Universidade do Minho
Mestrado Integrado em Engenharia Informática
Mestrado em Engenharia Informática

Unidade Curricular de Data Warehousing

Ano Letivo de 2018/2019

Sistema de Data Warehousing de aluguer de automóveis

Adriana Guedes, Diogo Soares, João Cabo, José Silva

janeiro, 2019

DW

Data de Receção	
Responsável	
Avaliação	
Observações	

Sistema de Data Warehousing de aluguer de automóveis

Adriana Guedes, Diogo Soares, João Cabo, José Silva

janeiro, 2019

Resumo

Um sistema de *data warehousing* (SDW) pode ser definido como um sistema de informação especializado e especialmente desenhado para servir de suporte a decisões, do gestor, ou outro tipo de entidade encarregue de tomar decisões. Os sistemas de data warehousing assumem um papel essencial no planeamento de atividades de uma empresa, pois, devido à sua especificação, originam soluções otimizadas que são importantes no momento de tomada de decisão. Neste documento pretende-se, relatar de forma detalhada o desenvolvimento de um sistema de *data warehousing* de uma empresa de aluguer de automóveis, em particular, descrever a análise, o planeamento, a conceção e implementação do sistema, realizar uma análise dimensional dos dados, e, relatar a modelação e caracterização de sistemas de extração, transformação e integração dos dados num *data warehouse* (DW).

Área de Aplicação: Desenho e arquitetura de Sistemas de *Data Warehousing*
Palavras-Chave: Sistemas de Suporte à Decisão, Sistemas de *Data Warehousing*, *Data Warehouses*, Bases de Dados Relacionais, *NoSQL, MySQL, Neo4j, .csv*.

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1 Contextualização

Desde os primeiros registos de deslocação humana até aos dias de hoje, sempre foi possível verificar um grande investimento na procura por métodos e meios que tornem as viagens o mais eficiente e cômodas possível. O resultado é constantemente apresentado no nosso dia-a-dia, quer nas vastas redes de estradas que ligam cidades e países, às enormes redes férreas e ligações aéreas, que permitem uma rápida mobilização de bens e pessoas pelo mundo inteiro de forma prática.

No caso concreto da empresa *Belos Automóveis*, é fornecido aos seus clientes um serviço de aluguer de veículos com um único estabelecimento físico na cidade de Braga, localizado junto à estação de comboios. Este local tornou-se uma referência no aluguer de veículos a turistas que chegam de todos os cantos do mundo e procuram uma forma de se mobilizarem facilmente pela cidade, assim como para empresas que procuram uma frota de veículos ideal para o seu negócio, sem a necessidade de recorrer à sua compra.

Um aluguer é relativo a um único veículo e possui um custo fixo diário, associado ao mesmo veículo. O preço final do aluguer é dependente da duração do mesmo em número de dias, definido na realização do contrato.

A crescente evolução tecnológica que acompanhamos nos últimos anos, leva a que este serviço tenha sentido a necessidade de se adaptar, criando recentemente uma plataforma online que permite o aluguer de veículos através do preenchimento de um formulário no website, e que, após o respetivo pagamento, gera uma referência de levantamento do veículo.

O aluguer de veículos de luxo está apenas disponível no estabelecimento físico diretamente com o gerente que possui os registos de aluguer deste tipo de veículos.

1.2 Motivação e Objetivos

Para uma empresa, é importante esboçar uma estratégia de desenvolvimento e inovação de negócio, de forma a que consiga tirar o máximo proveito das oportunidades que surgem no mercado. Um ponto importante para o sucesso passa pela correta definição de estratégias, pois para alcançar o sucesso, é necessário tomar boas decisões.

A empresa *Belos Automóveis,* por questões de organização, apresenta a informação recolhida em várias bases de dados. Como consequência, houve a necessidade de centralizar e armazenar os seus dados, de forma a que, nas tomadas de decisão, possam ser analisados todos os dados que sejam considerados relevantes. Assim, a empresa DWing foi contactada, com o intuito de criar um sistema de informação que vá auxiliar na identificação de características dos veículos mais populares nos alugueres realizados, de forma a conseguirem gerir os veículos que têm à disposição tendo em conta as preferências dos clientes, bem como de onde provêm os seus clientes, de forma a investir em publicidade nesses mesmos locais ou posterior abertura de uma filial, tendo em vista o aumento do lucro da empresa. O grande objetivo é que, através do d*ata warehouse* desenvolvido, seja possível identificar aspetos que influenciam o aumento do número de clientes da *Belos Automóveis.* Assim, surge a necessidade de serem criados mecanismo de acesso aos dados, fazendo o tratamento dos mesmos, permitindo no final apresentar os dados que são realmente relevantes para a tomada de decisão.

1.3 Viabilidade da implementação do sistema de data warehousing

A empresa *Belos Automóveis* conta no seu sistema com três bases de dados diferentes, uma base de dados relacional que contém a informação relativa a alugueres ocorridos em stands, uma base de dados não relacional dos alugueres efetuados no site da empresa e uma base de dados csv dos alugueres de carros de luxo destinada apenas ao gerente, visto que é o único com permissões a tratar da transação deste tipo de carros.

Posto isto, surgiu a necessidade de desenvolver um *data warehouse* para acompanhar e auxiliar as tomadas de decisão do gestor da empresa. É necessário analisar os alugueres de forma a estruturar uma estratégia que, consequentemente, leve ao aumento do lucro da *Belos Automóveis*.

Com a informação recolhida, e devidamente processada, será possível à empresa elaborar estratégias de *marketing* apropriadas para atrair clientes, com base na análise das informações relativas aos alugueres. Através dos dados fornecidos pelo *data warehouse* será possível identificar, por exemplo, a cidade mais comum onde residem os nossos clientes, e utilizar como fator de decisão para o investimento de publicidade. Será também possível averiguar o tipo de carros mais procurados ou as épocas onde os serviços são mais procurados, podendo assim regular o inventário dos veículos com o objetivo de lucro.

2. Planeamento do projeto

Este capítulo é direcionado para as questões relacionadas com o planeamento do projeto, isto é, pelas pessoas responsáveis pelo desenvolvimento do d*ata warehouse*.

2.1 Definição da identidade do projeto

Como todos os projetos precisam de uma identidade, após uma reunião com todos os responsáveis do projeto acordou-se que o nome seria DWing.

Figura 1 - Imagem da empresa

2.2 Identificação dos recursos necessários

Para a realização deste projeto é fundamental possuir os recursos necessários, pois sem estes nunca seria possível realizá-lo. Dito isto, é então imprescindível cada elemento do projeto ter uma máquina com o *software* Pentaho Data Integration, *Word*, *MySQL* e *Neo4j*.

2.3 Plano de desenvolvimento

Preparar o plano de desenvolvimento é um dos aspetos mais importantes que a equipa teve em conta, pois um bom planeamento levará a uma maior taxa sucesso do projeto que está em causa. Torna-se então imprescindível perceber as tarefas que são necessárias à realização do mesmo e, consoante a dificuldade destas, regular o tempo despendido para cada tarefa. Devemos ajustar o tempo sempre com um

período de conclusão de tarefa alargado, para prevenir atrasos que possam surgir. Tendo isto em conta, chegamos ao seguinte planeamento:

Tarefa	Data de Início	Data de Conclusão	Duração
Levantamento de requisitos	01-11-2018	08-11-2018	6d
Construção da matriz de decisão	09-11-2018	16-11-2018	6d
Seleção do data mart a desenvolver	17-11-2018	23-11-2018	6d
Escolha do Grão	24-11-2018	26-11-2018	2d
Escolha das dimensões de análise	25-11-2018	30-11-2018	6d
Desenvolvimento do diagrama das tabelas de factos	03-12-2018	06-12-2018	4d
Documentar as tabelas de facto	07-12-2018	14-12-2018	6d
Projetar o detalhe das dimensões	13-12-2018	20-12-2018	6d
Projeção do ETL com o modelo BPMN	04-01-2019	09-01-2019	4d
Construção do ETL com a ferramente kettel	09-01-2019	17-01-2019	7d
Documentação do processo ETL	17-01-2019	22-01-2019	4d

Figura 2 - Plano de desenvolvimento

2.4 Definição da equipa de desenvolvimento

Para o sucesso deste projeto foi necessário construir uma equipa capaz de dar resposta às dificuldades presentes na construção de um sistema de data warehouse. Assim, foi criada uma equipa constituída por quatro alunos que frequentam a Unidade Curricular de Data Warehousing.

3. Levantamento e Análise de Requisitos

3.1 Apresentação do método de aquisição de requisitos

Para ser possível perceber que funcionalidades e que tipo de informação o data warehouse deveria conter, foi necessário realizar, numa fase inicial, uma reunião com o nosso cliente, de forma a levantar os requisitos que seria necessário ter em conta. Além deste levantamento de dados, foram analisados também diversos serviços semelhantes de aluguer de automóveis e, por fim, foram realizadas várias entrevistas com os potenciais clientes que o serviço poderá ter.

Após uma análise dos requisitos levantados do cliente, conseguimos perceber que existem diversos requisitos que são essenciais para o sucesso do serviço, e que serão apresentados nas seções seguintes.

3.2 Requisitos levantados

Cada Cliente

- Possui informação sobre a sua profissão
- Possui informação sobre a sua cidade
- Possui informação sobre o seu país
- Pode possuir vários alugueres

Cada Veículo

- Tem informação sobre a marca
- Tem informação sobre o tipo de combustível
- Tem informação sobre o preço por dia
- Tem informação sobre o número de lugares
- Que se encontra no website não se encontra apresentado no stand

Cada Aluguer

- Possui um preço que não sofre alterações independentemente do dia da semana (fins-de-semana incluídos).
- Exige um novo contrato do cliente se pretender prolongar o aluguer do automóvel
- Tem um limite de dias de aluguer que são definidos na hora do aluguer.

- O número de dias de aluguer, após serem definidos no ato de aluguer, não é variável.
- O número mínimo de dias de aluguer é de um dia
- É relativo apenas a um automóvel
- Exige que o cliente se encontre registado no stand
- Realizado através do website é efetuado através de um formulário

A Empresa

- Possui um único stand existente e situa-se em Braga.
- Permite alugar veículos, quer no stand, quer através da sua página web.
- Possui os registos de todos os seus clientes na fonte relacional.

3.3 Revisão dos requisitos com os utilizadores

Após este levantamento, foi realizada uma apresentação com o responsável do stand onde foram apresentados e discutidos os diversos requisitos. Ao ser dada a aprovação por parte do cliente, partiu-se para a modelação dimensional do sistema

4. Modelação dimensional do Sistema

A modelação dimensional deve ser vista como a base de todos os processos de decisão do sistema. Nesta modelação é efetuado um estudo das necessidades que o negócio espera resolver com a implementação do *data warehouse*, assim como a realidade dos dados fornecidos pelas fontes.

Nesta etapa deve-se então desenvolver os esquemas dimensionais especializados no auxílio da tomada de decisões por parte dos agentes de decisão, da forma mais simples possível, fornecendo a informação pretendida relativamente aos requisitos de cada área de negócio envolvida.

4.1 Apresentação da metodologia de desenvolvimento

A metodologia de desenvolvimento utilizada foi a técnica dos *"4 passos"* descritos por *Kimball,* seguindo uma abordagem *bottom-up* na qual os processos do negócio são descritos permitindo **(1) construir a matriz de decisão**. Após a identificação das áreas de negócio para cada uma é efetuada a **(2) escolha do grão das tabelas de factos**, o menor nível de dados capturado por cada processo do negócio estudado. Após a definição do grão, a **(3) identificação e caracterização das dimensões** permite obter contexto descritivo através dos seus atributos sobre os quais queremos analisar os factos. Por fim são **(4) definidas as medidas dos factos**, ou seja, as medidas de um evento físico observável que pretendemos integrar em cada facto.

4.2 Definição e Caracterização dos *Data Marts*

De acordo com os requisitos do cliente foi possível chegar à implementação de um único *Data Mart* que apresenta ao utilizador uma visão global de todo o seu negócio. Neste *Data Mart* é possível encontrar facilmente toda a informação necessária para uma análise dos alugueres efetuados na loja e site da empresa Belos Veículos.

4.2.1 Esquematização da matriz de decisão

Caracterização de Data Mart Comercial	
Identificação: Comercial	
Descrição Geral: Informação para suporte à tomada de decisão na àrea dos alugueres da *"DWing"* fornecendo elementos de dados do aluguer de veiculos na loja e plataforma online para gestão das ações de marketing e da frota de veiculos.	
Estrutura Base	
Tabela de Factos >>	TF-Alugueres
<< Dimensões	
Calendário	**X**
Cliente	**X**
Veiculo	**X**
Número de Dimensões	3
Tipo	Transacional
Periodicidade	Diária
Descrição	Transações de alugueres de veiculos na empresa "Belos Veiculos".
Utilidade Estratégica	Avaliação do desempenho comercial de cada uma das plataformas de aluguer de veículos. Identificar e caracterizar stock de veículos para melhorar base de negociação com fornecedores. Estabelecer um perfil dos clientes para otimizar acções comerciais.
Utilizadores	
Gestor da empresa "Belos Veiculos" e gestores comerciais.	
Observações	
Nada a assinalar.	

Tabela 1- Matriz de decisão do Data Mart Comercial

4.2.2 Definição do grão

Visto apenas existir a necessidade de desenvolver uma tabela de factos TF-Alugueres, através do grão da mesma, definimos o detalhe das estruturas de dados. No caso estudado o grão corresponde ao **aluguer** de **um veículo** por parte de **um cliente** numa **certa data**.

4.3 Definição e caracterização das tabelas de factos

Na modelação dimensional a tabela de factos permite armazenar as informações relativas a um evento do processo de negócio definido, tendo uma linha do mesmo uma relação 1 para 1 com o evento descrito no grão. As características presentes devem ser a simplicidade que permitirá ao analista uma navegação fácil e rápida sobre os dados. A tabela de facto desenvolvida é relativa a um aluguer de um veículo por um dado cliente numa certa hora e plataforma, característica que torna a tabela de factos do tipo **transacional.**

4.3.1 Tabela de facto

Caracterização da tabela de factos						
Identificação	TF-Alugueres.					
Descrição	Tabela que acolhe todos os registos de aluguer de veiculos realizados na loja e plataforma online da "Belos Automóveis".					
Data Mart	Comercial.					
Tipo	Transacional.					
Utilidade estratégica	Avaliação do desempenho comercial de cada uma das plataformas de aluguer de veículos. Identificar e caracterizar stock de veículos para melhorar base de negociação com fornecedores. Estabelecer um perfil dos clientes para otimizar acções comerciais.					
Povoamento	Realizado diariamente entre a uma e as sete horas da manhã, iniciando-se preferencialmente às duas da manhã.					
Dimensão inicial	3.7KR					
Crescimento	0.5%/mês					
Periodo de dados	Ultimos 5 anos de dados de aluguer de veiculos. Os registos anteriores ficarão em arquivo fisico.					
Atributos						
Dimensões						
Nr	**Identificação**	**Chave**	**Tipo**	**Dominio**	**Descrição**	**Exemplo**
1	idCliente	S	V	String (45)	Código relativo ao cliente que efetuou o aluguer do veiculo.	123456
2	data_inicio	S	RP	Data	Código da data referente ao inicio do aluguer do veiculo.	2018/06/01
3	idVeiculo	S	V	String (45)	Código referente ao veiculo da loja "Belos Veiculos" alugado.	48-28-OV
4	data_fim	S	RP	Data	Código da data referente ao fim do aluguer do veiculo.	2018/06/01
Medidas						
Nr	**Identificação**	**Dominio**	**Tipo**	**Descrição**		**Exemplos**
1	numeroDias	Inteiro	A (avg)	Número de dias do aluguer - (data_fim - data_inicio)		7
2	preco	Float	A (sum)	Preço do aluguer do veiculo.		5,99
Indices						
Nr	**Identificação**	**Tipo**	**Descrição**			
1	idAluguer	Primária	Único, ordenado fisicamente de forma crescente.			
2	idCliente	Secundário	Ordenado de forma crescente.			
3	data_inicio	Secundário	Ordenado de forma crescente.			
4	data_fim	Secundário	Ordenado de forma crescente.			
5	idVeiculo	Secundário	Ordenado de forma crescente.			
Perfis de utilização						
Gestor da empresa "Belos Veiculos" e gestores comerciais.						
Observações						
Nada a assinalar.						

Tabela 2 - Caracterização da tabela de factos TF-Alugueres

4.4 Definição e caracterização das Dimensões

As dimensões fornecem os atributos necessários para suportar a perspetiva de análise numa dada categoria. Assim podemos explorar uma tabela de factos segundo várias perspetivas consoante a necessidade. Os atributos de cada medida são assim selecionados com base nas necessidades apresentadas pelos agentes de decisão.

Dimensões do Data Mart Comercial			
Nr	**Identificação**	**Descrição**	**Esquema**
1	Calendário	A dimensão Calendário é composta por atributos que sustentam análises temporais ao longo de um dia, mês, ano, etc.	dim_Calendario (com diferentes papeis)
2	Cliente	Identificação e caracterização dos clientes da empresa "Belos Veiculos"	dim_Cliente (com Variação), dim_Cliente_HST (Histórico) e periodicidade diária
3	Veiculo	Identificação e caracterização dos veiculos da empresa "Belos Veiculos"	dim_Veiculo (com variação), dim_Veiculo-HST (Histórico) e periodicidade diária

Tabela 3 - Descrição geral das dimensões

4.4.1 Dimensão Calendário

Caracterização da dimensão Calendário							
Identificação	dim_Calendario						
Descrição	Calendário do ano e os seus atributos						
Tipo	Com diferentes papéis *(role-playing dimension)*						
Dimensão	3.7 KR (registos gerados no povoamento)						
Crescimento	Não cresce. O povoamento desta dimensão é feito na fase de arranque do data warehouse para um periodo de 10 anos.						
Atributos							
Nr	**Identificação**	**Descrição**	**Chave (tipo)**	**Dominio (tamanho)**	**V / H / P**	**Variação**	**Exemplos**
1	idCalendario	Código de identificação da data.	S	Data	---	---	2018/06/01
2	mes	Número do mês.	N	Inteiro	---	---	6
3	ano	Número do ano.	N	Inteiro	---	---	2018
4	trimestre	Número do trimestre.	N	Inteiro	---	---	3
5	fimdesemana	Indicação se é fim de semana ou não.	N	String (3)	---	---	SIM
Indices							
Nr	**Identificação**	**Indice**	**Tipo**				
1	idCalendario	Primário	Único, ordenado fisicamente de forma crescente.				
Hierarquia							
Nr	**Identificação**	**Esquema**					
1	H1	idCalendario -> mes -> trimestre -> ano -> ALL					
2	H2	idCalendário -> fimdesemana -> ALL					
Perfis de Utilização							
Gestor da empresa "Belos Veiculos" e gestores comerciais.							
Observações							
Nada a assinalar.							

Tabela 4 - Caracterização da dimensão Calendário

4.4.2 Dimensão Cliente

Caracterização da dimensão Cliente							
Identificação	dim_Cliente						
Descrição	Caracterização dos clientes da empresa "Belos Veiculos"						
Tipo	Com variação						
Dimensão	10R						
Crescimento	10%/Ano						
Atributos							
Nr	**Identificação**	**Descrição**	**Chave (tipo)**	**Dominio (tamanho)**	**V / H / P**	**Variação**	**Exemplos**
1	idCliente	Código de identificação do cliente.	S	String (45)	---	---	12345
2	profissao	Profissão do cliente.	A	String (45)	S / S / ?	Tipo 4	CAMIONISTA
3	cidade	Cidade da morada do cliente.	A	String (45)	S / S / ?	Tipo 4	BRAGA
4	pais	Pais da morada do cliente.	A	String (45)	S / S / ?	Tipo 4	PORTUGAL
5	idade	Idade do cliente.	A	Inteiro	S / N / A	Tipo 1	23
Indices							
Nr	**Identificação**	**Indice**	**Tipo**				
1	idCliente	Primário	Único, ordenado fisicamente de forma crescente.				
2	profissao	Secundário	Ordenado de forma crescente.				
3	cidade	Secundário	Ordenado de forma crescente.				
4	pais	Secundário	Ordenado de forma crescente.				
5	idade	Secundário	Ordenado de forma crescente.				
Hierarquia							
Nr	**Identificação**	**Esquema**					
1	H1	idCliente -> profissao -> ALL					
2	H2	idCliente -> cidade -> pais -> ALL					
3	H3	idCliente -> idade -> ALL					
Perfis de Utilização							
Gestor da empresa "Belos Veiculos" e gestores comerciais.							
Observações							
Nada a assinalar.							

Tabela 5 - Caracterização da dimensão Cliente

4.4.3 Dimensão Veículo

Caracterização da dimensão Veiculo							
Identificação	dim_Veiculo						
Descrição	Caracterização dos veiculos da empresa "Belos Veiculos"						
Tipo	Com variação						
Dimensão	20R						
Crescimento	0,5%/ano						
Atributos							
Nr	**Identificação**	**Descrição**	**Chave (tipo)**	**Dominio (tamanho)**	**V / H / P**	**Variação**	**Exemplos**
1	idVeiculo	Código de identificação do veiculo.	S	String (45)	---	---	48-28-OV
2	marca	Marca do veiculo.	A	String (45)	---	---	OPEL
3	combustivel	Tipo de combustivel do veiculo.	A	String (45)	---	---	GASOLINA
4	preco	Preço do aluguer diário do veiculo.	A	Float	S / S / ?	Tipo 4	149,99
5	numero_lugares	Nº de lugares do veiculo.	A	Inteiro	---	---	5
Indices							
Nr	**Identificação**	**Indice**	**Tipo**				
1	idVeiculo	Primário	Único, ordenado fisicamente de forma crescente.				
2	marca	Secundário	Ordenado de forma crescente.				
3	combustivel	Secundário	Ordenado de forma crescente.				
4	preco	Secundário	Ordenado de forma crescente.				
5	numero_lugares	Secundário	Ordenado de forma crescente.				
Hierarquia							
Nr	**Identificação**	**Esquema**					
1	H1	idVeiculo -> marca -> ALL					
2	H2	idVeiculo -> combustivel -> ALL					
3	H3	idVeiculo -> preco -> ALL					
4	H4	idVeiculo -> numero_lugares -> ALL					
Perfis de Utilização							
Gestor da empresa "Belos Veiculos" e gestores comerciais.							
Observações							
Nada a assinalar.							

Tabela 6 - Caracterização da dimensão Veículo

4.5 Esquematização do esquema dimensional

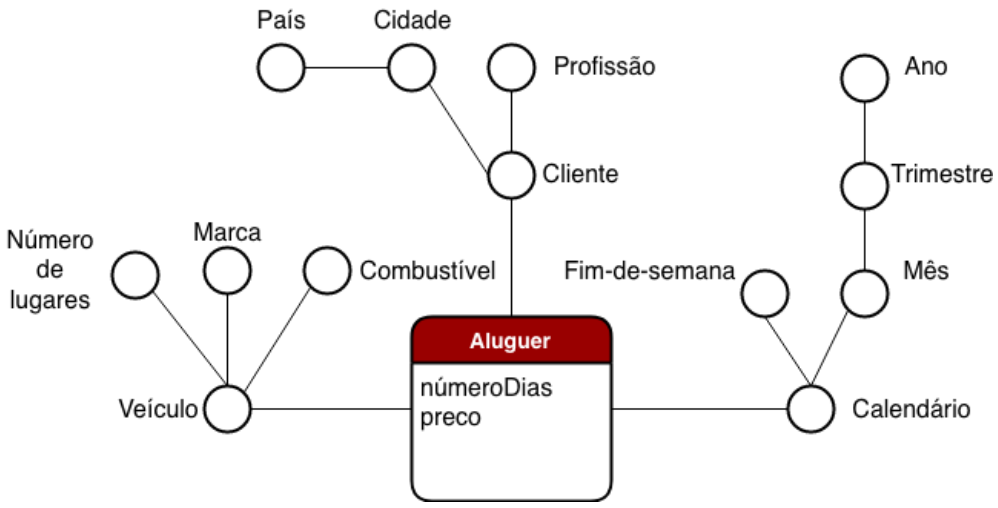

Figura 3 - Esquema dimensional conceptual

4.6 Revisão do esquema dimensional desenvolvido

Após o desenvolvimento do modelo dimensional do sistema foi marcada uma reunião com o cliente para verificar se o mesmo concordava com o modelo proposto segundo os requisitos impostos pela empresa Belos Veículos.

Na conclusão da reunião foi aprovado o modelo, podendo a equipa de desenvolvimento prosseguir para a modelação do processo de ETL.

5. Caracterização das Fontes de Informação

Neste capítulo será feita uma caracterização das fontes, onde serão explorados os aspetos mais relevantes das mesmas. Esta caracterização irá auxiliar a modelação e implementação do sistema ETL, cuja função será extrair os dados da fonte, tratar os dados tendo em conta as regras do negócio estudado e no final carregar esses mesmos dados para o *data warehouse*.

5.1 Identificação e descrição das fontes de informação do sistema

A empresa Belos Automóveis apresenta três fontes de informação distintas: uma que guarda a informação relativa aos alugueres efetuados no *stand* físico (relacional), uma que guarda a informação relativa aos alugueres efetuados pela plataforma da empresa (não-relacional), e a terceira, e última, guarda a informação relativa ao aluguer de carros de luxo (ficheiro *csv*) efetuados pelo gerente da empresa Belos Veículos.

5.1.1 Primeira fonte de dados

A primeira fonte de dados que a empresa *Belos Automóveis* nos disponibilizou corresponde aos registos dos alugueres, efetuados presencialmente no *stand*. Esta base de dados contém a informação relativa aos clientes e veículos disponíveis para aluguer, guardando também a informação correspondente a cada aluguer efetuado. Relativamente ao aluguer, é registado o veículo envolvido, o número de dias pretendidos para o aluguer e qual o valor pago pelo cliente pelo serviço contratado.

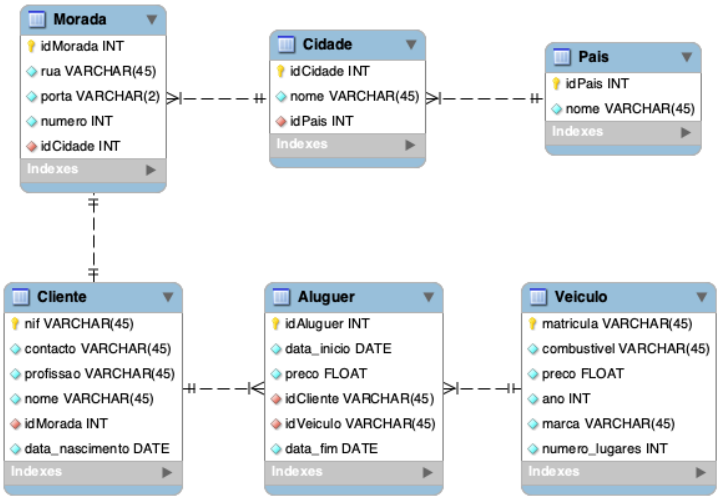

Figura 4 - Fonte de dados relacional

5.1.2 Segunda fonte de dados

A segunda fonte de dados é relativa aos alugueres realizados através da plataforma *online* da empresa. À semelhança da primeira fonte de dados, nesta podemos encontrar informação relativa aos utilizadores que utilizam a plataforma, bem como acerca dos alugueres por estes efetuados. Esta é a fonte de dados não-relacional, definida em *Neo4j*.

Na figura que se segue podemos ter uma melhor perceção de como é a organização da fonte de dados relativa à plataforma *online*, estando identificados os tipos de nodos presentes e os relacionamentos entre estes.

Figura 5 - Fonte de dados não-relacional

5.1.3 Terceira fonte de dados

A terceira fonte de dados disponibilizada corresponde a um ficheiro CSV que, retrata a informação relativa a alugueres efetuados a carros de luxo. Este tipo de aluguer apenas é efetuado pelo gerente da loja, que interage diretamente com o ficheiro de registo. Este ficheiro apresenta a seguinte informação:

- Data do aluguer
- Nome do cliente
- NIF do cliente
- Matrícula
- Marca
- Dias de aluguer
- Valor do aluguer

5.2 Análise dos dados das fontes - relatório de qualidade e de disponibilidade

É necessário fazer uma verificação nos dados existentes a fim de perceber se estes podem ser utilizados para análise, e caso não estejam num bom estado perceber se a falha pode ser corrigida.

5.2.1 Fonte de dados relacional

Ao analisarmos a primeira fonte de dados encontramos vários aspetos que requerem alguma atenção. Começando pelo cliente, é necessário verificar se os valores atribuídos às profissões dos mesmos não são redundantes, isto é, se a mesma profissão não é representada de maneira diferente em dois clientes diferentes. O mesmo acontece com o veículo, que será necessário analisar as marcas e verificar que a mesma marca não é representada de maneira diferente.

Em relação aos alugueres, o número de dias de um aluguer terá sempre que ser maior que zero. Assim, será necessário averiguar se não ocorreu nenhum erro na introdução destes valores, e caso se encontre erros contactar a empresa e averiguar possível correção.

5.2.2 Fonte de dados não-relacional

Como a base de dados da segunda loja está alocada numa base de dados *NoSQL* e como não existem uma estrutura prévia das tabelas presentes é necessário analisar cada linha da tabela para averiguar a estrutura presente em casa uma delas.

Existem também os casos enumerados a cima que é preciso também tomar em conta quando se for a extrair os dados.

5.2.3 Fonte de dados csv

Esta fonte de dados é a fonte de dados cujo gerente interage diretamente. Esta apenas guarda os registos de alugueres de carros de luxo. Como tal, em qualquer um dos campos que compõem esta terceira fonte, poderão haver erros na escrita. Assim, é necessário fazer uma revisão de todo o ficheiro e verificar que, por exemplo, a marca de um carro não é escrita de maneira diferente em dois casos distintos.

5.3 Desenvolvimento do esquema de mapeamento de dados - source-to-target data map

Target				Source				Transformação
BD	Tabela	Tipo	Coluna	BD	Tabela	Coluna	Tipo	
dw	dim_Cliente	dimensão	profissao	dw_font	Cliente	profissao	VARCHAR(45)	Direto
dw	dim_Cliente	dimensão	cidade	dw_font	Cidade	nome	VARCHAR(45)	É necessário juntar a tabela do Cliente com a Morada e posteriormente com a Cidade de forma a conseguirmos retirar o nome da mesma
dw	dim_Cliente	dimensão	pais	dw_font	Pais	nome	VARCHAR(45)	Igual ao atributo cidade, mas ainda requer uma junção com a tabela Pais
dw	dim_Cliente	dimensão	idade	dw_font	Cliente	data_nascimento	DATE	Calcular a idade através da data de nascimento do cliente

Tabela 7 - Tabela de mapeamento da dimensão dim_Cliente para a primeira fonte

Target				Source			Transformação
BD	Tabela	Tipo Tabela	Coluna	BD	Nodo	Propriedade	
dw	dim_Cliente	dimensão	profissao	dw_neo	Cliente	profissao	Direto
dw	dim_Cliente	dimensão	cidade	dw_neo	Cliente	cidade	Direto
dw	dim_Cliente	dimensão	pais	dw_neo	Cliente	pais	Direto
dw	dim_Cliente	dimensão	idade	dw_neo	Cliente	datanascimento	Fazer o cálculo da idade através da data de nascimento

Tabela 8 - Tabela de mapeamento da dimensão dim_Cliente para a segunda fonte

Target				Source				Transformação
BD	Tabela	Tipo	Coluna	BD	Tabela	Coluna	Tipo	
dw	dim_Veiculo	dimensão	marca	dw_font	Veiculo	marca	VARCHAR(45)	Direto
dw	dim_Veiculo	dimensão	combustivel	dw_font	Veiculo	combustivel	VARCHAR(45)	Direto
dw	dim_Veiculo	dimensão	preco	dw_font	Veiculo	preco	FLOAT	Direto
dw	dim_Veiculo	dimensão	numero_lugares	dw_font	Veiculo	numero_lugares	INT	Direto

Tabela 9 - Tabela de mapeamento da dimensão dim_Veiculo para a primeira fonte

Target				Source				Transformação
BD	Tabela	Tipo	Coluna	BD	Nodo	Propriedade	Tipo	
dw	dim_Veiculo	dimensão	marca	dw_neo	Veiculo	marca	String	Direto
dw	dim_Veiculo	dimensão	combustivel	dw_neo	Veiculo	combustivel	String	Direto
dw	dim_Veiculo	dimensão	preco	dw_neo	Veiculo	preco	String	Direto
dw	dim_Veiculo	dimensão	numero_lugares	dw_neo	Veiculo	numero_lugares	String	Direto

Tabela 10 - Tabela de mapeamento da dimensão dim_Veiculo para a segunda fonte

Target				Source				Transformação
BD	Tabela	Tipo	Coluna	BD	Tabela	Coluna	Tipo	
dw	fact_Rent	facto	numeroDias	dw_font	Rent	noDays	INT	Calculado a partir da data fim e de início
dw	fact_Rent	facto	numeroDias	dw_neo	Rent	days	int	Calculado a partir da data fim e de início
dw	fact_Rent	facto	numeroDias	dw_csv	-	dias_aluguer	Number	Calculado a partir da data fim e de início
dw	fact_Rent	facto	preco	dw_font	Rent	value	FLOAT	Direto
dw	fact_Rent	facto	preco	dw_neo	Rent	price	float	Direto
dw	fact_Rent	facto	preco	dw_csv	-	valor_aluguer	Décimal	Direto

Tabela 11 - Tabela de mapeamento relativa à tabela de factos

6. Modelação do Sistema de Povoamento

Após o desenvolvimento da modelação dimensional do sistema é necessário descrever o processo de povoamento do mesmo. Este povoamento está decomposto em duas fases, o **carregamento inicial** no qual os dados existentes nas diversas fontes são carregados para o data warehouse, e o **carregamento regular**, efetuado diariamente como descrito na modelação dimensional do sistema onde são inseridos os novos dados no sistema de data warehouse.

6.1 Esquematização conceptual e caracterização do sistema de povoamento

Nesta fase é descrita a modelação conceptual do sistema de povoamento em BPMN assim como uma descrição de cada um dos passos do processo ETL realizado.

6.1.1 Extração

Na fase de extração do processo ETL os dados necessários ao povoamento das dimensões e tabelas de factos são carregados de cada uma das fontes e colocados em tabelas correspondentes na área de retenção. Isto permite a extração de dados das três fontes em paralelo como descrito no diagrama seguinte.

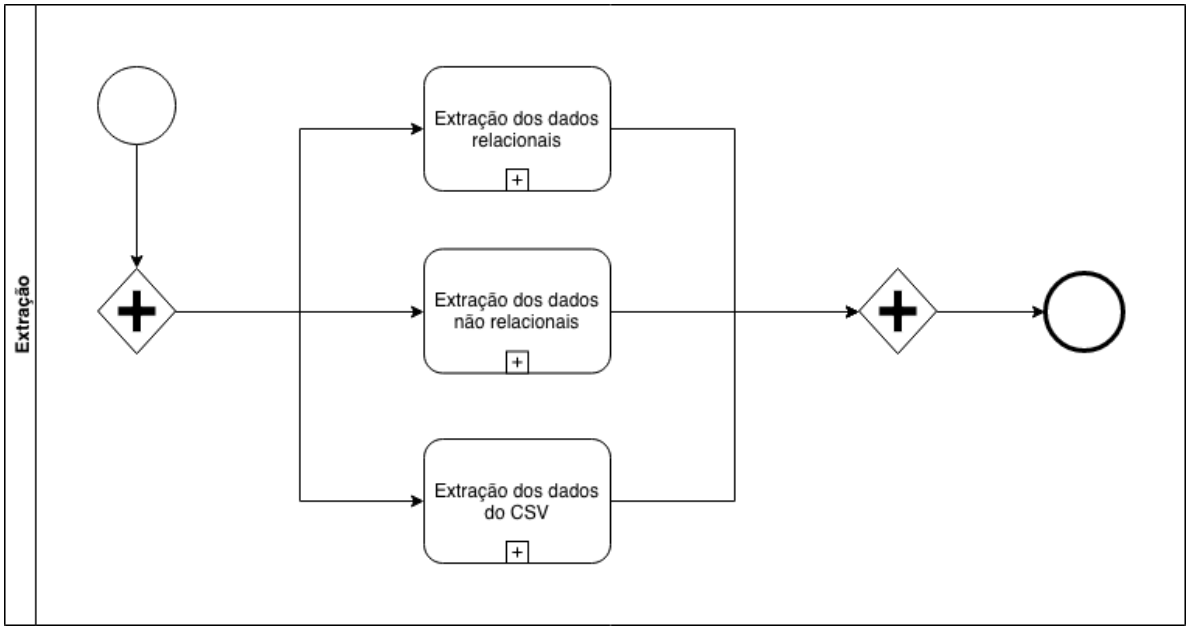

Figura 6 - Processo de extração das diversas fontes

O processo de extração da fonte relacional descrito de seguida é igual quer no carregamento inicial quer no carregamento regular onde são lidas as várias tabelas e substituído o código identificador de pais e cidade pelos respetivos nomes. A diferença entre os dois carregamentos está nas tabelas fonte que fornecem os dados. No carregamento inicial os dados são lidos das tabelas operacionais pois não se encontram no data warehouse. Após este carregamento inicial um *trigger* é responsável por adicionar em tabelas de auditoria os novos dados com a indicação da operação que os origina ('I' para inserção ou 'U' para atualização), tabelas essas que serão a fonte de dados no carregamento regular.

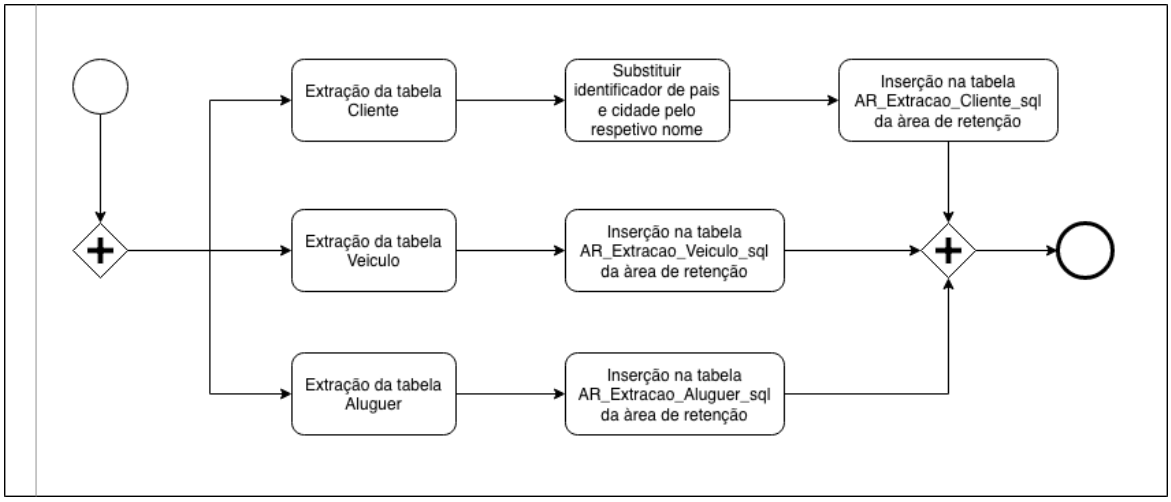

Figura 7- Processo de extração para a fonte de dados MySQL

O carregamento a partir das fontes Neo4J e csv não requerem "join" de valores pois os campos necessários vêm diretamente da fonte no formato desejado sendo preenchidos através de um formulário que garante a correção dos dados como descrito pelo cliente.

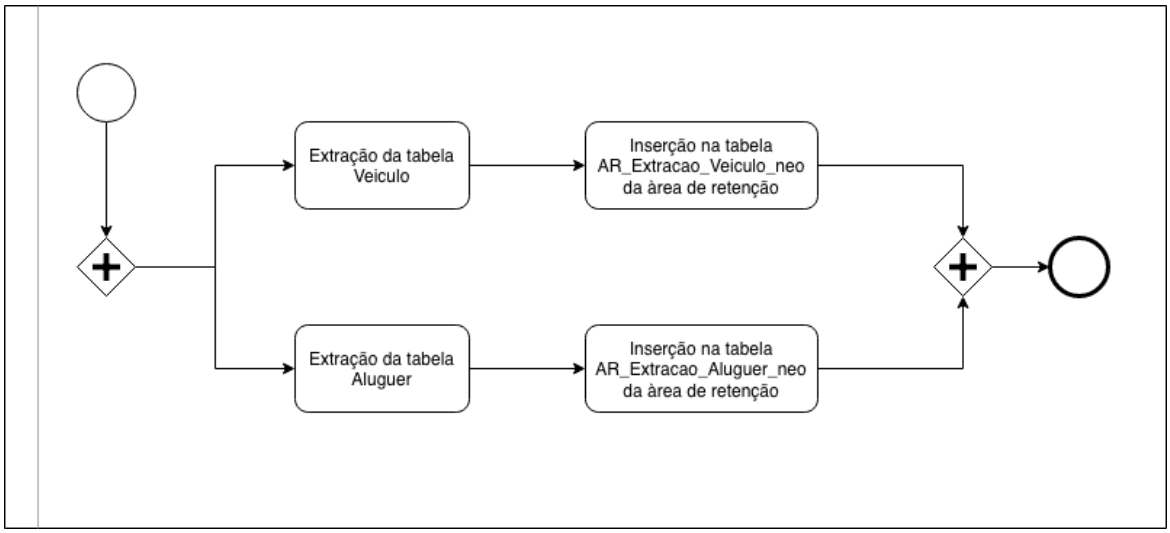

Figura 8 - Processo de extração para a fonte de dados Neo4J

O processo de adicionar valores à dimensão Calendário é um caso particular efetuado diretamente de um ficheiro csv. Isto acontece visto se tratar de uma Dimensão sem variação e que como referido será pré-povoada no carregamento inicial para um período de 10 anos.

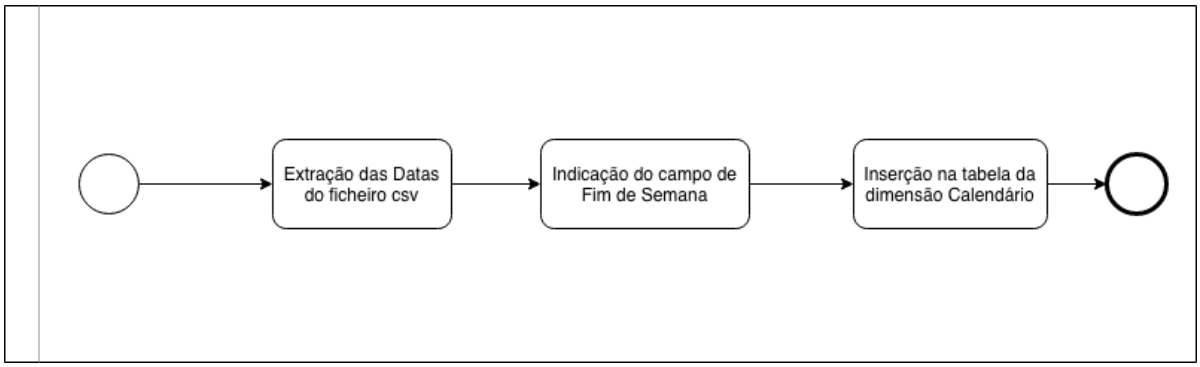

Figura 9 - Processo de extração e carregamento da dimensão Calendário

6.1.2 Limpeza

A fase de limpeza num processo ETL tem como objetivo a homogeneização dos dados das diversas fontes. Como identificado anteriormente existe a necessidade de tratar dados relativos ao Cliente (profissão) e aos Veículos (marca e combustível) sendo efetuada a capitalização dos valores dos mesmos provenientes das várias fontes.

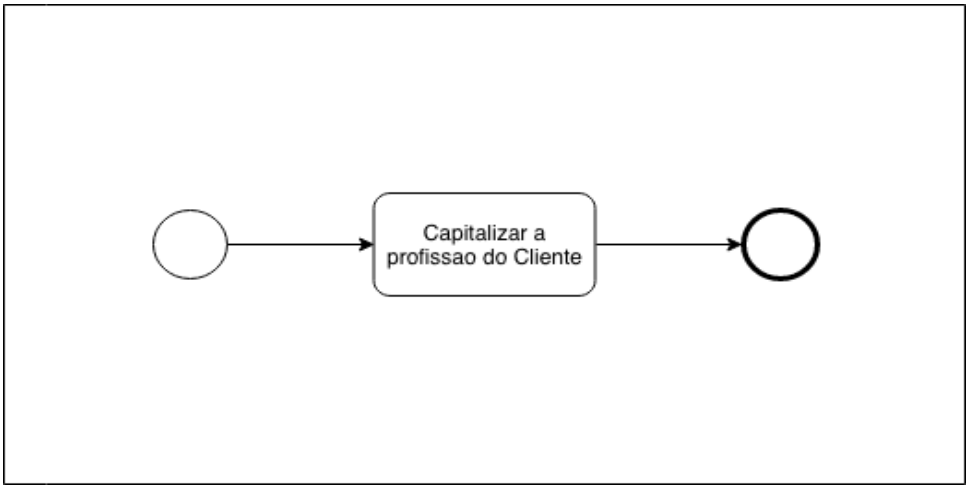

Figura 10 - Processo de limpeza dos Clientes da fonte MySQL

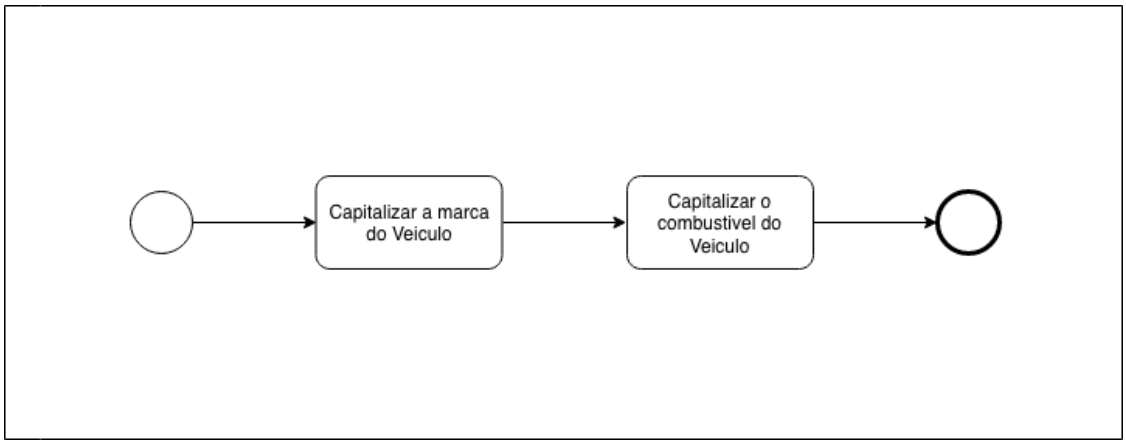

Figura 11 - Processo de limpeza do Veículo na fonte de dados MySQL e Neo4J

6.1.3 Conciliação

Na fase de conciliação como o próprio nome indica, os dados das diversas fontes são integrados entre si em pré tabelas de dimensão e factos ficando assim preparados para a inserção final no data warehouse na fase de carregamento.

Como as fontes de dados são homogéneas na implementação, isto é, não existe a possibilidade de duplicação de chaves visto o registo dos utilizadores se encontrar na fonte relacional e os veículos das fontes onde são registados serem diferentes este processo apenas efetua o calculo das medidas e o povoamento das tabelas de pré dimensões e pré tabela de factos.

6.1.4 Carregamento

Na fase final do processo ETL é efetuado o carregamento dos dados de forma direta a partir das pré tabelas de dimensões e factos, pois os dados já se encontram com o mesmo tipo e formato do suposto no sistema de data warehouse. O único caso especial é a variação na inserção entre valores de *update* e valores de inserção conforme descrito para os atributos com variação.

6.2 Descrição e caracterização dos elementos de dados utilizados para suporte ao povoamento

Para auxiliar o povoamento como referido foi definida uma área de retenção com tabelas que armazenaram os dados durante as diferentes fases do processo ETL.

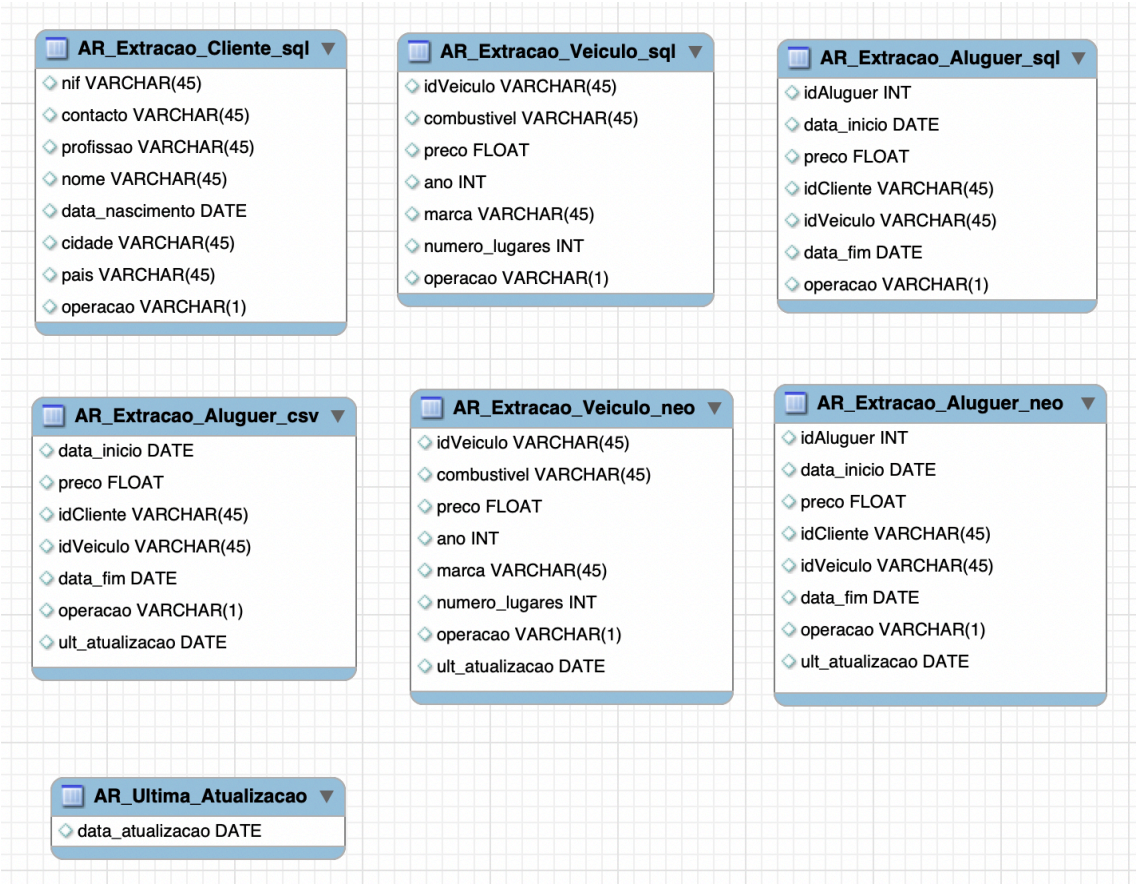

Figura 12 - Tabelas de extração da área de retenção

Inicialmente os valores extraídos são colocados em tabelas de extração representadas acima. Estas tabelas são iguais nos dois tipos de carregamento, porém a tabela de controlo AR_Ultima_Atualizacao apenas tem utilidade no carregamento regular fornecendo informação acerca da marca temporal da ultima extração, definindo o conjunto de valores que serão carregados no próximo processo de extração.

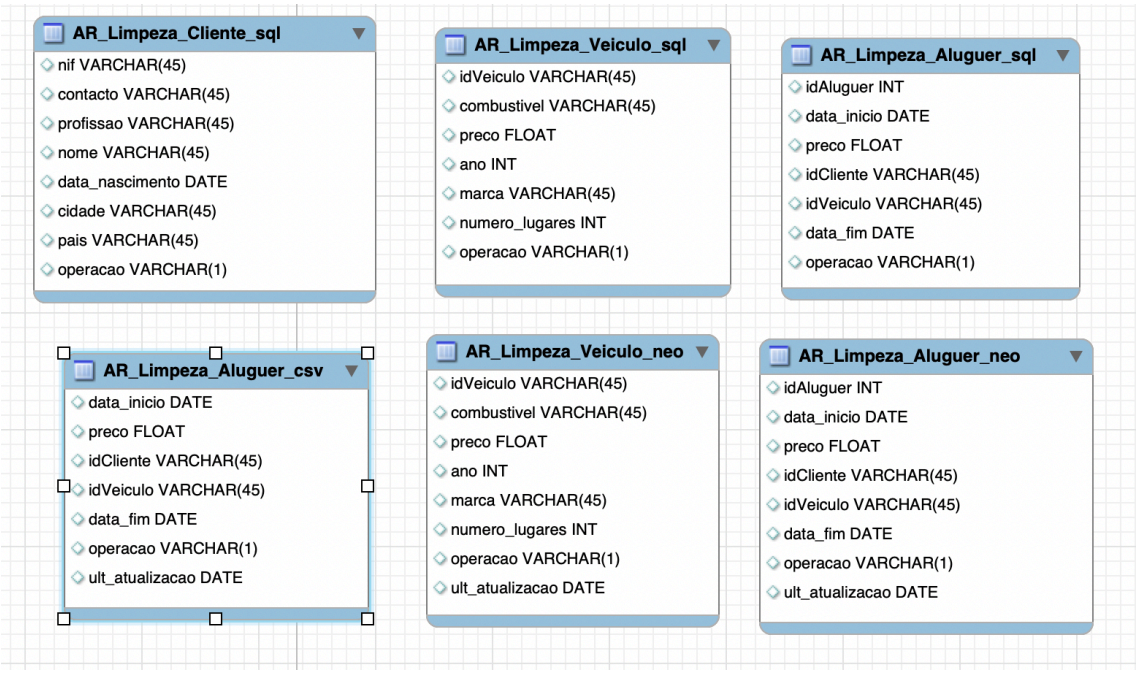

Figura 13 - Tabelas de limpeza da área de retenção

Para a fase de limpeza foram criadas tabelas de limpeza na área de retenção que irão transferir os registos da tabela de extração e efetuar as transformações necessárias para obter os respetivos valores tratados.

Para a fase de conciliação foram criadas pré tabelas de dimensões, histórico e facto.

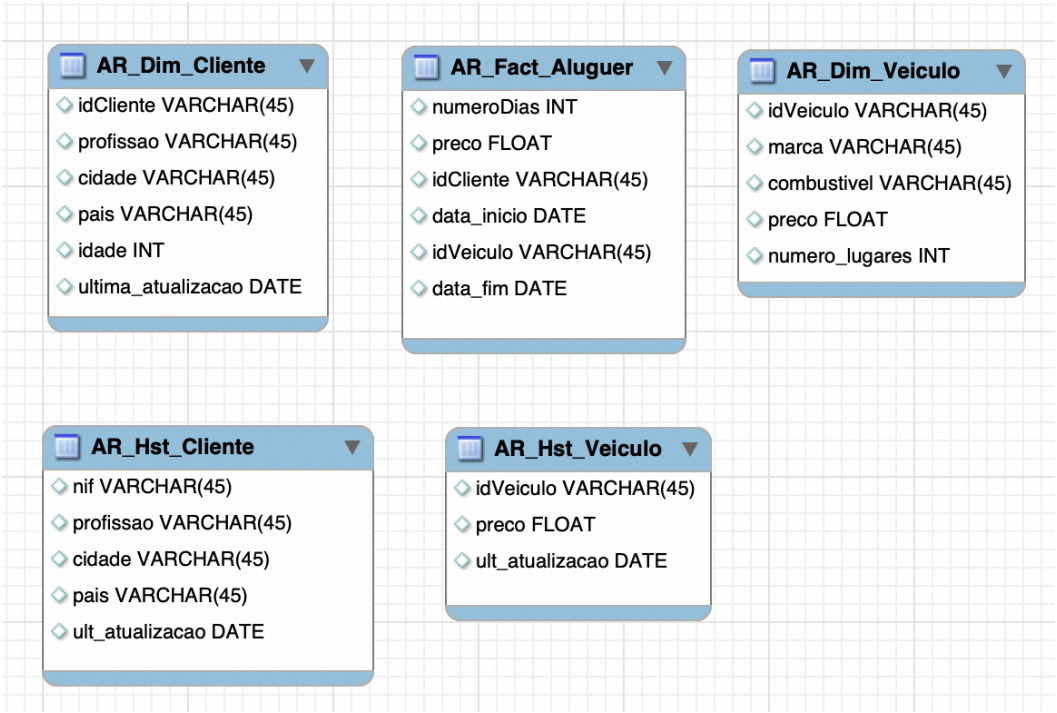

Figura 14 - Tabelas de conciliação da área de retenção

7. Implementação do Sistema de Data Warehousing

Concluída a modelação do sistema de data warehousing e confirmada a solução desenvolvida ao cliente é necessária a implementação do sistema. Para tal foi utilizada a ferramenta *Pentaho Kettle* que permite desenvolver todo o processo de ETL de forma simples, tendo que os sistemas de armazenamento de dados sido desenvolvidos no *MySQL.*

7.1 Implementação do Sistema de Povoamento

A implementação do sistema de povoamento é repartida em dois Jobs correspondentes ao carregamento inicial e final, representados respetivamente nas seguintes figuras.

Figura 16 - Processo geral do carregamento inicial

Figura 15 - Processo geral do carregamento regular

Como referido na modelação do sistema o job do carregamento inicial efetua o carregamento da dimensão Calendário, efetuando posteriormente o processo ETL descrito tendo como fonte de dados os sistemas operacionais da empresa Belos Veículos. O carregamento regular difere neste processo de extração da fonte relacional efetuando a leitura de registos da tabela de auditoria que salvaguarda os registos em conjunto com um atributo **operação** que na fase de conciliação determina a inserção na dimensão ou na tabela histórico da mesma.

Para a fonte de dados NoSQL e csv devido à não existência de *triggers*, foi adicionado um *timestamp* aos registos permitindo que através das tabelas de controlo na área de retenção seja possível extrair apenas as operações efetuadas desde a última extração, garantindo a consistência dos dados presentes no data warehouse.

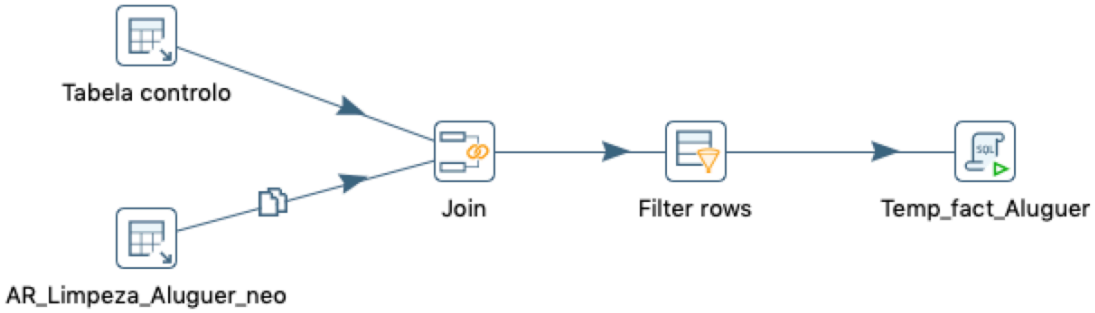

Figura 17 - Processo de extração dos alugueres da fonte Neo4j no carregamento regular

A criação de tabelas intermédias entre cada fase do processo ETL permite garantir a continuação do processo mesmo na ocorrência de falhas, fator importante para garantir que o processo ETL finaliza no tempo que assim tem para o fazer, mesmo na ocorrência de uma falha. Tal seria difícil de alcançar num cenário com um grande volume de dados em que ocorresse uma falha a meio do processo sem este mecanismo de recuperação.

7.2 Análise da execução do sistema de povoamento

Concluída a implementação do data warehouse, a equipa de desenvolvimento efetuou testes ao sistema desenvolvido para garantir o correto funcionamento do mesmo. Inicialmente foi efetuado o carregamento inicial para garantir que todo o processo decorria como esperado. Tal é confirmado pela contagem do número de registos e verificação da correta transformação dos valores que tal requeriam, através do uso de registos inseridos num cenário de povoamento teste efetuado pela equipa.

Após a confirmação de que o carregamento inicial contemplava os requisitos para a consistência do sistema, foi efetuado um cenário de teste ao processo de um carregamento regular, com a inserção de vários registos nas dimensões e facto, assim como *update* a valores definidos como atributos com histórico. Na sequência da execução destes testes verificamos que apenas os novos registos estavam a ser inseridos garantindo o correto funcionamento do mecanismo da tabela de controlo e auditoria.

Apesar dos novos registos serem inseridos sem problema, deparamo-nos com um mau funcionamento na atualização de atributos com histórico devido a um erro no *procedure* desenvolvido.

Assim, apesar deste percalço na atualização de valores com histórico, todo o restante sistema de *data warehouse* se mostra consistente nas duas fases que tornam a implementação viável para entrega ao cliente.

8. Conclusões e Trabalho Futuro

A implementação de um sistema de data warehouse provou ser um desafio com um algum nível de dificuldade, de forma a poder satisfazer todos os requisitos do cliente. O desenvolvimento de um *data warehouse* é um processo trabalhoso que requer um planeamento cuidadoso e atento em que uma pequena falha na modelação poderá causar grande impacto no sistema desenvolvido.

O sistema final permite que o agente de decisão da *Belos Automóveis* retire conclusões sobre o número de alugueres, quais as zonas mais comuns de onde provêm os seus clientes e gerir o stock de veículos disponíveis para aluguer consoante a época do ano, obtendo a visão geral do negócio que o cliente tão pretendia alcançar.

Apesar disso, no decorrer do desenvolvimento do projeto, mais especificamente na implementação do sistema ETL, ficou clara a importância do planeamento deste tipo de sistemas. A complexidade que apresentam requerem muita atenção na hora da implementação, que com o suporte do planeamento realizado torna mais fácil reconhecer com que dados precisamos de trabalhar tendo em conta a funcionalidade que queremos ver implementada. Ao longo do desenvolvimento do ETL deparamo-nos com alguns problemas, e consequentemente dificuldades em perceber como os corrigir, o que provocou um certo atraso no desenvolvimento do projeto.

Na eventualidade de uma continuação deste projeto, a equipa detetou que existem certos aspetos que necessitam de ser melhorados, como por exemplo (1) a validação dos valores provenientes das fontes, (2) a recuperação do processo de ETL em caso de falha e (3) o tratamento de valores nulos no csv. Existem também pequenos erros na definição dos sistemas de dados utilizados neste projeto que necessitam de ser corrigidos para um correto funcionamento do sistema implementado.

Referências

Golfarelli, M. and Rizzi, S. (2009). *Data warehouse design: modern principles and methodologies*. 1st ed. New York: McGraw-Hill.

Kimball, R., Margy, M.*, The Data Warehouse Toolkit*, *3rd Edition (Wiley, 2013)*

Lista de Siglas e Acrónimos

De seguida apresentamos as siglas que são referidas durante este relatório

SDW - Sistema de Data Warehousing

DW - Data Warehouse

Csv - Comma-Separated Values

ETL – Extract Transform Load