informação útil dos mesmos. Esta ferramenta de análise e visualização de dados permite avaliar a informação de uma forma interativa e visual, com a vantagem da partilha de dashboards entre todos os elementos do grupo. Desta forma, com o objetivo de responder às perguntas já definidas, o grupo criou um relatório referente ao datamart existente. Segue-se cada pergunta a ser respondida, juntamente com o gráfico obtido e uma breve explicação.

1. - Qual o número médio de vendas por mês?

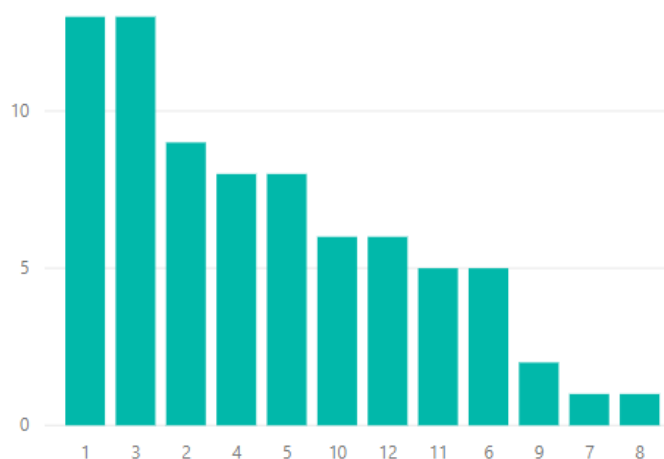


Figura 29 - Média do número de vendas por mês

É possível observar, através do gráfico, que os meses com um maior volume de vendas são os de Janeiro (1) e Março (3).

2. – Qual é o número de vendas por país de destino?

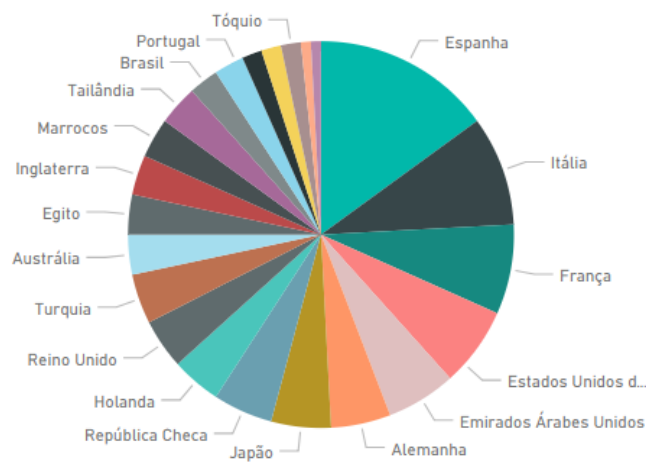


Figura 30 - Número de vendas por país destino

De acordo com o diagrama, o país destino favorito dos clientes é Espanha, seguido por Itália.

3. – Qual é o número de vendas por local de cliente?

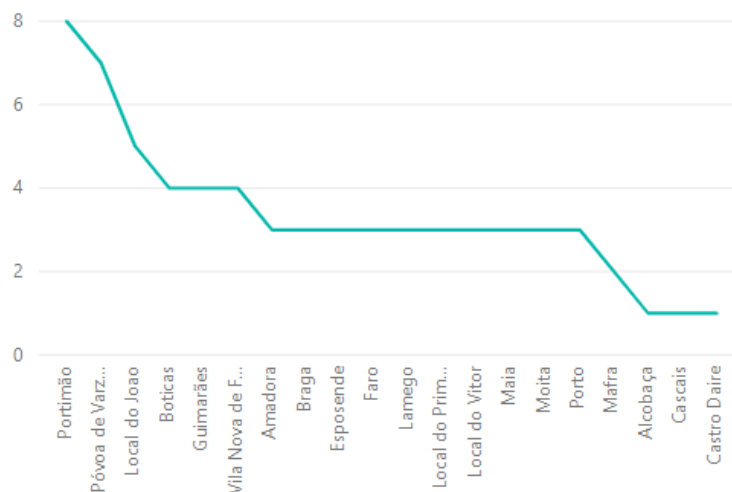


Figura 31 - Número de vendas por local

Após a leitura do esquema chega-se à conclusão que a localidade que realizou um maior número de vendas foi a de Portimão, com oito.

4. – Qual é número de vendas por cliente?

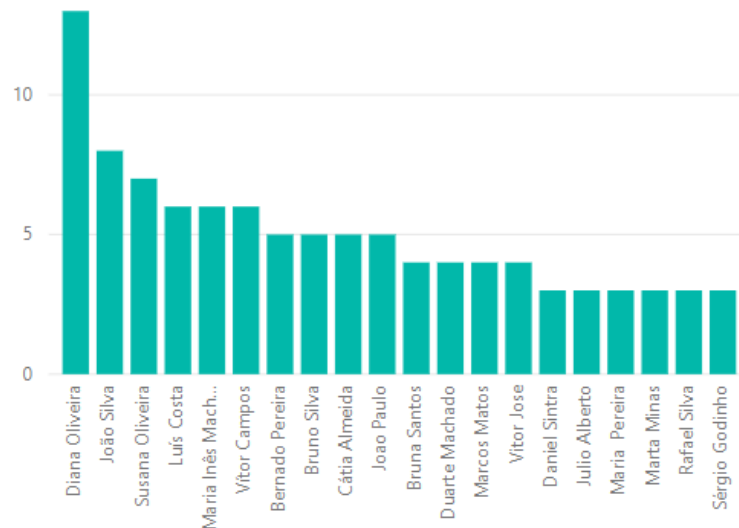


Figura 32 - Número de vendas por cliente

Após a análise do diagrama, é possível concluir que a cliente Diana Oliveira foi aquela que realizou mais compras nas três agências.

Em suma, todos os indicadores criados permitiram responder às perguntas colocadas no início do projeto com bastante clareza. Para além disso, estas revelaram-se de bastante utilidade na descoberta de tendências que, numa base de dados operacional, dificilmente seriam visíveis.

9. Conclusões e Trabalho Futuro

A ideia de criar um Data Warehouse que servisse de apoio à decisão da agência de viagens 'Belo Mundo', deixando ao critério do grupo esta escolha do tema e todo o trabalho envolvido na criação e fundamentação do mesmo foi, sem dúvida, desafiante. Foi possível constatar que todo o esforço envolvido aquando do surgimento e definição de uma ideia exige tanto ou mais trabalho quanto pô-la em prática. Na primeira fase, apenas foi necessário elaborar um modelo geral da ideia a desenvolver, esclarecer alguns pontos e medidas de sucesso e viabilidade, e analisar os recursos necessários associados ao sistema de decisão. Ainda assim, sem uma boa de fundamentação e sem o estabelecimento de metas, todos os passos futuros acabariam por levantar demasiadas dúvidas e poderiam acabar por divergir, tanto da ideia inicial, como entre os elementos do grupo. Vê-se aqui, também, a importância da definição precoce do grão neste tipo de sistemas, uma vez que, sem isto, nada é implementado coerentemente.

Já numa fase posterior, foi requerido que fossem especificados os requisitos de descrição, exploração e controlo, para além da modelação dimensional e análise das fontes. O grande desafio consistiu na descrição completa e detalhada de todos os requisitos, com a consequente análise da área de negócio, uma vez que estes são a base de todo o sistema. O grupo viu-se, com o passar do tempo, a refazer e a aperfeiçoar estes tópicos. Uma vez estes concluídos, foram elaborados diagramas (BPMN) para modelação do processo de ETL, sendo que a equipa optou por este alto nível, de forma a domar a complexidade que este processo pode atingir. Tanto os diagramas como a implementação em *Pentaho* fluíram naturalmente, uma vez que já havia uma boa base para o projeto. Ainda assim, a dificuldade destas etapas foi bastante elevada, uma vez que a equipa não possuía qualquer conhecimento na área, e viu-se obrigada a tomar decisões, primeiramente, da constituição de processos de ETL, e, consequentemente, de construção de estruturas de apoio que suportassem.

Na terceira e última fase, o grupo procedeu ao teste e análise dos dados obtidos, que é o grande objetivo do desenvolvimento de sistemas de suporte à decisão. O Power BI revelou-se bastante útil para esta etapa, pois permitiu ao grupo responder com sucesso a todos os indicadores propostos com bastante clareza. Para além disso, foram descobertas tendências interessantes que, numa base de dados operacional, dificilmente saltariam à vista.

Com a realização deste projeto, foi adquirida experiência relevante para a vida profissional futura dos membros do grupo. De facto, as dificuldades foram muitas, especialmente nas tomadas de decisão durante o processo de ETL. O grupo considera que

este foi dos trabalhos mais complexos que teve. Ainda assim, a equipa tirou proveito da oportunidade de colocar em prática os conhecimentos teóricos adquiridos anteriormente, através da realização de um Data Warehouse e dentro de um contexto real. A abordagem simplista permitiu concluir com sucesso este trabalho e reconhece-se que, caso se tivesse optado por uma abordagem mais alargada e realista, as dificuldades ascenderiam de forma radicalmente vertical. Num futuro, com mais conhecimento e experiência na área, provavelmente seriam tomadas outras decisões de implementação mais eficientes, úteis num projeto de maior dimensão.

Referências

- Data Warehouse Design: Modern Principles and Methodologies de Golfarelli, M., Rizzi, S.;
- The Data Warehouse Lifecycle Toolkit – Practical Techniques for Building Data Warehouse and Business Intelligence Systems de Kimball, R., Reeves, L., Ross, M., Thornthwait, W.

Lista de Siglas e Acrónimos

BD	Base de Dados
DW	Data Warehouse
OLTP	<i>On-Line Analytical Processing</i>