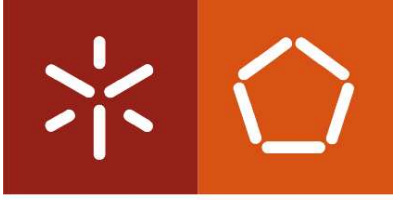


Universidade do Minho

Escola de Engenharia

André Correia Morgado

Data Warehouses Espaciais – Projeto e implementação

Outubro de 2013

Universidade do Minho
Escola de Engenharia
Departamento de Informática

André Correia Morgado

Data Warehouses Espaciais – Projeto e implementação

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de
Professor Orlando Belo

Outubro de 2013

Agradecimentos

Quero agradecer a todas as pessoas que direta ou indiretamente contribuíram para a realização desta dissertação. Em especial:

Ao professor e orientador Orlando Belo pela disponibilidade demonstrada e pelos concelhos e sugestões dadas ao longo deste processo.

A todos os meus colegas que frequentaram comigo, tanto o mestrado como a licenciatura.

A todos os amigos que me incentivaram e motivaram neste processo.

À minha namorada, Débora Filipe, por toda a ajuda e força que deu ao longo deste tempo, e por nunca me ter deixado desistir.

Por fim, aos meus pais, Paula Correia e João Morgado, por me terem proporcionado esta oportunidade, pela dedicação e pelo incentivo que sempre me deram.

Resumo

Data Warehouses Espaciais

Segundo um estudo realizado pela empresa *International Data Corporation (IDC)* (Adrian Bridgwater, 2009), o mercado dos *data warehouses* tem tido um grande crescimento. Cada vez mais as empresas procuram guardar todos os dados relacionados com o seu negócio, de forma a obter o máximo de conhecimento possível, podendo, assim, tomar melhor decisões relacionadas com o seu negócio. Os *data warehouses* aparecem como uma ferramenta útil para suporte a processos de tomada de decisão. A capacidade dos *data warehouses* guardarem grandes quantidades de dados relativos ao negócio da empresa e permitirem aos agentes de decisão acederem de forma simples e fácil a esses dados, fazem deles uma ferramenta de eleição para o processo de tomada de decisão. A partir dos dados presentes num data warehouse pode-se efetuar relatórios e análises detalhadas. Apesar de serem ferramentas muito poderosas, os *data warehouses* ditos convencionais ainda contêm limitações relativamente à capacidade de guardar e analisar dados com características geográficas. Estas ferramentas capazes de lidar com este tipo de características são largamente utilizadas pelas empresas em muitos domínios de aplicação, como as telecomunicações ou a segurança, que com auxilio desta ferramenta conseguem descobrir qual o melhor local onde instalar antenas ou, então, por entidades Governamentais, de forma a descobrir as zonas do seu país, com a maior criminalidade. Ao longo desta dissertação, pretende-se entender o processo de construção de um *data warehouse* espacial desde a sua fase de levantamentos e análises de requisitos até à sua fase de implementação, sendo, por fim, transformado um *data warehouse* convencional num *data warehouse* espacial recorrendo a toda a informação obtida ao longo do processo.

Abstract

Spatial Data Warehouses

According to a study by International Data Corporation (IDC) (Adrian Bridgwater, 2009), the market for data warehouses has been a great growth. Increasingly companies are looking to save all data related to your business in order to get the maximum possible knowledge, and thus can take better decisions related to your business. Data warehouses appear as a useful tool to support decision-making processes. The ability of the data warehouses keep large amounts of data relating to the business and enable decision-makers to access simple and easy way to such data, makes them a tool of choice for the process of decision making. From the data presented in a data warehouse can make reports and detailed analyzes. Although they are very powerful tools, said conventional data warehouses have limitations regarding the ability to store and analyze data with geographical features. These tools able to cope with this kind of features are widely used by companies in many application domains, such as telecommunications and security, with the help of this tool you can find out what the best place where to install antennas or else by governmental entities, in order to discover the areas of their country, with the highest crime. Throughout this dissertation, we intend to understand the process of building a spatial data warehouse from its phase of surveys and analysis of requirements until its implementation phase, and finally turned a conventional data warehouse in a data warehouse using spatial all the information obtained during the process.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 *Data Warehouses* nas Organizações

Tendo em conta o estado atual da economia, o processo de tomada de decisão por parte dos agentes responsáveis pelas empresas assume um papel fulcral no crescimento e subsistência das suas organizações. Algumas empresas possuem, por exemplo, vários departamentos em que cada um deles detém uma base de dados independente, que quando consideradas como um todo possuem informação repetida. Essas bases de dados são bastante uteis para as operações do dia-a-dia. Contudo, quando é necessário efetuar um relatório contendo informação dessas bases de dados, essa tarefa apresenta alguns constrangimentos que atrasam o processo de tomada de decisão (Bédard et al. 2001). Entre outras, tais circunstâncias levaram muitas empresas a apostarem na implementação de um *data warehouse* (Inmon,2002), como ficou provado pelo estudo realizado por Paul Gray, em que 90% das empresas que faziam parte da *Fortune 500*, tinham ou pretendiam ter um *data warehouse* (Gray e Israel, 1999).

A eficiência das operações das organizações é bastante importante para o sucesso da empresa, contudo, para que a organização assuma um papel de destaque no mercado é necessário que a tomada de decisões seja baseada em conhecimento sobre o mercado e sobre

as suas próprias tendências. Os sistemas de suporte à decisão, como os *data warehouses*, assumem um papel de destaque nesta área, visto fornecerem ferramentas rápidas e eficientes que ajudam os agentes de decisão a compreender padrões e comportamentos das suas atividades e elementos do negócio.

As organizações, cada vez mais, guardam mais dados. Estes, depois de armazenados nos seus *data warehouses*, constituem a base e a principal fonte de informação para os seus agentes de decisão (Rivest et al., 2001). Contudo, apesar dos *data warehouses* convencionais serem capazes de lidar com vários tipos e formatos de dados, não são capazes de guardar e processar dados espaciais, não sendo, assim, possível, em alguns cenários aplicacionais, retirar uma parte importante do conhecimento contido na informação recolhida pelas organizações. Um estudo estimou que 80% dos dados, guardados nas bases de dados das organizações, contêm uma componente espacial (Franklin, 1992), que pode ser caracterizada por propriedades como posição, forma, orientação ou tamanho. Por exemplo, a informação sobre ruas, cidades, países de um dado cliente, é tratada pelos *data warehouses* convencionais, como dados alfanuméricos, não permitindo assim a georreferenciação dos dados com conteúdo espacial. Isto faz com que tais dados não sejam representados sobre um sistema de referência espacial como, por exemplo, o sistema de coordenadas, o que traduz numa possível perda de conhecimento, tornando-se bastante difícil dessa forma descobrir padrões e relacionamentos entre os dados espaciais.

1.2 Motivação

No seguimento da necessidade de criação de uma ferramenta orientada para os dados espaciais, surgiram as bases de dados espaciais com capacidade de guardar e manipular este tipo de dados, respeitando a sua própria natureza, fornecendo, para isso, um conjunto de funções e operadores específicos que permitem verificar as relações topológicas entre os dados espaciais (Shekar e Chawla, 2003). Tendo como principal função o suporte às operações do dia-a-dia, esta ferramenta não possui as capacidades analíticas dos *data warehouses***.** Para que fosse possível extrair todo o conhecimento dos dados espaciais, seria necessário dispor de uma

ferramenta que tivesse a capacidade de guardar, manipular e mostrar todos os tipos de dados geográficos. Esta capacidade é fornecida pelos sistemas de informação geográfica que, apesar de serem bastantes uteis para visualizarem dados espaciais, são normalmente lentos a processar queries complexas (Bédard, 2003). Seria necessário, assim, uma ferramenta que combinasse a capacidade dos *data warehouses* com a capacidade de representar dados geográficos.

Desta necessidade, surgiram os *data warehouse* espaciais, como ferramenta capaz de fornecer suporte para dados espaciais como a consultas espaciais. Contudo, ainda é uma área pouco explorada e com bastante potencial para as organizações. Devido à pouca utilização deste tipo de *data warehouses* por parte das organizações, esta é uma área pouco explorada pelos investigadores e que ainda não reúne um consenso. É esse fator a principal motivação desta dissertação, perceber a complexidade dos dados espaciais e de que forma afetam a construção de um *data warehouse*.

1.3 Âmbito da Dissertação

Segundo *Malinowski* (Malinowski, 2006), a ausência de uma metodologia para a modelação de bases de dados espaciais, combinado, com a inexistência de um consenso relacionado com a modelação de *data warehouses* espaciais, leva a que o termo *data warehouse* espacial seja por vezes mal aplicado. O termo é muitas vezes aplicado, quando existe uma grande quantidade de dados espaciais presentes nos sistemas operacionais ou no *data warehouse*. Contudo, um *data warehouse* espacial consiste numa ferramenta de suporte à decisão com capacidade de analisar tanto dados tabulares, como dados espaciais, sendo que estes últimos são representados num mapa por forma a fornecer aos agentes de decisão maior facilidade de perceção de padrões e tendências.

Este trabalho propõe analisar o processo de desenho e construção de um *data warehouse* espacial, dando particular atenção aos aspetos relacionados com a sua modelação dimensional. O foco na modelação dimensional prendesse com o facto de ser, o principal ponto diferencial na

construção de um *data warehouse* convencional quando comparada com a construção de um *data warehouse* espacial.

1.4 Objetivos

Com este trabalho pretende-se construir um *data warehouse* espacial capaz de fornecer suporte a processos de análise espacial num determinado contexto aplicacional. O processo de construção do *data warehouse* espacial deve abranger todos os passos da construção de um *data warehouse* convencional, desde a sua fase de levantamento de requisitos, até à criação do processo povoamento do *data warehouse*, passando obviamente por todas as fases intermédias que um projeto deste género usualmente considera. O objetivo desta dissertação pode-se dividir nos seguintes sub-objetivos:

- Análise do estado da arte dos modelos propostos para modelação de um data warehouse espacial.
- Caracterização das diferenças processuais que podem ser encontradas entre a construção de um *data warehouse* convencional e a de um *data warehouse* espacial.
- Transformação de um *data warehouse* convencional num *data warehouse* espacial para um dado domínio de aplicação.
- Construção de todos os processos necessários para o povoamento do data warehouse espacial, dando principal destaque ao conceitos relacionados com os dados espaciais.

1.5 Metodologia de investigação

Por forma a cumprir todos os objetivos propostos anteriormente, o desenvolvimento dos trabalhos desta dissertação tiveram como base a metodologia de investigação apresentada em *Carr*, (2006). O primeiro passo desta metodologia consiste na identificação de um problema no qual, posteriormente, possa ser formulada uma hipótese que será usada durante o desenvolvimento.

Após concluído este primeiro passo, a informação que aí foi recolhida será reavaliada continuamente por forma a desenvolver uma solução para o problema em causa.

Por fim, será possível analisar os resultados obtidos com base na documentação produzida ao longo de todo processo.

Por forma a conseguir alcançar os objetivos desta metodologia, foi necessário seguir alguma diretrizes, em particular:

- Definição e especificação do problema, tendo especial atenção aos seus detalhes;
- Atualização constante do conteúdo referente ao estado da arte;
- Modelação e implementação de todo o sistema;
- Analise dos resultados e concretização das conclusões tendo como base os objetivos propostos;

1.6 Estrutura do documento

Para além do presente capítulo, esta dissertação incorpora um conjunto de mais quatro capítulos, nomeadamente:

- Capítulo 2 - **Modelação de *Data Warehouses* Espaciais**, que apresenta um "estado da arte" sobre a modelação dimensional de *data warehouses* espaciais e discute os aspetos mais pertinentes na realização prática destes sistemas.
- Capítulo 3 - **Implementação de um *Data Warehouse* Espacial**, que descreve o processo de criação de um *data warehouse* espacial, focando-se essencialmente na sua componente espacial.
- Capítulo 4 - **Um Caso de Estudo**, no qual se descreve o processo utilizado para a transformação de um *data warehouse* convencional, num *data warehouse* espacial, fazendo-se referência às várias decisões tomadas ao longo do processo.
- Capítulo 5 - **Conclusões e Trabalho Futuro**, capítulo no qual é realizada uma síntese de todo o trabalho realizado no âmbito desta dissertação, bem como se indica algumas linhas de orientação para a realização de trabalho futuro no âmbito do presente trabalho.

Capítulo 2

Modelação de Data Warehouses Espaciais

2.1 *Data Warehouses* Espaciais

Várias propostas foram feitas ao longo dos anos por vários investigadores tendo em conta a utilização de dados espácias num ambiente típico de *data warehousing*, isto é, a criação de *data warehouses* espaciais.

Um *data warehouse* espacial (*Bédard et al.*,2001) combina todas as capacidades do modelo dimensional com a capacidade de suportar operações sobre dados geográficos (*Malinowski and Zimányi*, 2008). Um *data warehouse* espacial é uma "coleção de dados espaciais e não espaciais, orientados por assunto, não voláteis, integrados e variáveis no tempo" (*Stefanovic et al.*, 1998). Analisando esta definição, facilmente se verifica que esta é bastante semelhante à definição de um *data warehouse* convencional, apresentada por *Bill Inmon* (*Inmon*,2002). Assim, os *data warehouse* espaciais podem ser vistos como uma extensão dos *data warehouses* convencionais.

Apesar de terem surgido várias propostas ao longo dos anos em relação à modelação dimensional dos *data warehouses* espaciais, a investigação sobre este tipo particular de *data warehouses* ainda se encontra um passo atrás em relação aos *data warehouses* convencionais. Os motivos para que tal aconteça são vários. De referir (Maria Luisa Damiani & Stefano Spaccapietra, 2009):

- A falta de um mercado capaz de impulsionar a sua investigação, tal como acontece nas outras áreas, com maior mercado comparativamente aos *data warehouses* convencionais.
- O contexto espacial é complexo e peculiar, necessitando por isso de técnicas especializadas para a representação dos dados e, consequentemente do seu processamento.
- Só a partir de 1999, com a revisão do SQL, as tecnologias que tiveram como base de sua implementação o SQL-3, atingiram a maturidade necessária para lidar com dados espaciais.

Por estes motivos, a definição de um modelo dimensional para data warehouses espaciais continua a ser uma questão em aberto, visto não haver ainda um consenso generalizado quanto ao modelo a aplicar.

2.2 Modelos Dimensionais para Data Warehouses Espaciais

A investigação sobre *data warehouses* espaciais é relativamente recente. Apesar disso, desde a primeira proposta realizada por *Han et al* (1998), para um modelo multidimensional para data warehouses espaciais, vários modelos têm surgido entretanto, cada um deles, com características e conceitos distintos. De seguida, com intuito de entender as várias abordagens possíveis, iremos debruçar-nos com mais detalhe sobre algumas dessas propostas.

2.2.1 O Modelo de Han et al. (1998).

Em 1998, surgiu o primeiro modelo para um *data warehouse* espacial proposto por Han et al. (1998) ainda hoje, talvez, o mais significativo. Este modelo introduziu novos conceitos como a dimensão espacial e a medida espacial, fazendo assim uma clara distinção entre dimensões e medidas com características espaciais das não espaciais.

Basicamente, uma dimensão não espacial consiste numa dimensão que contem apenas dados não espaciais, isto é, dados nominais que não possuem representação cartográfica associada. Dimensões espaciais descrevem propriedades de factos que possuem a capacidade de serem representados cartograficamente, podendo, assim, os membros dessa dimensão, serem consultados e visualizados num mapa.

Além da distinção dos tipos das dimensões, foram ainda propostas dois tipos de medidas: as medidas numéricas e as medidas espaciais. As primeiras são medidas que possuem unicamente dados numéricos. Por sua vez, as medidas espaciais definem medidas " que contêm um ou uma coleção de pontos para objetos espaciais", não tendo, por isso, uma caracterização semântica, sendo unicamente constituída por um conjunto de geometrias. Para ilustrar estes conceitos, Han et al. (1998) consideraram um *data warehouse* espacial sobre dados meteorológicos.

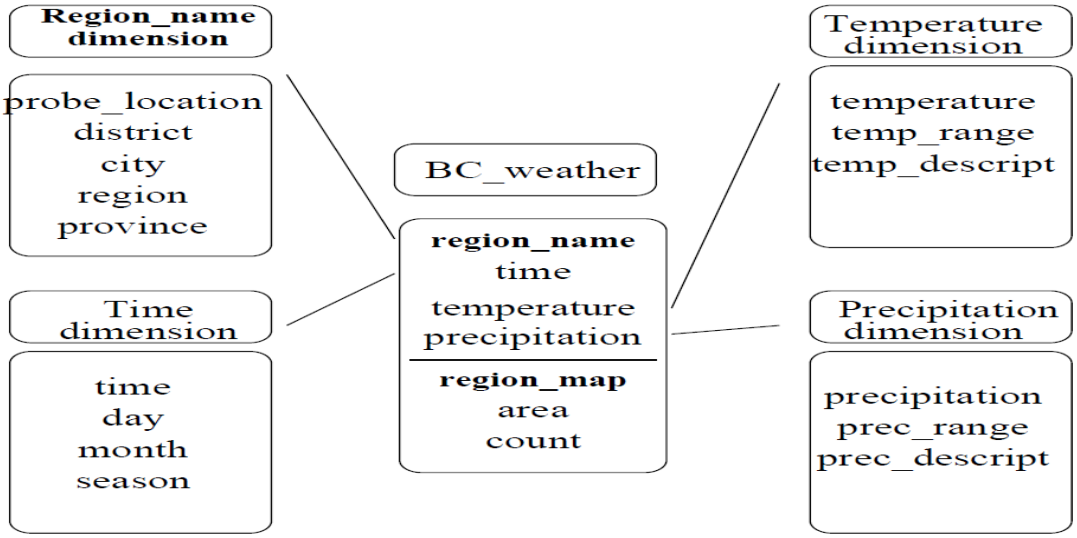


Figura 1 – *Data warehouse* utilizado no modelo de Han et al.

Este *data warehouse* espacial possui, um esquema em estrela com quatro dimensões: "Region_name", "Temperature", "Time" e "Precipitation" e uma tabela de factos com três medidas: "region_name", "area" e "count". Analisando em detalhe cada uma das dimensões e das medidas, facilmente se verifica que o *data warehouse* espacial, possui uma dimensão espacial, a dimensão "Region_name", visto que os dados da podem ser representados cartograficamente. Em relação às medidas, "area" e "count" são medidas numéricas, enquanto a medida "region_map" corresponde a uma medida espacial que representa uma coleção de pontos espaciais da região correspondente.

Para além destes conceitos, a exploração dos dados também foi abordada neste modelo. Os autores dão alguns exemplos de como generalizações podem levar dimensões espaciais a obterem resultados não espaciais, como é o caso de um estado dos 'Estados Unidos da América' que possui a capacidade de ser representado num mapa. Contudo, a sua generalização pode ter como resultado um valor não espacial como '*Pacific Northwest*'.

Porém, o modelo proposto possui uma limitação: os investigadores supuseram que a origem dos dados são fontes homogéneas, não tendo em conta o problema de atualizar o *data warehouse* e manter a sua consistência devido a esse facto.

2.2.2 A Proposta de Rivest et al. (2001).

Rivest, et al. (2001) usaram como base da sua investigação o modelo proposto por Han et al (1998). Ambos partilham a mesma ideia em relação às dimensões existentes, contudo, as principais diferenças entre os modelos em questão surgem nas medidas, tendo sido proposta uma extensão à sua definição. Rivest et al. (2001) propuseram 3 medidas espaciais em vez das 2 medidas, propostas anteriormente por Han et al. (1998).

A primeira medida consistia numa forma, ou um conjunto de formas geométricas, obtida a partir da combinação de múltiplas dimensões espaciais, ou seja, um conjunto de coordenadas que

necessitam de uma operação geométrica, e em que o seu resultado consiste numa nova forma geométrica. Os autores dão como exemplos casos das fronteiras políticas ou das bacias hidrográficas.
A segunda medida proposta, resulta da computação de medidas espaciais ou operadores topológicos, sendo este resultado guardado nas células de um cubo. Os exemplos desta medida são, por exemplo, superfícies ou distâncias.

A terceira, e última medida, consiste num conjunto de apontadores para formas geométricas guardadas numa noutra estrutura ou *software*.

Fazendo uma comparação das medidas propostas por este modelo, com o modelo proposto por Han et al. (1998), facilmente se verifica que foi proposta uma nova medida que tem em conta as medidas espaciais, que são calculadas usualmente através de métricas ou de operadores lógicos.

2.2.3 A Abordagem de Malinowski e Zimany (2008)

Malinwoski e Zimany (2008) propuseram uma abordagem diferente para o modelo conceptual, tendo como base o paradigma de modelação Entidade-Relacionamento, resultando no modelo *MultiDimER*. Para uma melhor compreensão desse modelo, eles utilizaram como exemplo a análise de custos de manutenção de uma autoestrada, tendo como eixos de análise, os municípios, os estados, o tempo, a autoestrada, em que se dividem em secções e que, por sua vez, se dividem em segmentos.

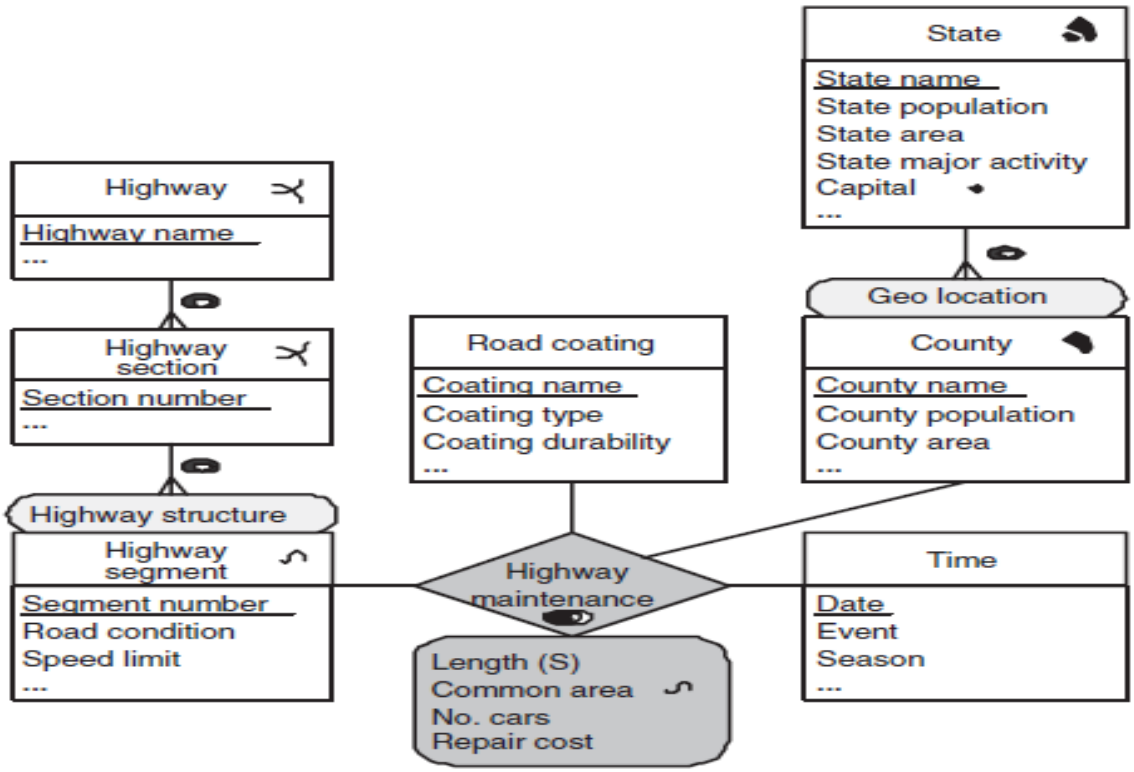


Figura 2 – *Data Warehouse* utilizado no modelo de Malinowski e Zimany (2008).

Inicialmente, os investigadores propuseram a definição de níveis espaciais, como sendo "um nível para o qual a aplicação precisa para manter as suas características espaciais", podendo essas características serem representadas por pontos, linhas, áreas, ou por um conjunto de essas representações. Como podemos observar, o data warehouse da figura 2 possui cinco níveis espaciais: "County", "State", "Highwat segment", "Highway section" e "Highway". A manutenção das características espaciais em todos os níveis é conseguida através da sua geometria, utilizando-se para a sua representação a notação MADS (Figura 3).

Figura 3 – Notação MADS utilizada para representar a geometria das entidades [1].

Em seguida, Malinowski e Zimmany (2008) definiram o conceito de hierarquia espacial, como sendo uma hierarquia que possui, pelo menos, um nível espacial. Cada hierarquia pode ter diferentes estruturas, dependendo dos critérios de análise. Para os distinguir, é utilizado o nome do critério, tal como pode ser constatado pelos critérios presentes na figura 2: "Geo location" e "Highway structure". O relacionamento entre dimensões possui uma cardinalidade que, para além de indicar o número de elementos de uma dimensão que se relaciona com uma outra, influencia também o tipo de hierarquia que se pode definir entre eles. Por exemplo, ao se analisar o relacionamento entre "County" e "State" presente na figura 2, verifica-se que possui um relacionamento de um-para-muitos, o que revela que um estado possui vários municípios. Numa hierarquia espacial, dois níveis espaciais estão relacionados através de um relacionamento topológico, sendo esse relacionamento representado por pictogramas (Figura 4).

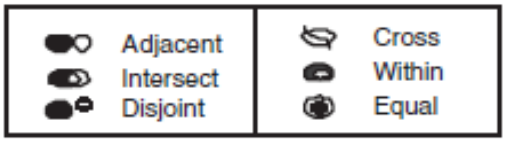


Figura 4 – Notação MADS utilizada para representar as relações entre as entidades [1].

Tendo em conta as situações do mundo real, os vários tipos de relacionamentos topológicos foram também considerados pelos autores. Utilizando novamente o exemplo de "State" e "County", verifica-se pelo pictograma utilizado na figura 4, que a geometria de cada município está incluída na geometria do estado correspondente.

O seguinte conceito introduzido pelos autores, foi o relacionamento entre os factos, *fact relationship*, que representa o relacionamento entre a tabela de factos e as dimensões de primeira linha. Este relacionamento pode ser espacial, se pelo menos duas das dimensões forem espaciais. Analisando novamente a figura 2, mais concretamente o relacionamento "Highway maintenance", consegue-se ver que duas das três dimensões presentes são, de facto, dimensões espaciais ("Highway segment" e "County"). Nos relacionamentos espaciais entre os factos, poderá necessitar a inclusão de um predicado espacial (Figura 4).

Após a explicação sobre os relacionamentos entre as dimensões, os autores focam a sua atenção nas medidas que podem ser incluídas num relacionamento. Tal como Hans et al. (1998), os autores reconhecem a existência de dois tipos de medidas, medidas não espaciais (*thematic measures*) e medidas espaciais. As medidas não espaciais são usadas para realizar análises quantitativas e são valores numéricos, enquanto que as medidas espaciais podem ser representadas por geometrias ou ser calculadas utilizando operadores espaciais, podendo esses valores calculados serem, por exemplo, uma distância ou uma área. Caso a medida seja calculada utilizando operadores espaciais, os autores utilizam o símbolo "(S)", tal como se pode ver no caso da medida comprimento (*Length*) (Figura 2), que representa o comprimento da parte do segmento da autoestrada que pertence a um dado município. A outra medida especial apresentada na figura 2, é a medida que representa a área comum (*Common area*). Esta corresponde à parte que eles têm em comum, sendo identificado o tipo de geografia resultante, utilizando-se os pictogramas apresentados anteriormente. Para todas as medidas é necessário especificar a operação a utilizar para o seu cálculo. Neste caso, os autores assumem que, por omissão, a operação a realizar para medidas com representação numérica é a operação de união geométrica para medidas com representação geométrica.

2.2.4 O Modelo de Bâazaoui Zghal et al. (2003).

Bâazaoui Zghal et al. (2003) começaram por clarificar os aspetos relacionados com a informação espacial e as suas características, uma vez que, esta informação, envolve tipos de dados tão específicos e, como tal, devem ser tidos em conta na construção de um *data warehouse* espacial.

Os dados geográficos podem ser objetos espaciais, tais como coordenadas de um objeto ou objetos não espaciais. Contudo, estes objetos não espaciais podem representar entidades geográficas, sendo, portanto, uma descrição dessas entidades, como por exemplo, o nome de uma cidade. Aquando da aquisição e integração dos dados espaciais, as características especiais desses dados devem ser tidas em conta: a riqueza semântica, a precisão dos procedimentos e a multiplicidade das representações geométricas. Tendo em conta todas estas características, os investigadores referidos

sugerem um modelo para a generalização de todos os dados geográficos, recorrendo para isso aos diagramas de classe da linguagem de modelação UML.

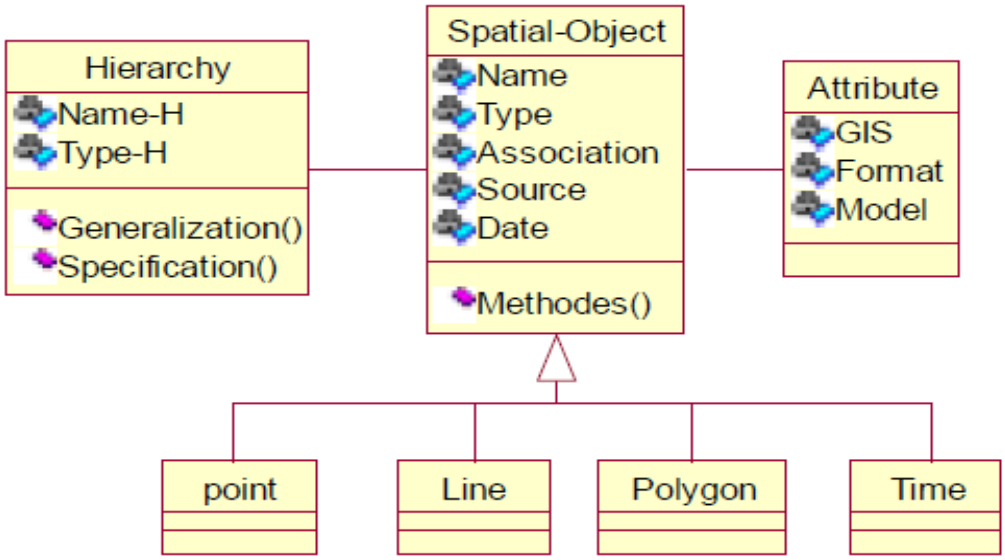


Figura 5 – Modelo de generalização dos dados geográficos.

Usualmente, um objeto geográfico é caracterizado pelo seu nome, pela sua fonte e pela sua data. Esta, permite ter uma uma noção das sucessivas variações do objeto ao longo do tempo. Por sua vez, cada um desses objetos, pode ser composto por outros objetos – objetos compostos. Estes são identificados pelo seu nome, tipo, tamanho e o local, sendo o local um endereço físico no qual se pode encontrar o objeto. Para além destas características, os objetos podem possuir objetos inferiores, que podem ser pontos, linhas, polígonos ou tempo, que determinam os métodos de análise a usar pelas ferramentas SOLAP.

Após uma análise à informação geográfica, Bâazaoui Zghal et al. (2003) propuseram um meta modelo para a construção de *data mart* espaciais. Cada *data mart* espacial é caracterizado pelo seu esquema multidimensional, formado por dimensões e medidas, onde, por sua vez, podem ser espaciais ou não espaciais. Tendo em conta estes aspetos, propuseram um novo modelo, recorrendo novamente ao UML para a sua especificação (Figura 6).

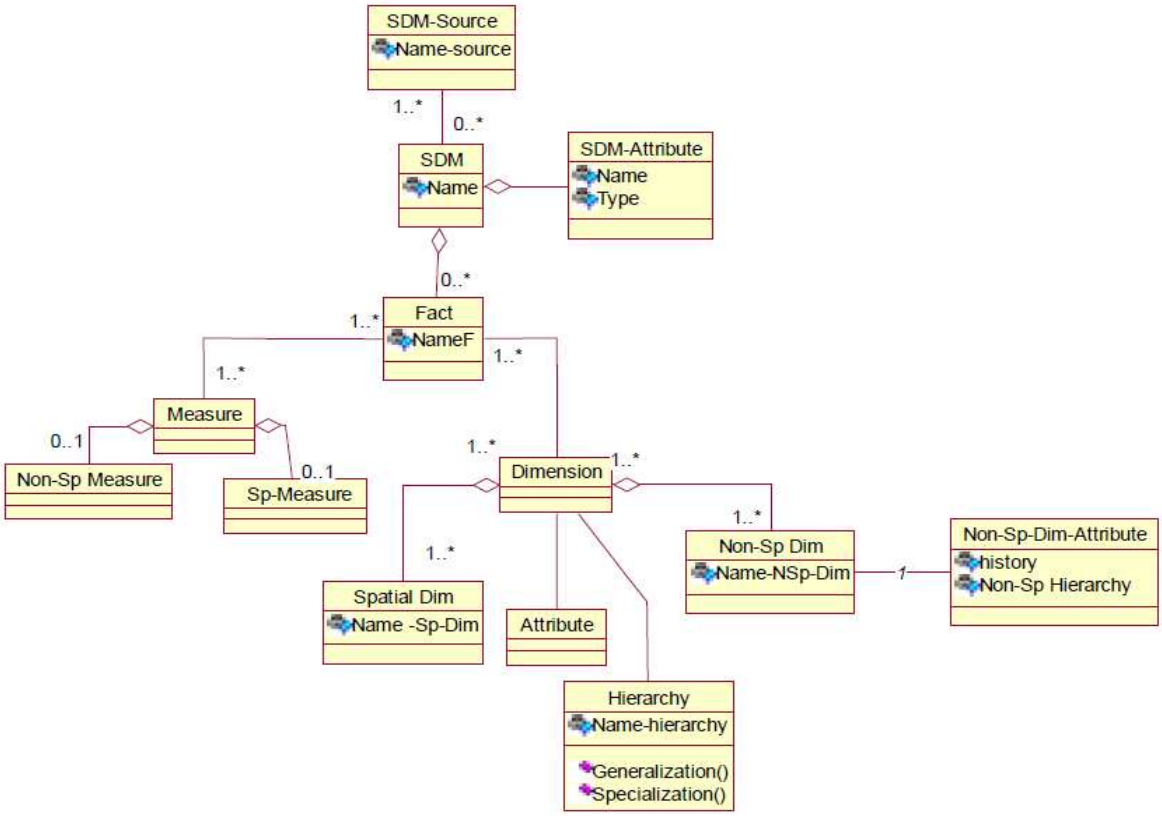


Figura 6 – Meta modelo para construção de *data mart* espacial.

Os dados espaciais, oriundos de fontes de dados externas, são utilizados para a criação do *data mart* espacial, representado nesse modelo pela classe SDM (*Spatial Data Mart*). Cada *data mart* especial é caracterizado pelo seu nome, um atributo que nunca é alterado ao longo da vida do objeto. Para além desta característica, pode ainda ser composto por uma ou mais tabelas de factos. Cada tabela de factos pode conter várias medidas e estar relacionada com várias dimensões, podendo estas serem, por sua vez, espaciais ou não. Para a modelação das hierarquias, os investigadores propõem uma classe Hierarquia (*Hierarchy*), que é necessária para manter um histórico acerca dos vários níveis da hierarquia para as dimensões e medidas devido à natureza dos dados espaciais. Cada dimensão e medida possui os seus parâmetros organizados de acordo com uma determinada hierarquia. Depois, após uma breve explicação do modelo proposto, os investigadores identificaram 4 tipos de medidas: espacial não geométrica, espacial geométrica para não geométrica, totalmente geométrica e temporal em vez das 3 dimensões identificadas anteriormente por Han. As medidas referidas são:

- **Temporal**, que caracteriza a organização do tempo do caso de estudo de acordo com as necessidades do utilizador. Como sabemos, as referências temporais são elementos fundamentais dos sistemas de informação. É difícil imaginar um sistema de informação sem datas e durações das atividades em análise. Através das referências temporais podemos prever acontecimentos futuros ou explicar o estado atual da situação.
- **Dimensão Geométrica Não Espacial**, constituída unicamente por dados não geométricos. Estes dados, apesar de não serem geométricos, podem conter informação capaz de ser localizada no espaço. Como exemplo, a dimensão pode conter nomes de municípios, que por sua vez, podem ser generalizados em país. Contudo a informação do país pode não ser geométrica. Estes casos podem ser implementados, desde que não haja a necessidade de representar países cartograficamente.
- **Dimensão Espacial Geométrica para Não Geométrica**, consiste numa dimensão que, através de uma generalização, passa de dimensão geométrica para não geométrica.
- **Dimensão Totalmente Geométrica**, que consiste numa dimensão em que todos os níveis da hierarquia são dados geométricos e mantêm-se geométricos mesmo após generalizações. Os investigadores dão como exemplo desta dimensão os polígonos de análise da precipitação, que são elementos espaciais e mesmo efetuando generalizações, passando a análise de 10-100 milímetros para 100-200 milímetros, as suas características espaciais mantêm-se.

Em termos de medidas do modelo, os investigadores identificaram 2 tipos de medidas:

- **Medidas Numéricas -** os dados destas medidas são unicamente do tipo numérico.
- **Medidas Espaciais -** os dados destas medidas são um conjunto de pontos para objetos espaciais.

2.3 Síntese de Características dos Modelos Analisados

Neste capítulo foram analisados os modelos mais significativos propostos para a modelação dimensional de um *data warehouse* espacial, tendo-se dado especial atenção aos elementos de maior relevância de cada um dos modelos apresentados, como foram, por exemplo, os casos das dimensões, das medidas e das hierarquias. Em seguida apresenta-se uma breve síntese das características dos vários modelos que foram apresentados e analisados (Tabela 1).

	Han et al. (1998)	**Rivest et al. (2001)**	**Malinowski and Zimany (2008)**	**Bâazaoui Zghal et al. (2003)**
Hierarquias explícitas nas dimensões	Não possui	Não possui	Possui	Possui
Tratamento simétrico das medidas e dimensões	Permite	Permite	Permite	Não permite
Relações de muitos para muitos entre dimensões e factos	Permite	Permite	Permite	Permite
Manipular alterações e tempo	Permite	Permite	Permite	Permite

Tabela 1 – Síntese de características dos vários modelos multidimensionais para DW Espaciais.

Analisando a tabela 1, e tendo em conta toda a informação obtida dos vários modelos estudados, facilmente se conclui que o modelo mais completo apresentado até à data, é o modelo proposto por Malinowski e Zimányi (2008). Apesar de ser bastante completo, este modelo baseia-se num modelo entidade-relacionamento, sendo indicado, em particular, para a modelação de bases de dados relacionais. Contudo, este modelo fornece uma base bastante concreta para a obtenção de informação de todos os assuntos relacionados com dimensões, medidas e relacionamentos, entre outros aspetos, constituindo assim uma boa fonte de informação a ter em conta na modelação de um *data warehouse* espacial.

Capítulo 3

Implementação de um DW Espacial

3.1 O processo de Implementação

O desenho e implementação de um *data warehouse* convencional é um processo complexo. Contudo, existem metodologias definidas amplamente testadas e comprovadas. O mesmo, porém, não acontece em relação aos *data warehouses* espaciais. Podemos facilmente comprovar isso, tendo em conta a existência de tão poucas publicações relacionadas com este tema. Para além deste facto, nas publicações disponíveis, e tanto quanto pudemos constatar, não existe um consenso em relação ao processo a seguir, aos passos a realizar, visto que alguns autores consideram que a implementação de um *data warehouse* deve seguir os mesmos passos que a implementação de uma base de dados relacional, enquanto que outros utilizam um outro tipo de metodologia (Malinowski e Zimányi, 2008). Neste trabalho, a análise das várias fases de construção de um *data warehouse* espacial terá como base o modelo proposto por Ralph Kimball e Ross (2002), para a implementação de *data warehouses* convencionais, e o modelo proposto por Malinowski e Zimányi (2008) para a implementação de *data warehouses* espaciais, bem como, obviamente, todo o conhecimento obtido através dos vários modelos analisados previamente.

3.2 Abordagens ao Desenho e Implementação de DWE.

Segundo Malinowski e Zimányi (2008) existem duas abordagens possíveis para o desenho de um *data warehouse* espacial. Essas abordagens foram designadas por *early inclusion* e *late inclusion*. A abordagem *early inclusion* (ou inclusão precoce), consiste em analisar os diversos elementos espaciais e introduzi-los no esquema do *data warehouse* espacial, desde a fase inicial do seu desenho. A abordagem *late inclusion* (ou inclusão tardia) consiste na criação de um esquema conceptual não espacial e, após a sua conclusão, fazer a introdução dos vários elementos espaciais no esquema resultante.

A escolha de uma das abordagens referidas para o desenho do esquema do *data warehouse* espacial depende, em grande medida, do conhecimento dos utilizadores sobre os conceitos relacionados com os dados espaciais e sobre a sua existência nos sistemas já desenvolvidos. Caso o utilizador seja um conhecedor dos vários aspetos relacionados com a análise de dados espaciais, a abordagem aconselhada será a inclusão dos elementos espaciais precocemente, permitindo, desta forma, que todos os problemas relacionados com os dados espaciais sejam tratados desde o início. Caso o utilizador não tenha esse conhecimento ou os dados geográficos tenham de ser obtidos através de sistemas externos pelo facto de a organização não possuir os mesmos nos seus sistemas operacionais, a inclusão tardia dos elementos espaciais surge como uma boa opção, uma vez que permite a criação rápida de um esquema base no qual será incluído no futuro o suporte aos dados espaciais.

3.3 Tipos de Abordagem para Elementos Espaciais

Tendo em conta as duas hipóteses de inclusão apresentadas para elementos espaciais, Malinowski e Zimányi (2008) propuseram três abordagens possíveis para o levantamento de requisitos, nomeadamente: *analysis-driven*, *source-driven*, e *analysis/source-driven*. Vejamos cada uma delas nas secções que se seguem.

3.3.1 A Abordagem *Analysis-Drive*

Segundo Malinowski e Zimányi (2008), a abordagem *analysis-driven* consiste em identificar os utilizadores chave do negócio, as suas necessidades e, por fim, construir uma documentação que dê suporte à construção do esquema conceptual do *data warehouse* espacial.

Caso se opte por uma inclusão de elementos espaciais, desde o momento inicial, os utilizadores podem indicar quais são dados espaciais que necessitam para efetuar as suas análises. Posto isto, o primeiro passo consiste assim em identificar utilizadores de vários níveis da empresa, para que a análise de requisitos vá de encontro aos objetivos da organização. Esses utilizadores vão ajudar a equipa de desenvolvimento a perceber as necessidades da organização que, por sua vez, vão colecionar toda a informação de forma a poder produzir um documento com todas as especificações necessárias. Após concluído esse documento, começa-se o desenvolvimento do esquema conceptual do *data warehouse* espacial. Tendo um esquema inicial do *data warehouse* já preparado, é necessário verificar que se os sistemas fonte contêm a informação necessária para dar o suporte espacial requerido. Caso contrário, será necessário obter esses dados de outras fontes externas. Por fim, faz-se o desenvolvimento do esquema final contendo todos elementos do *data warehouse*, quer estes sejam provenientes de fontes internas ou de fontes externas.

A outra hipótese de aproximação ao problema discutido, consiste na introdução dos dados espaciais apenas no final, sendo que dessa forma o processo de levantamento de requisitos segue os passos do desenvolvimento de um *data warehouse* convencional. Contudo, após a elaboração de um esquema final, é necessário, junto dos utilizadores, tentar obter informação acerca dos dados espaciais para que seja possível suportar futuros processos de análise espacial.

3.3.2 A Abordagem *Source-Driven*

A abordagem *source-driven* tem como principal foco as fontes de dados das organizações e, tal como na abordagem anterior, os elementos espaciais podem ser introduzidos desde o início do projeto ou apenas no fim. Esta opção, entre o *timing* de introdução dos dados espaciais, é dependente do tipo

das bases de dados que as organizações possuem. Caso sejam bases de dados espaciais, a melhor opção será introduzir, desde o início, os elementos espaciais no esquema de *data warehouse*. Caso contrário, a melhor opção será a introdução desses elementos no fim.

Na situação das fontes de dados possuírem dados espaciais, a abordagem mais correta será optar por incluir os dados espaciais no esquema conceptual desde do início. Inicialmente, identificam–se quais as fontes de dados e respetivos elementos espaciais que precisamos, de forma a avaliar a necessidade de recorrer a fontes externas para obtenção de dados espaciais. Após identificadas as fontes de dados, identificam-se os esquemas multidimensionais que se pretende criar. Por fim, constrói-se um documento com os metadados requeridos e todas as especificações necessárias.

Caso as fontes de dados não possuam dados espaciais, a abordagem mais adequada consiste em incluir os dados espaciais no fim do esquema conceptual estar concluído. O desenho do esquema conceptual segue a estrutura dos *data warehouses* convencionais. No final do processo de construção do esquema faremos a introdução dos respetivos dados espaciais de forma a que o esquema concebido seja capaz de fornecer suporte aos processos de análise espacial.

3.3.3 A Abordagem *Analysis/Source Driven*

A abordagem *analysis/source driven* consiste numa combinação das duas abordagens analisadas anteriormente, as quais podem ser aplicadas de forma paralela ou combinada. Caso se opte por aplicar paralelamente as duas abordagens, obtêm-se dois documentos de especificação no final, um para cada uma das abordagens. Caso se opte por combinar as abordagens, após o desenho do esquema inicial de cada abordagem, analisam-se os dois esquemas, de forma a se obter no final do processo um só esquema e, consequentemente, um só documento de especificação.

3.4 Levantamento de Requisitos

O processo de levantamento de requisitos neste domínio de trabalho é de supra importância, pelo simples facto de permitir identificar os aspetos fundamentais para o desenho e implementação de um *data warehouse*. Um dos elementos do documento de especificação, resultante da análise de

requisitos para a implementação de um *data warehouse* espacial, consiste na informação sobre os requisitos e necessidades dos utilizadores. Esta informação permite aos responsáveis pela implementação do projeto, definir um plano que permita corresponder às necessidades dos utilizadores, de forma eficaz, poupando, assim, dinheiro e tempo. Para além desse elemento, devem constar no documento também as análises de viabilidade, uma relação detalhada dos custos e o levantamento da informação existente nas fontes de dados (Kimball e Ross, 2002). Este último elemento assume, aqui, particular relevância, quando se está a desenhar um sistema de *data warehousing* espacial. O levantamento dos dados existentes nos sistemas operacionais permite identificar, desde o início, a existência de dados espaciais e se são suficientes para corresponder às necessidades dos utilizadores (Malinowski e Zimányi, 2008). Caso estes não existam, é necessário recorrer a sistemas de dados externos para se obterem, para que o futuro *data warehouse* espacial seja capaz de representar a sua informação em mapas.

3.5 Modelação Dimensional

Após o levantamento de requisitos e da validação e aceitação do documento resultante por parte dos elementos da organização, dá-se início ao processo de modelação dimensional. Tal como foi referido anteriormente, na literatura não se encontra um consenso, pelo menos que tivéssemos detetado, em torno desta questão. Apesar de o modelo proposto por Kimball e Ross (2002) ser específico para *data warehouses* convencionais é possível fazer a sua conjugação com outras propostas presentes na literatura, com o objetivo de desenvolver um esquema dimensional bem estruturado, capaz de fornecer suporte a análises espaciais.

3.5.1 Construção da Matriz Dimensional

De acordo com o modelo proposto por Kimball e Ross (2002), após concluído o levantamento de requisitos, que fornece informação acerca das necessidades de análise por partes dos agentes de decisão e da informação que consta nas fontes de dados, o próximo passo consiste em utilizar essa informação para identificar os vários *data marts* e dimensões existentes. Após a conclusão da análise,

constrói-se a matriz dimensional de forma a identificar dimensões partilhadas por vários data marts, podendo, dessa forma, efetuar uma melhor escolha sobre os processo(s) de negócio a modelar. A seguir é apresentado um exemplo de uma matriz dimensional.

Data Mart / **Dimensão**	**TF Encomendas**	**TF QoS**
Tempo	X	X
Encomenda	X	X
Cliente	X	X
Zona	X	
Produto	X	
Local	X	

Tabela 2 – Exemplo de uma matriz de dimensional.

Cada linha da matriz dimensional representa um processo de negócio e cada coluna representa uma dimensão presente no *data warehouse*. Analisando a matriz, verifica-se que existem dimensões partilhadas entre os dois *data marts*. A identificação desta dimensões partilhadas no inicio do projeto permite definir um padrão para as dimensões, de forma a que as várias equipas que participam na construção do projeto possam utilizar a dimensão previamente definida, com características capazes de fornecer suporte aos vários *data marts*, mantendo assim a consistência do projeto e evitando a redundância de dimensões que degrada a performance do *data warehouse*.

3.5.2 Definição do Grão

Após identificado o processo a modelar, é necessário definir o grão. A definição do grão, ou da granularidade dos dados, indica o grau de detalhe dos factos. Para a sua definição deve ser utilizada a informação do processo de negócio mais atómica possível, ou seja, informação que não possa ser divida ainda mais. Por norma, o grão representa o nível de detalhe mais pequeno do processo de negócio (Kimball e Ross, 2002).

Como exemplo de grão de um *data mart* temos por exemplo a venda de um produto, a um cliente, numa determinada data e numa determinada loja. A granularidade é bastante importante, visto que, um grão com mais detalhe permite ao utilizador visualizar a informação em qualquer nível de agregação. Contudo, esta opção pode levar a que seja necessário armazenar mais informação para

fornecer suporte a estas análises, o que leva a uma perda de performance. Por outro lado, se o nível de granularidade definido for muito alto, o utilizador não terá capacidade de efetuar consultas mais detalhadas.

3.5.3 Definição das Dimensões e Medidas

Depois de escolhido o grão, o próximo passo da modelação é a escolha das dimensões que se aplicam a cada linha da tabela de factos, ou seja, vão representar as possíveis perspetivas de análise sobre os factos e a sua descrição. A identificação do grão permite a identificação das dimensões de primeira linha, necessárias para fornecer suporte ao grão. Contudo, é sempre possível adicionar novas dimensões, desde que essas adições não violem o grão, caso contrário é necessário analisar o grão novamente para que acolha as novas dimensões (Kimball e Ross, 2002).

As dimensões fornecem contexto e significado às medidas da tabela de factos e por fim, permitem responder a questões sobre os factos como por exemplo: quando, como, quem, onde, entre outras. Cada registo da tabela da dimensão corresponde a um determinado elemento, isto é, uma linha da tabela da dimensão Cliente corresponde à informação de um determinado cliente. O nível de detalhe desta informação depende muito do nível de detalhe das análises do utilizador. No processo de definição das dimensões é necessário ter atenção a duas questões, desempenho e compreensão dos dados (Michelle, A., 2007). Dimensões com bastantes atributos permitem perceber melhor a informação relativa à dimensão, mas por outro lado, implica mais dados armazenados e uma degradação da performance.

Após a definição das dimensões, precisamos de identificar as medidas que vão integrar uma tabela de factos. Novamente, o grão é bastante importante porque analisando a sua definição facilmente se identificam várias medidas da tabela de factos. Porém, podem surgir medidas que levem a que a própria definição do grão e a escolha das dimensões sejam analisadas (Kimball e Ross, 2002).

Após a conclusão do método dos quatro passos é possível construir um diagrama da tabela de factos com os seus elementos e as medidas. Contudo, para cada medida pertencente à tabela de factos é necessário identificar se a mesma é agregável ou não, e por que função é agregável caso seja possível. Para se perceber melhor este conceito é usado como exemplo, a medida quantidade vendida de um produto de uma tabela de fatos de vendas. Esta medida por norma é agregável, isto é, fazendo

uma generalização com o intuito de saber a quantidade de um determinado produto que foi vendida na cidade 'X', o valor quantidade vai ser agregado utilizando para isso a função SUM() para obter o resultado. Como exemplo de uma medida não agregável temos o custo de um produto.

3.5.4 Caracterização das dimensões e seus atributos

No processo de caracterização das dimensões de um data warehouse espacial é necessário identificar os seus atributos, tendo especial atenção aos atributos com características espaciais. A caracterização dos atributos espaciais pode ser efetuada recorrendo à notação MADS (Malinowski e Zimányi, 2008), onde para além de facilmente se identificar os atributos com características espaciais é possível identificar o seu tipo de geometria. Após identificadas as dimensões espaciais e a sua geometria é necessário identificar se as dimensões são primitivas, compostas ou híbridas (Fidalgo et al., 2010), visto que a sua construção vai depender do seu tipo.

Dimensão Geográfica Primitiva

Segundo Fidalgo et al. (2010) a construção de uma dimensão espacial primitiva vai depender do suporte que a base de dados adotada fornece em termos de dados espaciais. Caso a base de dados não tenha a capacidade de guardar dados espaciais, é necessário guardar numa dimensão todos os valores das coordenadas geográficas, seguindo-se, por exemplo, uma abordagem relacional, tal como apresentada na Figura 7

Figura 7 - Exemplo de representação de dimensão geográfica primitiva segundo uma abordagem relacional.

Porém, caso a base de dados possua a capacidade de armazenar tipos de dados geográficos, a abordagem apropriada consiste em guardar toda a informação geográfica no próprio atributo geométrico. Uma dimensão deste tipo pode, assim, ser composta, pela sua chave e pelo atributo atrás referido – veja-se o exemplo apresentado na Figura 8.

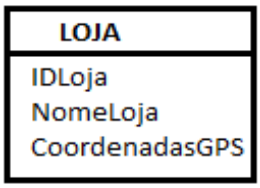


Figura 8 - Exemplo de representação de dimensão geográfica primitiva.

As dimensões geográficas primitivas lidam apenas com a informação geográfica e, devido a isso, não são apropriadas para consultas OLAP. Todavia, estas dimensões, são essenciais, entre outros aspetos, para fornecer suporte a operações geográficas.

Dimensão Geográfica Composta

Para fazer a construção de uma dimensão geográfica composta é necessário realizar uma análise para perceber a hierarquia dos dados espaciais. Caso isso não aconteça, pode conduzir a uma situação na qual ocorra redundância de dados, contendo as referências geográficas, que, por sua vez, implica um aumento significativo dos custos de armazenamento. Para exemplificar este tipo de dimensões, Fidalgo et al. (2010) utilizaram a hierarquia "Region → Property →Lot" na construção da dimensão composta. O conhecimento da hierarquia permite, também, que sejam usados métodos convencionais para agregar os vários níveis, em vez de se recorrer a métodos de indexação espacial.

Analisando o exemplo, facilmente se conclui que a dimensão geográfica composta possui um nome para cada elemento da sua hierarquia, bem como uma chave estrangeira para a dimensão primária correspondente de cada um desses elementos (Figura 9).

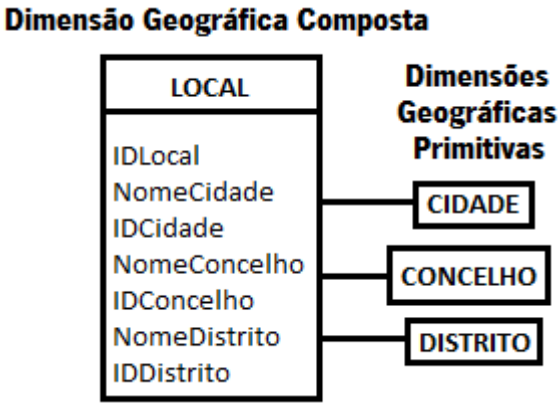


Figura 9 - Exemplo de uma dimensão geográfica composta.

Dimensões Híbridas

As dimensões híbridas podem ser decompostas em três tipos (Fidalgo et al., 2010): *micro*, *macro* e *joint*.

Uma micro dimensão híbrida é composta por dados espaciais e por dados convencionais. Contudo, os dados espaciais representam o grão da informação geográfica, isto é, são normalmente constituídos por pontos que representam o nível mais pequeno de detalhe, tal como acontece com as ruas ou com as casas, por exemplo. Devido a este facto, esta informação é raramente partilhada com outras dimensões, podendo haver exceções que levem a que essa informação seja partilhada ou que seja necessário fazer a sua replicação (exemplo: um empregado que também é um cliente). Em suma, este tipo dimensão tem atributos todos convencionais e uma chave estrangeira para a dimensão primitiva com a informação geográfica (Figura 10).

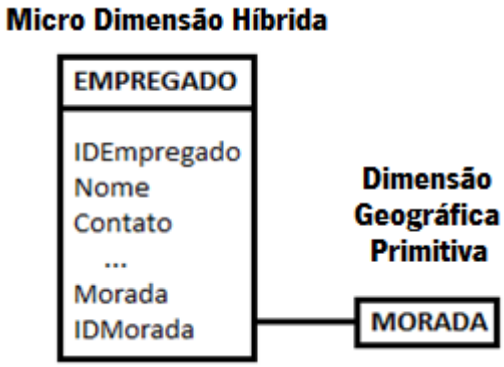


Figura 10 - Exemplo de uma micro dimensão híbrida.

Quanto às macro dimensões híbridas, estas partilham os seus dados geográficos com outras dimensões por exemplo: Pais, Região, Distrito e Cidade. Existem duas abordagens para a construção de uma macro dimensão híbrida.

A primeira abordagem consiste em criar uma dimensão com todos os atributos convencionais, mais uma chave estrangeira para uma dimensão geográfica composta (Figura 11). A utilização de uma dimensão composta, para além de reduzir o número de chaves estrangeiras na dimensão híbrida, permite também que seja utilizada como uma mini dimensão ou como uma *role-playing dimension*.

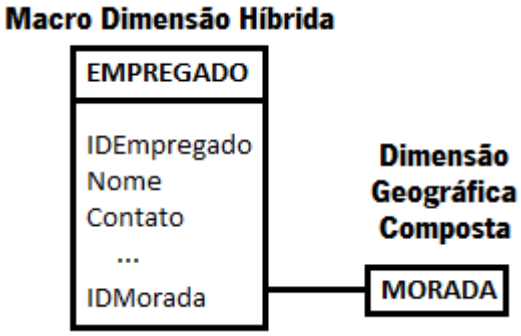


Figura 11 - Exemplo da 1ª abordagem possível para macro dimensões híbridas.

A outra abordagem possível consiste em criar uma dimensão com todos os elementos convencionais, mais uma chave estrangeira para a dimensão primitiva por cada elemento geográfico. Veja-se um destes casos na figura 12.

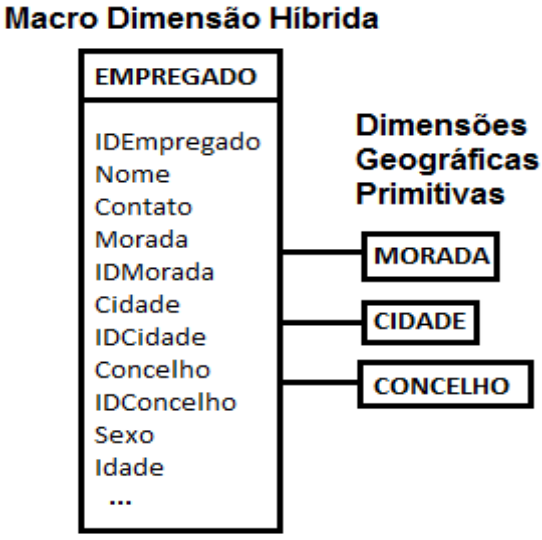


Figura 12 - Exemplo da 2ª abordagem possível para macro dimensões híbridas.

Uma joint dimensão híbrida é uma combinação de uma micro dimensão híbrida com uma macro dimensão híbrida, ou seja, é uma dimensão composta por atributos convencionais, uma chave estrangeira para uma dimensão geográfica composta e uma chave estrangeira para uma dimensão geográfica primitiva (Figura 13).

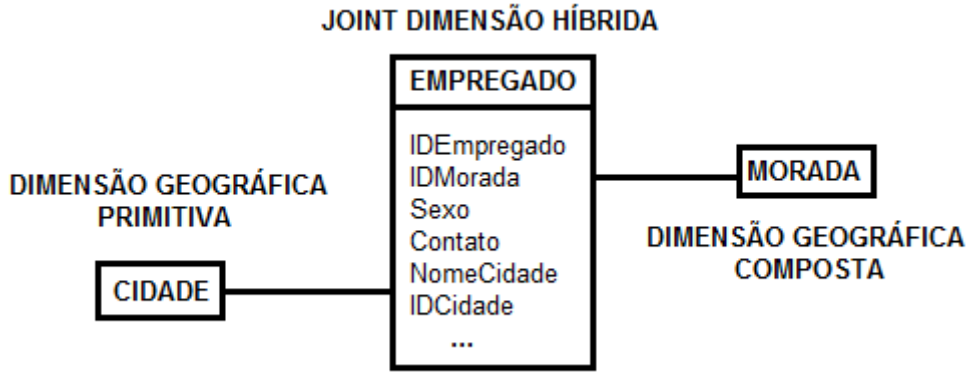


Figura 13 - Exemplo de uma joint dimensão híbrida.

3.5.4 Desenho lógico do *data mart*

Concluídos todos os passos anteriores, é possível construir o esquema logico do *data mart*, sendo que o esquema resultante pode ser um esquema em estrela ou um esquema em floco de neve, dependendo da existência de dimensões de segunda linha.

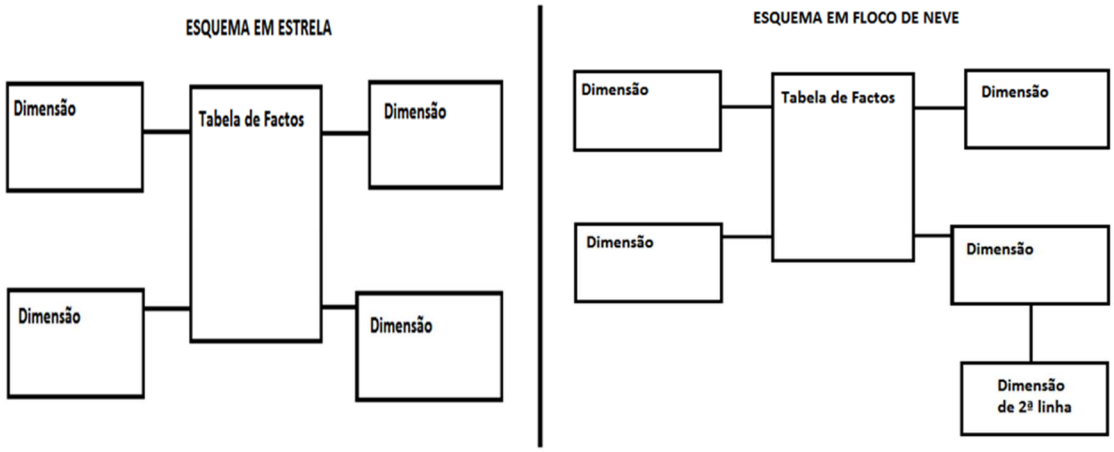


Figura 14 - Exemplo de um esquema em estrela e de um esquema em floco de neve.

3.6 As Bases de Dados

O suporte que uma base de dados fornece aos dados geográficos influencia o desenho das dimensões e seus atributos, mas, além disso, influencia também a sua construção. Contudo, a maioria das bases dados existentes no mercado possuem dois tipos de dados para armazenar dados espaciais: o tipo *geometry* e o tipo *geography*. O tipo de dados *geometry* representa um conjunto de dados num sistema de coordenadas Euclidiano (plano), enquanto o tipo de dados *geography* representa os dados num sistema de coordenadas de globo terreste (Alastair Aitchison, 2012). A escolha entre o tipo *geometry* e o tipo *geography* prende-se unicamente com o tipo de superfície onde serão visualizados os dados. Na figura 15 são apresentados os vários subtipos de dados suportados.

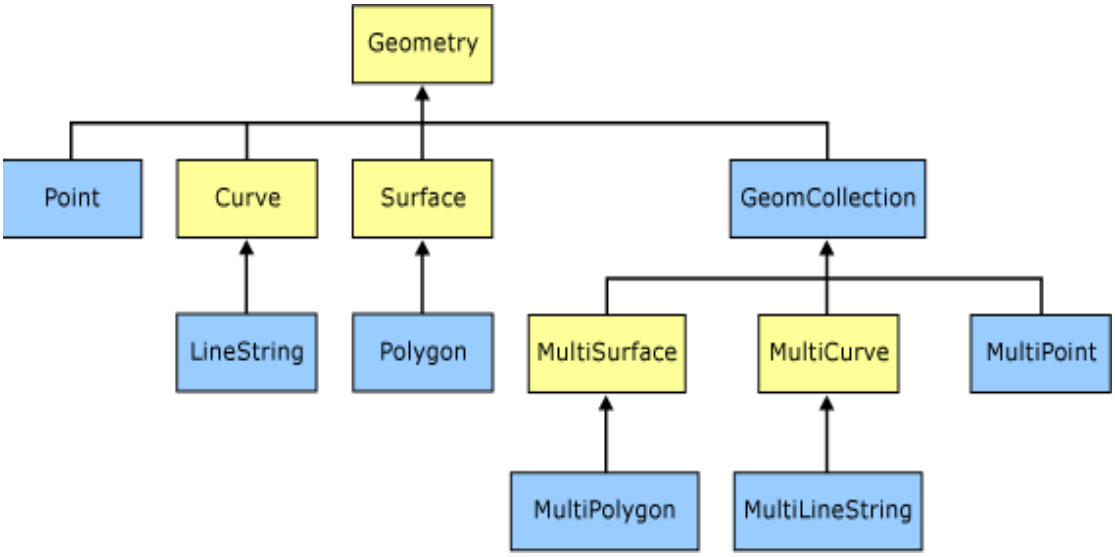


Figura 15 – Esquema dos vários tipos de dados (Fonte: http://technet.microsoft.com/en-us/library/bb964711(v=sql.105).aspx).

De seguida, uma breve apresentação dos vários tipos de dados geométricos. São eles (Alastair Aitchison, 2012):

GeomCollection, que é um objeto geométrico constituído por uma coleção de outros objetos geométricos.

Point, que é um objeto geométrico que não possui dimensão e que representa uma localização simples num sistema de georreferenciado. Um ponto possui um valor de coordenada 'X' e um valor de coordenada 'Y', sendo que num sistema georreferenciado essas coordenadas correspondem à latitude e à longitude.

Curve, que é um objeto geométrico unidimensional que pode ser classificado como simples, se não passar duas vezes pelo mesmo ponto, ou fechada, caso o seu ponto inicial seja igual ao seu ponto final. Uma curva simples e fechada é denominada por anel.

LineString, que é uma curva com interpolação linear entre pontos. Existem vários tipos de *LineString*, nomeadamente simples, não-simples, simples e fechado, e não-simples e fechado. Por não-simples entende-se *LineStrings* que passem pelo mesmo ponto mais do que uma vez e por fechadas *LineStrings* com o ponto inicial igual ao final. A figura 16 exemplifica cada um destes tipos.

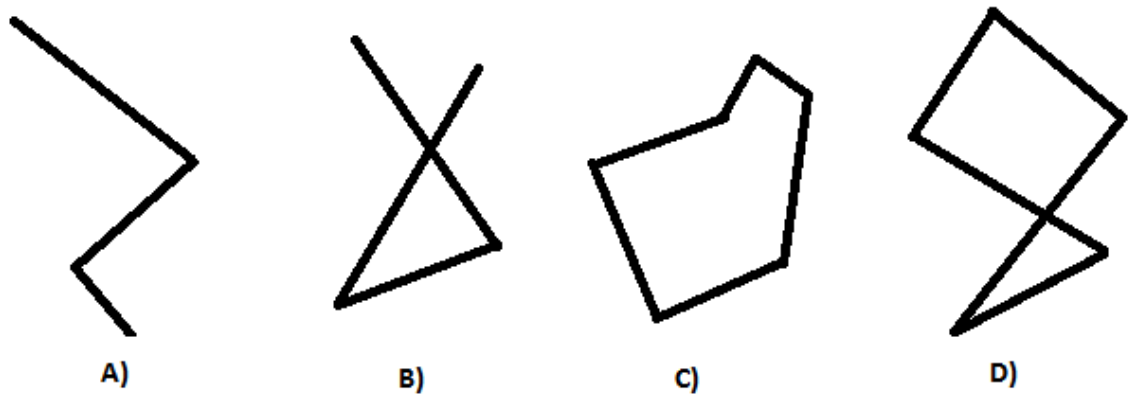


Figura 16 - A) LineString simples; **B)** LineString não-simples; **C)** LineString simples e fechada; **D)** LineString não-simples e fechada

Surface, que é um objeto geométrico bidimensional e tal como o próprio nome indica, corresponde a uma determinada superfície.

Polygon, que é um objeto geométrico que corresponde a uma superfície planar que possui determinadas regras: são fechados, não pode ter linhas cortadas, pontas ou cavidades e o seu interior é um conjunto de pontos conectados.

Quanto aos subtipos ***MultiPoint, MultiCurve, MultiSurface, MultiPolygon, MultiLineString*** estes possuem características muito semelhantes. Cada um deles é um objeto

geométrico composto por vários elementos do seu tipo, isto é, um *MultiPoint* é composto por vários pontos, um *MultiCurve* por várias curvas, um *MultiSurface* por várias superfícies, um *MultiPolygon* por vários polígonos, um *MultiLineString* por vários *LineStrings*.

3.7 *Spatial ETL*

O processo de ETL que permite a extração de dados de fontes externas, a transformação desses dados extraídos, para atender às necessidades de negócios, e o carregamento dos dados para um *data warehouse*. Este processo é considerado uma das fases mais críticas do *data warehouse*, uma vez que como alguns estudos comprovam, podem consumir 80% do tempo de desenvolvimento do projeto (Demarest, M., 1997).

O *Spatial ETL* consiste num conjunto de ferramentas de *software* que, para além das funcionalidades presentes nas ferramentas de ETL convencionais, possui a vantagem de manipular dados espaciais. De seguida serão analisadas as suas diversas fases, com o objetivo de entender as principais diferenças entre um ETL convencional e um *Spatial ETL*. Vejamos então cada uma delas:

1. **Extração (*Extract*) -** esta fase é bastante semelhante com a executada pelos sistemas de ETL convencionais. Contudo, possui alguns problemas adicionais, uma vez que os dados espaciais, mais do que qualquer outro tipo de dados, podem estar em vários formatos e padrões. Assim, é necessário um cuidado maior ao efetuar a extração dos dados, sendo que essa tarefa é facilitada com o recurso a *software* especializado, como é o caso do *Spatial ETL*.
2. **Transformação/Limpeza (*Transform*):** para além das funcionalidades presentes nas ferramentas ETL convencionais, as ferramentas *Spatial ETL* possuem várias funções específicas para dados espaciais, tais como:
    a) Projeção - a capacidade de converter dados espaciais entre vários sistemas de coordenadas

b) Transformações espaciais - a capacidade de modelar interações espaciais e calcular predicados espaciais.

c) Transformações topológicas - a capacidade de criar relações topológicas entre os conjuntos de dados diferentes.

d) Ressimbolização - a capacidade de alterar as características cartográficas de um recurso, tais como a cor ou linha.

e) Geocodificação - a capacidade de converter dados convencionais em dados espaciais.

3. **Carregamento (*Load*):** esta fase não sofre alterações em relação às ferramentas de ETL convencionais.

Capítulo 4

O caso de Estudo

4.1 Níveis e áreas do Data Warehouse Espacial

Para o caso de estudo desta dissertação foi escolhido o *data mart* (*InternetSales*) do *data warehouse AdventureWorks2012*, relativo às vendas de produtos realizadas através da Internet e que é amplamente usado no âmbito académico. Optou-se, assim, por usar um *data warehouse* já construído e povoado, tornando, possível desta forma focar todo o trabalho unicamente nas alterações necessárias para transformar um *data warehouse* convencional num data warehouse espacial. Apesar de se tratar de uma transformação de um *data warehouse* e não da construção de um *data warehouse* espacial de raiz, a maioria das fases de construção de um *data warehouse* espacial referidas no capítulo anterior foram utilizadas neste caso de estudo.

Para Mennecke e Higgins (1999) existem vários níveis possíveis de inclusão dos dados espaciais num *data warehouse*. Os investigadores, consideram que o primeiro nível de um *data warehouse* espacial corresponde a um *data warehouse* com dimensões espaciais, isto é, com atributos de localização, passíveis de serem georreferenciados, porém não possuindo suporte para o seu tratamento.

Ainda, para os investigadores, a grande maioria dos *data warehouses* atuais enquadram-se neste nível, que não oferecem qualquer capacidade de análise espacial, devido simplesmente ao facto de que as informações espaciais estarem representadas como texto.

Num segundo nível, os investigadores identificam os *data warehouses* com capacidade de visualizar e analisar informação espacial no próprio DW, com recurso a ferramentas GIS, separadas do ambiente de *data warehousing*. Este tipo de *data warehouse* tem como principal desvantagem, a necessidade de exportar os dados obtidos por consultas sobre um *data warehouse* para outro ficheiro e, em seguida, utilizando uma ferramenta GIS, incorporar esses dados para efetuar as análises espaciais. Este esquema de implementação do DW, além de excessivamente trabalhoso, é, apenas, utilizado em áreas de uma organização onde o espaço é tido como crítico (Gonzales, 1999).

O terceiro e último nível dos *data warehouses* espaciais, de acordo com Mennecke e Higgins (1999) corresponde, efetivamente, ao *data warehouse* espacial, tal como foi analisado no capítulo anterior. De acordo com os investigadores, os *data warehouses* espaciais devem permitir a consulta, a análise, a criação de modelos e, a visualização, tanto de dados espaciais como de dados convencionais. Desta forma, os dados espaciais tornam-se uma componente integral do DW espacial.

De acordo com o Yuan et al. (2000), os *data warehouses* espaciais, têm impacto em três áreas principais: informação geográfica na descoberta de conhecimento, descoberta de conhecimento geográfico na ciência da informação geográfica e descoberta de conhecimento geográfico na ciência da informação geográfica. No âmbito dos trabalhos desta dissertação, apenas focaremos a área de informação de geográfica na descoberta de conhecimento, uma vez que o principal objetivo de um *data warehouse* espacial, num ambiente de negócios é a visualização e descoberta de padrões sobre os dados e visualização da informação numa camada distinta da camada dos dados convencionais, não tanto, como acontece em outras situações, o tratamento e manipulação de mapas ou de outros aspetos relacionados com a geografia (Gonzales, 1999).

4.2 Data Mart "InternetSales"

A seleção do *data mart* "InternetSales*"* do *data warehouse AdventureWorks2012*, teve em conta vários fatores que consideramos serem pertinentes neste tipo de processo. Um desses fatores foi a necessidade de se obter um *data warehouse* com características específicas, como por exemplo, um *data warehouse* de uma organização que pudesse beneficiar de uma análise espacial dos seus dados. Outro fator foi a necessidade do desenho do *data warehouse* estar disponível para utilização em termos académicos, sendo essa uma das principais vantagens em específico. Após a escolha do *data warehouse*, passamos à seleção do *data mart.* Esta recaiu sobre o *data mart "*InternetSales*"*, devido às características das suas dimensões, isto é, são dimensões com características espaciais, representadas unicamente em texto. Utilizando este *data mart* seria, assim, possível, por exemplo, efetuar uma análise por cidade, dos produtos vendidos de uma determinada categoria e numa determinada época. Vamos então explicar de forma um pouco mais detalhada as dimensões e a representação do *data mart* em questão.

Este *data mart* tem como objetivo fornecer suporte adequado para analisar as vendas realizadas através da Internet. Para além da tabela de factos óbvia, o *data mart* é constituído por várias dimensões de suporte: uma dimensão temporal *DimDate*, que acolhe o calendário relacionado com as vendas, uma dimensão *DimCurrency*, que representa a moeda em que foi efetuada a compra, uma dimensão *DimPromotion*, que indica se existe alguma promoção associada ao produto vendido, uma dimensão *DimProduct*, que representa o produto vendido, que por sua vez tem associado uma subcategoria e uma categoria, representado no *data mart* pelas dimensões *DimProductSubCategory* e *DimProductCategory*, respetivamente. Possui ainda uma dimensão *DimCustomer*, que representa a informação relativa ao cliente, uma dimensão *DimSalesTerritory*, que representa as áreas de venda dos produtos e, por fim, uma dimensão *DimGreography*, que representa toda a informação geográfica relativa aos clientes (Figura 17).

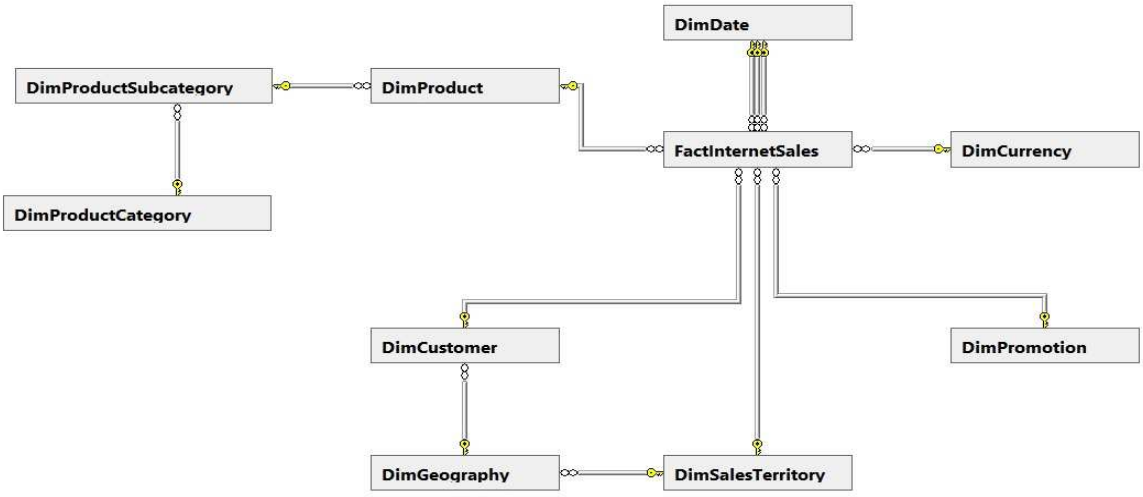


Figura 17 – Representação simplificada do *data mart* "InternetSales".

4.3 O Levantamento de Requisitos

O processo de levantamento de requisitos é bastante importante na construção de um *data warehouse* espacial, uma vez que se destina a recolher as necessidades dos utilizadores e a informação nos sistemas operacionais necessárias ao seu suporte. Para além disso, através do levantamento de requisitos é possível detetar a necessidade de obtenção de dados espaciais específicos para o *data warehouse*, bem como estabelecer a sua qualidade. Segundo Indulska e Orlowska (2002) nem todas as aplicações requerem a mesma qualidade de dados para o seu processamento. A maioria das organizações não necessita de informações precisas e de mapas detalhados para as suas análises. O simples facto de incluir as coordenadas na informação do cliente representa, por si só, um incremento substancial na capacidade de análise espacial. A informação georreferenciada para a maioria das organizações pode ser obtida a partir dos códigos postais, podendo ser unicamente ser representada no mapa da cidade, em vez de se representarem as ruas. Esta análise permite reduzir o custo de aquisição de dados espaciais, uma vez que esses dados estão normalmente disponíveis de forma gratuita, sem que isso represente qualquer tipo de perda na qualidade dos dados (Indulska e Orlowska, 2002).

No processo de levantamento de requisitos para o caso do nosso *data mart*, foi necessário analisar todas as dimensões do *data mart*, por forma a identificar os atributos com características espaciais. Após terminada esta análise, concluiu-se que existia informação geográfica presente nas próprias dimensões. Contudo, seria necessário efetuar algumas alterações ao seu esquema, que implicariam que as novas dimensões necessitariam de dados espaciais. Isto requereu uma nova análise das fontes de dados para se descobrir a necessidade de recorrer a fontes externas para povoar as novas dimensões. Nessa análise concluiu-se que não se possuía os dados necessários, tendo-se levantado a necessidade de recorrer a fontes de dados externas.

4.4 Modelação dimensional

Com a utilização de um data warehouse, já construído e povoado, o processo de modelação dimensional não seguiu os trâmites usuais. Assim neste processo, foi necessário identificar as dimensões com características espaciais, alterá-las de acordo com os seus atributos e, por fim, desenhar um esquema final do *data warehouse* espacial.

4.5 Alguns aspetos a ter em conta

Após a análise efetuada no capítulo de implementação de um *data warehouse* espacial e de ter sido demonstrado um dos possíveis métodos para o desenho multidimensional do *data warehouse* espacial, é preciso relevar alguns outros aspetos que podem influenciar o próprio desenho do *data warehouse*. Alguns desse aspetos a ter conta no desenho do *data warehouse* espacial relacionam-se com a redundância dos dados e a performance do *data warehouse* aquando da realização de queries SOLAP (Thiago et al., 2009).

Han et al Han et al (1998) foram os primeiros a propor uma *framework* para suporte a dimensões e medidas espaciais. Contudo, o modelo que sugeriram apresentava problemas de redundância de informação, visto que, todos os níveis do atributo espacial deveriam possuir características geométricas representando o objeto espacial. Mais tarde, surgiu um outro método, proposto por Fidalgo et al. (2010). Este método teve em especial atenção o aspeto da redundância dos dados. Nele foram propostas as dimensões geográficas primitivas, que correspondiam tipicamente a *outriggers*. Dessa forma, a informação geográfica poderia ser partilhada por várias dimensões, evitando fazer a replicação dessa informação. Para além desta vantagem, a informação geográfica é uma informação pesada, isto é, a consulta sobre os seus dados é mais complexa do que uma consulta sobre dados convencionais e, consequentemente, mais demorada. A criação de *outrigers* permite que esta informação seja utilizada quando necessária, permitindo assim maior rapidez nas consultas às dimensões não espaciais. Contudo, a utilização de *outriggers* deve ser bem ponderada, uma vez que,

para consultar a informação presente nessas dimensões é necessário efetuar um maior número de junções entre as dimensões, conduzindo a uma perda de performance (Margy Ross, 2008).

4.6 Identificação das dimensões espaciais

Apesar de terem sido analisadas todas as dimensões existentes no *data mart*, no nosso caso serão apenas indicadas em seguida as dimensões com características espaciais. Para identificar os atributos com características espaciais será utilizada a notação MADS (Malinowski e Zimányi, 2008).

4.6.1 A Dimensão "DimSalesTerritory"

A dimensão "DimSalesTerritory" representa a informação relativa ao território das vendas efetuadas, tendo como objetivo permitir identificar a região onde foi efetuada a compra.

Esta dimensão é constituída pelos seguintes atributos:

- "SalesTerritoryKey", que representa a chave da dimensão.
- "SalesTerritoryAlternateKey", que representa a chave alternada da dimensão.
- "SalesTerritoryRegion", que representa a região do país.
- "SalesTerritoryCountry", que representa o país em que foi efetuada a venda.
- "SalesTerritoryGroup", que representa o continente a que pertence o país.

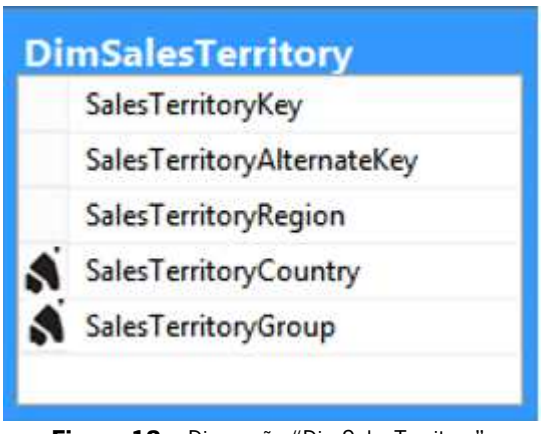


Figura 18 – Dimensão "DimSalesTerritory".

Tanto o atributo "SalesTerritoryCountry" como o atributo "SalesTerritoryGroup" possuem características geográficas e, como tal, a capacidade de serem representados num mapa. Tal como se pode ver na figura 18, ambos os atributos são representados por um conjunto de superfícies ("Surface Bag").

4.6.2 A Dimensão "DimGeography"

A dimensão "DimGeography" representa a informação relativa à cidade e ao país, permitindo dessa forma identificar a cidade e o país onde foi efetuada a compra.

Esta dimensão é constituída pelos seguintes atributos:

- "GeographyKey", que representa a chave da dimensão.
- "City", que representa o nome da cidade.
- "CountryRegionCode", que representa o código do país.
- "CountryRegionName", que representa o nome do país.
- "PostalCode", que representa o código postal da cidade.
- "SalesTerritoryKey", que é a chave estrangeira para a dimensão "DimSalesTerritory".

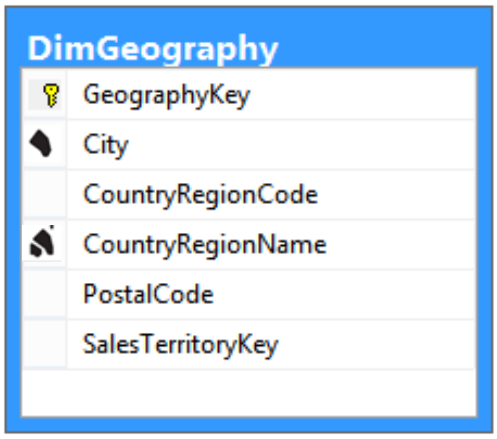


Figura 19 - Dimensão "DimGeography".

Nesta dimensão existem dois atributos espaciais: o atributo cidade, que é representado por uma superfície ("Surface"); e o atributo "CountryRegionName", que é representado por um conjunto de superfícies ("Surface Bag").

4.6.3 A Dimensão "DimCustomer"

A dimensão "DimCustomer" representa toda a informação relativa ao cliente, tendo como principal objetivo identificar o cliente que efetuou uma determinada compra.

Esta dimensão é constituída por um vasto número de atributos, portanto, apenas será indicado o atributo com características espaciais:

- "AddressLine", que representa a rua da residência do cliente.

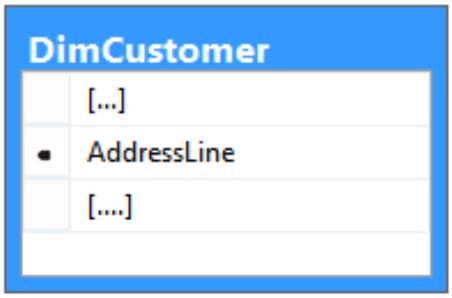


Figura 20 – "Dimensão DimCustomer".

Apenas um atributo possui características espaciais, o atributo que contém a morada do cliente e que é representado por um ponto ("Point").

4.7 Transformação das Dimensões

Após analisadas as dimensões e identificados os atributos espaciais e seus tipos, é necessário efetuar algumas alterações às dimensões, de forma a que possam armazenar informação espacial. De seguida, serão demonstradas as várias alterações efetuadas às dimensões.

4.7.1 A Dimensão "DimCustomer"

Esta dimensão possui apenas um atributo espacial, "AddressLine", que pode ser analisado de forma convencional ou de forma espacial. Este atributo dará origem a uma dimensão geográfica primitiva com os dados geográficos respetivos, a dimensão "DimAddressGeo". No entanto, é necessário adicionar um atributo à dimensão "DimCustomer", que será uma chave estrangeira para a nova dimensão, tornando-a numa micro dimensão híbrida, como é demonstrado de seguida.

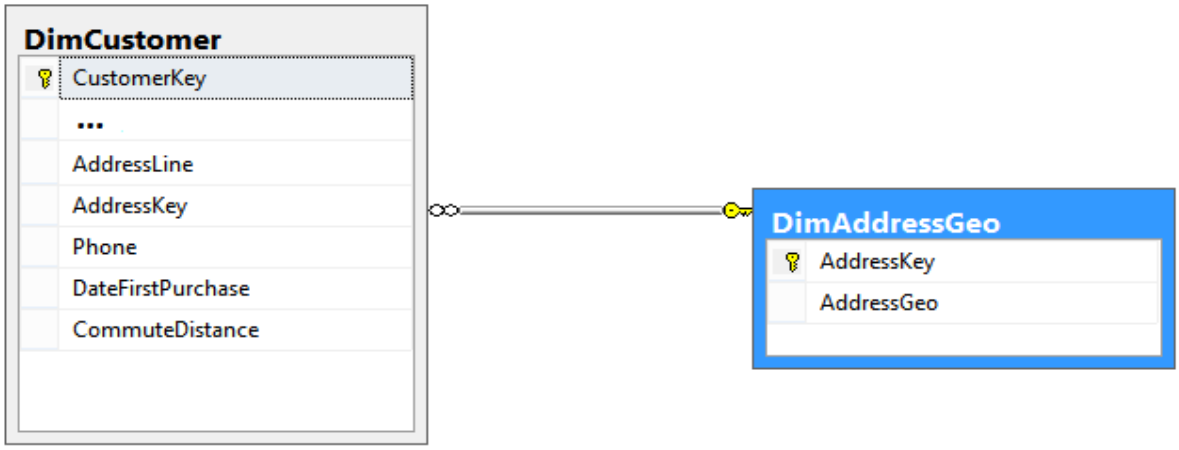


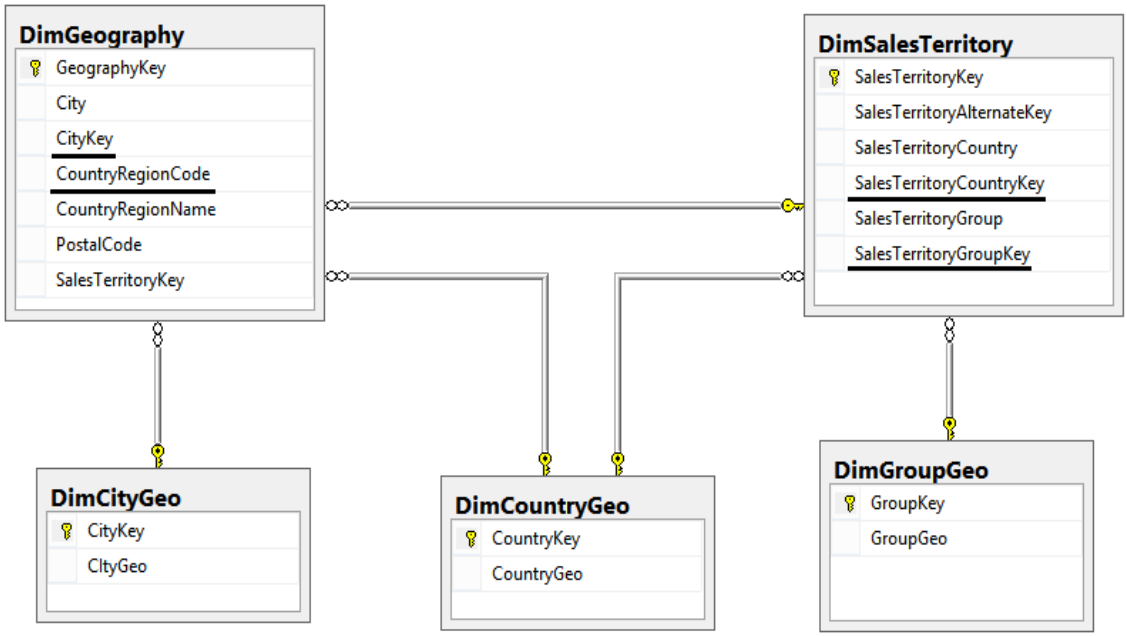
Figura 21 – Dimensão "DimCustomer" transformada.

4.7.2 A Dimensão "DimGeography" e a Dimensão "DimSalesTerritory"

O motivo pelo qual as dimensões "DimGeography" e "DimSalesTerritory" serão apresentadas em conjunto, prende-se com o facto de ambas partilharem uma nova dimensão, contendo a informação relativa ao país. A dimensão "DimGeography" possui dois atributos espaciais, o atributo "CountryRegionName" e o atributo "City". O atributo "City" dará origem a uma nova dimensão com os

dados espaciais, "DimCityGeo", enquanto o atributo "CountryRegionName" dará origem à dimensão "DimCountryGeo". Assim, será necessário adicionar à dimensão "DimGeography" uma chave estrangeira para a dimensão "DimCityGeo". Porém, não será necessário adicionar um novo atributo visto que o atributo "CountryRegionCode" poderá ser reutilizado como chave estrangeira para a dimensão "DimCountryGeo". Mas, tal como referido anteriormente, a dimensão "DimCountryGeo" será partilhada também pela dimensão "DimSalesTerritory". Esta última dimensão é constituída pelo atributo "SalesTerritoryCountry", que estará ligado à dimensão "DimCountryGeo", e, ainda, pelo atributo "SalesTerritoryGroup", que dará origem a uma nova dimensão "DimGroupGeo" com a informação espacial. Além desta nova dimensão, é necessário adicionar dois atributos à dimensão "DimSalesTerritory", que serão as chaves estrangeiras para as respetivas dimensões. Estas alterações levam a que as dimensões analisadas se transformem em macro dimensões híbridas. De seguida, na figura 22, apresenta-se o esquema resultante de todas as alterações realizadas.

Figura 22 – Dimensão "DimGeography" e dimensão "DimSalesTerritory" transformadas.



4.8 Desenho Final do *Data Mart*

Após todas as análises e alterações terem sido concluídas, desenvolvemos o desenho final *data mart* (figura 23). De forma a representar as alterações efetuadas, nessa figura, estão destacadas as novas dimensões espaciais, a vermelho, e as dimensões onde foram adicionados novos atributos, a azul.

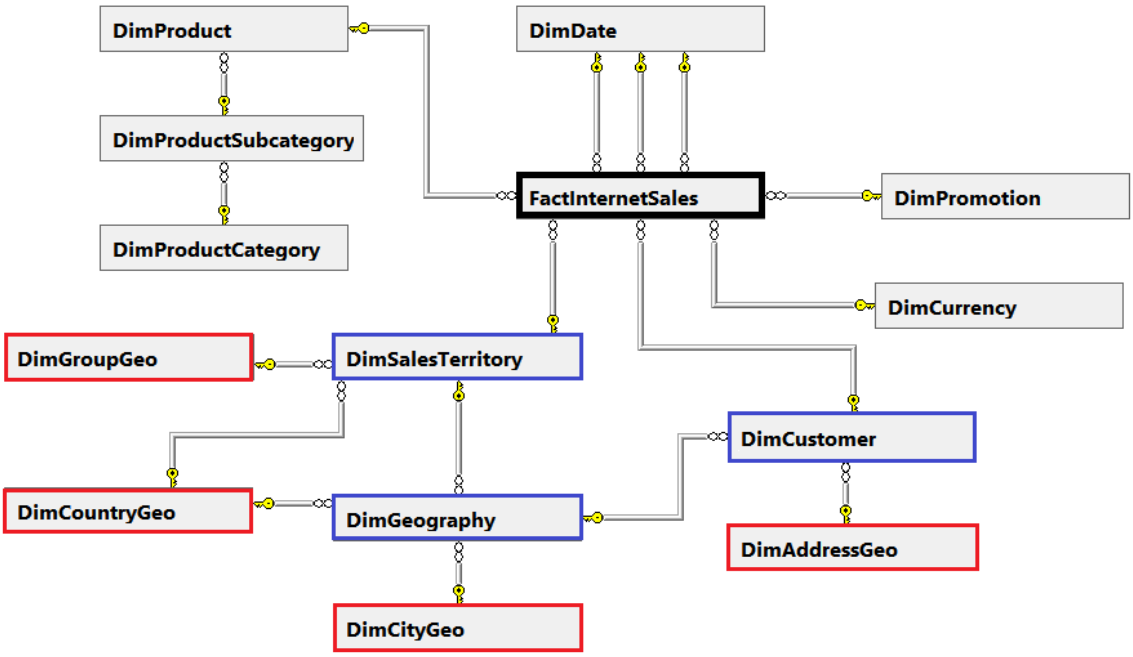


Figura 23 – *Data mart FactinternetSales* final.

Como podemos observar na Figura 23, o esquema resultante das alterações efetuadas consiste num esquema em *snowflake*¸ onde foram adicionadas quatro dimensões de segunda linha, "DimGroupGeo", "DimCountryGeo", "DimCityGeo" e "DimAddressGeo", com o objetivo de fornecer suporte às análises espaciais do *data mart*. As dimensões "DimSalesTerritory", "DimGeography" e "DimCustomer" sofreram também alterações ou foram adicionados novos atributos que consistem em chaves estrangeiras para as novas dimensões.

4.9 Povoamento do *Data Warehouse*

O povoamento do *data warehouse* é considerado o processo mais crítico de todo o processo de construção de um *data warehouse*, onde a sua complexidade está bastante dependente da heterogeneidade das bases de dados. Nesta fase, são processados os mapeamentos sintáticos e semânticos entre os esquemas, criando assim uma visão unificada e concretizada das fontes (João Ferreira et al., 2010). A qualidade dos dados, presentes no ciclo de desenvolvimento, tem um papel fulcral no fornecimento dos dados necessários para suportar os processos de análise por parte dos agentes de decisão. A utilização de ferramentas ETL ajuda na construção deste processo, contudo, continua a ser algo complexo. De seguida, será descrito todo a trabalho realizado nesta área.

4.9.1 Uma Fonte de Dados Externa

Após analisar as fontes de dados à nossa disposição, verificou-se que não existia a informação necessária para corresponder às necessidades de análise dos agentes de decisão e, portanto, seria necessário recorrer a uma fonte de dados externa.

Depois de efetuar várias pesquisas sobre esta área, verificou-se a existência de um vasto leque de opções. Mas, a maioria dessas opções designavam serviços pagos e bastante complexos. Assim, optou-se por construir uma base de dados com a informação espacial necessária para o povoamento das dimensões. Esta base de dados é constituída por quatro tabelas:

- "Street" – a tabela que contém a informação espacial relativa às moradas dos clientes.
- "Cities" – a tabela que contém a informação espacial relativa às cidades.
- "Countries" – a tabela que contém a informação espacial relativa aos países.
- "Continent" – a tabela que contém a informação espacial relativa aos continentes.

O Povoamento das Tabelas "Continent", "Countries" e "Cities"

Para o povoamento das tabelas "Continent". "Countries" e "Cities" foi necessário obter dados relativos à geometria referente a cada um dos casos. Para tal, foi utilizada uma aplicação com o nome *shape2sql* (http://www.sharpgis.net/page/shape2sql.aspx). Esta aplicação (Figura 24) permite inserir, numa tabela do *SQL Server*, dados relativos a um *shapefile*. *Shapefiles* são ficheiros que contêm informação geoespacial sobre a forma de vetores. A informação geoespacial contém a geometria do objeto (ponto, linha, polígonos) e atributos que o descrevem.

A figura seguinte corresponde ao interface da aplicação.

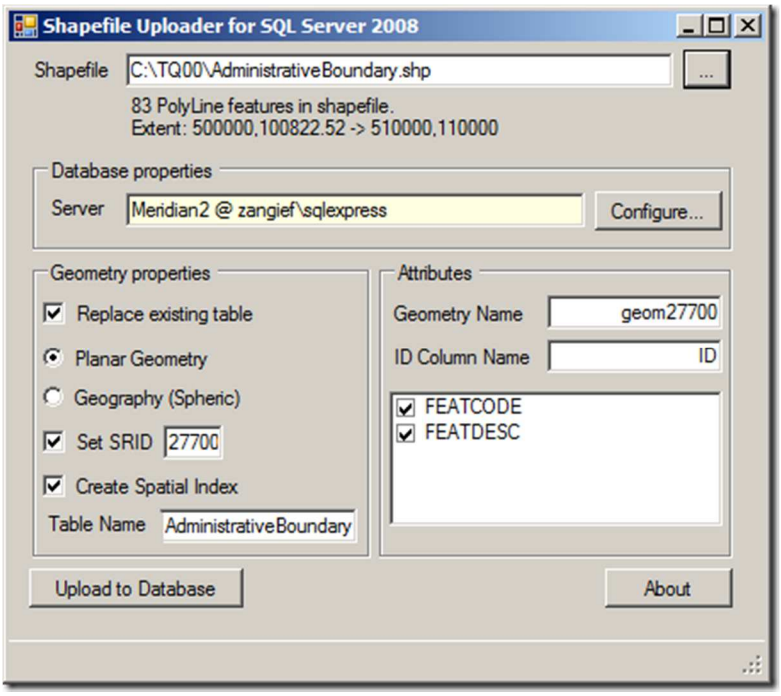


Figura 24 – Interface da aplicação shape2sql.

As principais opções da aplicação *shape2sql* podem-se resumir em:

- *Replace existing table* – caso esta opção seja selecionada e a tabela em questão exista, esta é reescrita, caso contrário, a informação é adicionada à tabela existente.
- *Planar Geometry vs Geography* – esta indica se os dados geográficos são planares ou esféricos.
- *Set SRID* – define o identificador de referência espacial.

- *Create Spatial Index* – adiciona um índice espacial à tabela.
- *Table Name* – define o nome da tabela na base de dados.
- *Geometry Name* – define o nome da coluna geométrica na base de dados.
- *ID Column Name* – define o nome da chave primária da tabela.
- *Attribute Columns* – seleciona as colunas a adicionar à tabela.

O resultado da execução desta aplicação consiste, numa tabela contendo a informação geográfica ou geométrica e todos os atributos selecionados na opção de "Attribute Columns".

Povoamento da tabela street

Para o povoamento desta tabela, optou-se por uma abordagem um pouco diferente, uma vez que esta tabela possui a informação sobre as ruas dos clientes, que estão representadas por pontos. A abordagem consistiu, inicialmente, em obter todas as ruas de todos os clientes, em conjunto com a cidade respetiva e o país do *data warehouse* original. Para isso utilizámos a seguinte querie:

```
SELECT DISTINCT c.AddressLine, d.City, d.CountryRegionName
FROM AdventureWorksDW2012Source.dbo.DimCustomer AS c,
AdventureWorksDW2012Source.dbo.DimGeography AS d
WHERE c.GeographyKey=d.GeographyKey;
```

Após a obtenção dos resultados, estes foram colocados num ficheiro de texto para serem tratados mais tarde. Após a sua extração, recorreu-se a um *website* (http://www.findlatitudeandlongitude.com/batch-geocode/) para fazer a geocodificação das ruas em coordenadas GPS. De seguida, apresenta-se um pequeno exemplo de uma geocodificação resultante do *website* referido:

Input : 1343 Apple Drive, Lemon Grove, United States

OutPut : 32.714813,-116.990822

Após este processo estar concluído, foi necessário inserir toda a informação presente no ficheiro de texto, na tabela street.

4.9.2 Processo de ETL

Após a obtenção da informação geográfica espacial necessária, desenvolveu-se um sistema ETL para o povoamento do *data warehouse* espacial. Este sistema tem como fonte de dados a base de dados, criada previamente com os dados espaciais, e o *data warehouse* convencional povoado, poupando, assim, o esforço da criação de um ETL específico para o tratamento desses dados. O ETL construído faz o povoamento das dimensões espaciais adicionadas e a atualização das dimensões existentes com as chaves estrangeiras para as respetivas dimensões espaciais. Para além destas bases de dados foi utilizada uma área de retenção - *data staging área* -, para o tratamento da informação das fontes de dados e posterior carregamento para o *data warehouse*. Para a realização desta tarefa, foi utilizado o *Microsoft Visual Studio*, no qual foi criado um projeto *Integration Services*. Todo o processo de povoamento e de limpeza, tanto do data warehouse como da DSA, foi criado recorrendo, unicamente, a SQL. De seguida, serão explicados os principais passos do processo de ETL.

Limpeza da *Data Staging Area* e do *Data Warehouse* Espacial

As primeiras tarefas a executar no ETL estão relacionadas com a limpeza da *data stating area* e do *data warehouse* final. Este passo é necessário por forma a testar, repetidamente, o código presente nas várias fases do processo. A *data stating area* é apenas um local temporário para armazenar a informação e que deve ser limpa no início e no fim do processo ETL. Por sua vez, a limpeza do *data warehouse* é necessária, devido ao facto da informação necessária ser extraída do *data warehouse* original e ser colocada posteriormente no novo *data warehouse* desenvolvido.

Figura 25 – SQL Tasks de limpeza da DSA e do DWE.

Povoamento do Data Warehouse com a Informação Original

Tal como referido anteriormente, optou-se por utilizar a informação presente no data warehouse original, de forma a focar toda a atenção na parte geográfica. Desta forma, evitamos extrair, tratar e carregar novamente os dados existentes. Esta tarefa consiste, então, em selecionar toda a informação das dimensões do *data warehouse* original e inserir no *data warehouse* espacial.

Figura 26 – Representação da SQL *Task* de povoamento dos dados originais.

O Povoamento das Dimensões Espaciais

Após concluído o processo de povoamento do *data warehouse* espacial, com os dados originais, tem início o processo de povoamento das dimensões espaciais. Este povoamento é feito em paralelo, por forma a otimizar o processo de ETL. O fluxo de cada uma das dimensões é bastante semelhante ao das restantes.

O primeiro passo consiste em extrair a informação geográfica, presente na base de dados, para tabelas de auditoria. Após estra extração ter sido realizada, esta é inserida em tabelas de equivalência, com a respetiva chave de substituição já associada e uma descrição do objeto em causa. Depois são inseridos os registos devidamente tratados nas dimensões geográficas. De seguida, é necessário atualizar os registos das dimensões com atributos que sejam chaves estrangeiras das novas dimensões. Para tal, é necessário efetuar uma consulta a uma tabela de equivalência, de forma a identificar o atributo correspondente ao registo em causa, sendo, de seguida, feita a atualização dessa informação.

Tal como já referido, o fluxo é bastante semelhante no povoamento de todas as dimensões. Assim, de seguida, apenas se explicará, detalhadamente, o povoamento de uma das dimensões.

O processo de povoamento da dimensão “DimCountryGeo” inicia-se com a extração dos dados relativos ao país, para a tabela de auditoria “audDimCountry”. Após essa extração ter sido realizada, a informação do país e a respetiva chave de substituição é inserida na tabela “equiDimCountry”, sendo, de seguida, inseridos os registos na dimensão. Depois, é necessário atualizar os registos das

dimensões "DimGeography" e "DimSalesTerritory", mais concretamente, os atributos *CountryRegionCode* e *SalesTerritoryCountryKey* respetivamente, correspondendo à chave estrangeira para a tabela de dimensão "DimCountryGeo" - a chave estrangeira é obtida através de uma consulta à tabela de equivalência "equiDimCountry". De seguida, é realizado um *update* aos registos de ambas as dimensões, com as chaves estrangeiras correspondentes ao atributo país do registo em causa.

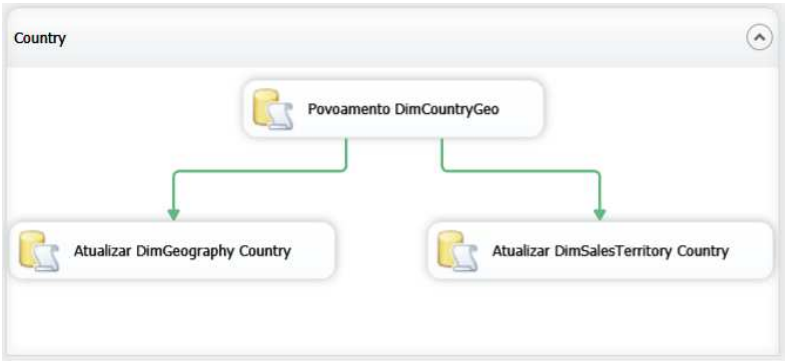


Figura 27 – Representação das SQL *Tasks* de povoamento e atualização das dimensões relacionadas com o país.

Limpeza da *Data Staging Area*

Por fim, após todo o processo de povoamento das dimensões, é necessário efetuar a limpeza das tabelas da DSA, visto que estas tabelas são tabelas meramente auxiliares, utilizadas durante o processo de povoamento das dimensões.

Figura 28 – SQL *Task* de limpeza da DSA final.

4.10 Resultados Obtidos

Após estar concluído o processo de extração, limpeza e carregamento para o *data warehouse*, foi necessário validar todos os resultados desse processo. Essa validação foi efetuada através de um conjunto de consultas especificamente orientadas às várias dimensões, de forma a verificar se o processo de ETL foi concluído com sucesso e se os resultados obtidos eram os pretendidos. Nas próximas secções serão demonstradas os resultados das consultas realizadas sobre as dimensões espaciais em que incidiu todo o foco do processo de ETL. Para visualização dos dados geográficos foi utilizada uma das potencialidades do *SQL Server Management Studio*, a ferramenta *Spatial Tools*.

4.10.1 A Dimensão "DimAddressGeo"

Para a verificar os resultados obtidos do processo ETL, em relação a esta dimensão, efetuou-se a seguinte *querie*, de forma a validar, principalmente, o conteúdo do atributo geográfico.

```
SELECT * FROM AdventureWorksSDW2012.dbo.DimAddressGeo;
```

Como resultado desta *querie* foi obtida a seguinte figura, representados na tab *Spatial Results*, presente no SQL *Server Management Studio*.

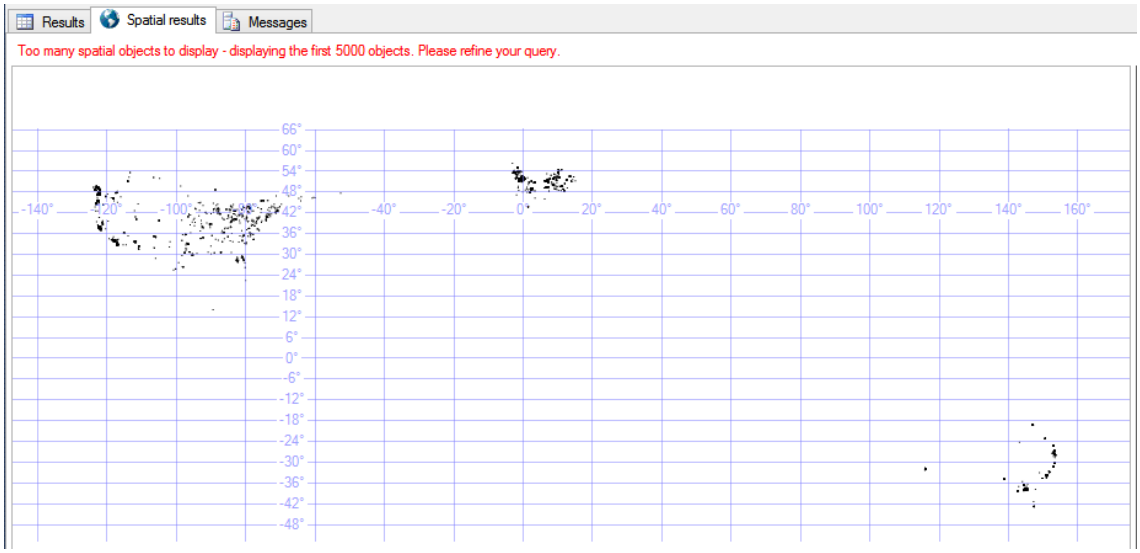


Figura 29 – Mapa resultante da consulta efetuada à dimensão "DimAddressGeo".

Tal como a figura 29 demonstra, este resultado não apresenta todos os registos existentes na dimensão. Contudo, facilmente se pode comprovar que, os registos presentes na dimensão "DimAddressGeo", possuem a capacidade de serem representados num sistema georreferenciado.

De seguida apresenta-se uma breve análise dos vários componentes que constituem o valor geométrico do atributo espacial "AddressGeo" (Alastair Aitchison, 2012).

Valor: 0x00000000010C8E01D9EBDD4D50401630815B77774240

- **0x:** identificador na notação hexadecimal.
- **00:** indica a ordem dos bytes. 0x00 corresponde a uma ordem *little-endian*.
- **00000001:** indica o tipo de geometria do registo, sendo que o valor 1 corresponde a um ponto.
- **0C8E01D9EBDD4D504:** valor da coordenada X (65,21667).
- **01630815B77774240:** valor da coordenada Y (36,93333).

4.10.2 A Dimensão "DimCityGeo"

Para validar os registos da dimensão "DimCityGeo" efetuou-se a seguinte *querie*:

```
SELECT * FROM AdventureWorksSDW2012.dbo.DimCityGeo;
```

A partir dos resultados obtidos para esta *querie*, obteve-se os vários pontos correspondentes às várias cidades referidas no *data warehouse* (Figura 30).

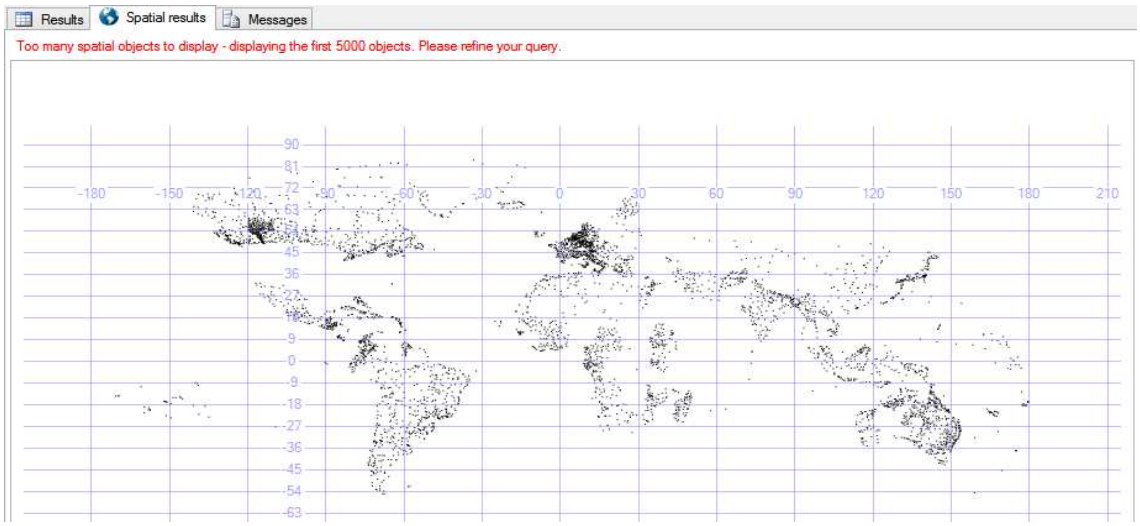


Figura 30 – Mapa resultante da consulta efetuada à dimensão "DimCityGeo".

Tal como na dimensão anterior, podemos verificar que não é possível visualizar todos os resultados disponibilizados pela *querie* apresentada, devido à limitação da ferramenta utilizada apenas permite mostrar 5000 registos.

4.10.3 A Dimensão "DimCountryGeo"

Para validar os registos da dimensão "DimCountryGeo" desenvolvemos a seguinte *querie*:

```
SELECT * FROM AdventureWorksSDW2012.dbo.DimCountryGeo;
```

Como resultado, obtivemos uma representação dos vários países referidos no *data warehouse* (Figura 31).

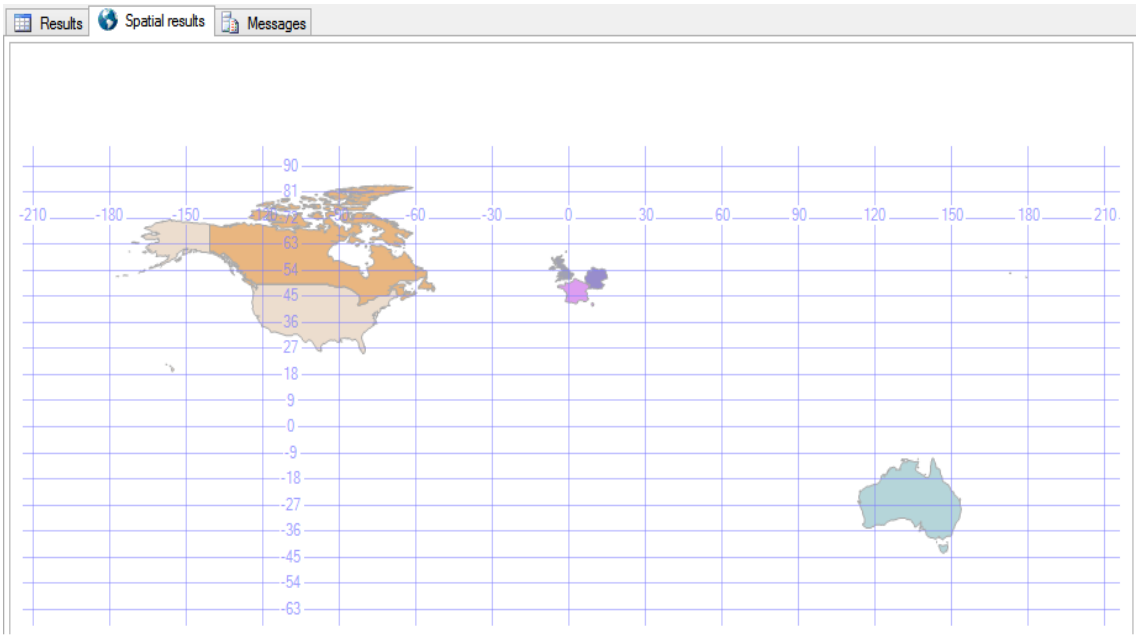


Figura 31 – Mapa resultante da consulta efetuada à dimensão "DimCountryGeo".

4.10.4 A Dimensão "DimGroupGeo"

Para validar os registos da dimensão "DimGroupGeo" efetuou-se a seguinte *querie*:

```
SELECT * FROM AdventureWorksSDW2012.dbo.DimGroupGeo;
```

Como resultado, obteve-se uma representação dos vários continentes referidos no *data warehouse* (Figura 32).

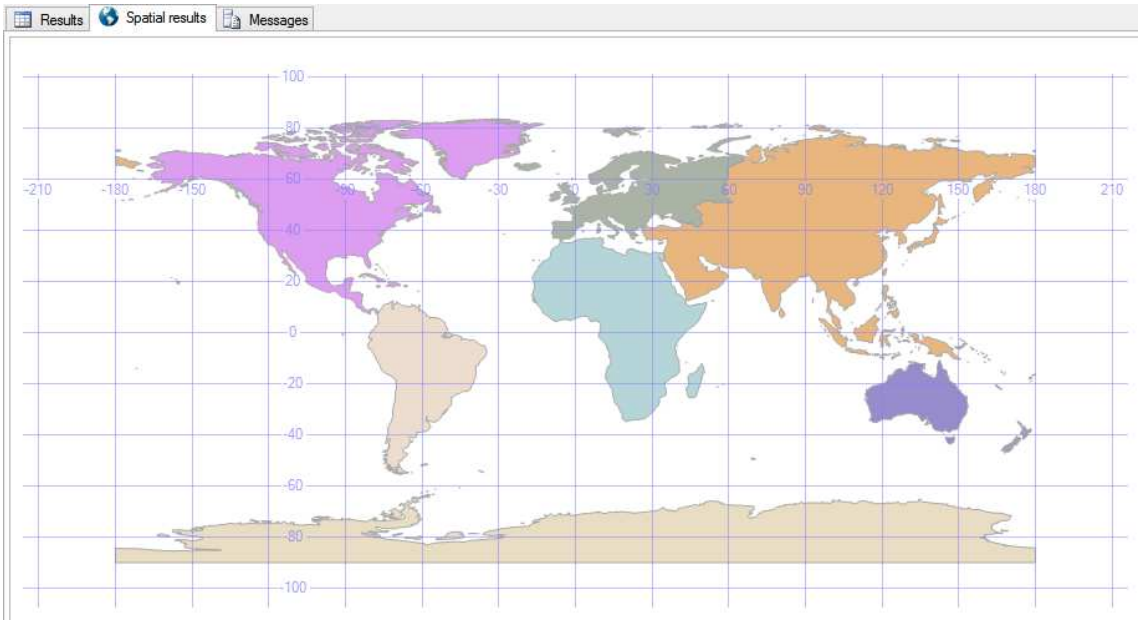


Figura 32 – Mapa resultante da consulta efetuada à dimensão "DimAddressGeo".

Capítulo 5

Conclusão e Trabalho Futuro

5.1 Conclusão

O desenho e implementação de um *data warehouse* espacial é um processo complexo e ao mesmo tempo subjetivo. Apesar da existência de várias propostas na literatura, não existe um consenso sobre a forma como se deve implementar este tipo de *data warehouse*, o que, obviamente, dificulta um pouco mais todo o seu processo. Ao longo desta dissertação, foram apresentados os vários métodos propostos pelos investigadores, bem como todo o processo de transformação de um *data warehouse* convencional num *data warehouse* espacial.

Os *data warehouses* constituem parte integrante do processo de tomada de decisão por parte de várias organizações. Contudo, apesar dos *data warehouses*, possuírem uma grande quantidade de dados espaciais, estes continuam sem ser devidamente explorados e analisados devido às limitações dos *data warehouses* convencionais. Era necessário analisar a complexidade dos dados espaciais e os vários métodos propostos na literatura para a integração destes dados no *data warehouse*, tornando-os parte integrante das análises dos agentes de decisão e tirando o máximo conhecimento deles.

Após ter sido realizado um levantamento de várias propostas para o desenho de um *data warehouse* espacial, facilmente se verifica que, apesar de não existir um consenso total sobre a melhor metodologia para o desenho do *data warehouse*, existem alguns pontos em comum entre as várias propostas. Todos os investigadores distinguem as dimensões não espaciais, das dimensões que

possuem atributos com características espaciais. Para além deste ponto em comum, os investigadores distinguem as medidas espaciais das medidas não espaciais. Estas análises aos vários modelos foram bastante importantes e permitiram perceber de que forma os dados espaciais são identificados e caracterizados. Contudo, nas publicações disponíveis, não foi possível identificar uma metodologia para todas as fases de um projeto de *data warehouse* espacial.

Para a realização deste trabalho e consequente obtenção de resultados utilizou-se vários modelos propostos por investigadores, devido à limitação anteriormente identificada. Por forma a que todo o processo de desenho e implementação do *data warehouse* seguisse uma metodologia amplamente estudada e testada, foi utilizado o método de Kimball combinado com outros métodos, para a criação do *data warehouse*. Este método, apesar de ser mais apropriado para *data warehouses* convencionais, ele fornece uma boa base para o desenho e implementação de *data warehouses*. Outro método utilizado neste processo foi o método proposto por Fidalgo para o desenho de *data mart*. Além destes métodos, foi utilizada toda a informação recolhida das várias propostas dos investigadores por forma a que modelo seguido fosse o mais completo possível. Esta combinação de modelos permitiu a construção de uma solução de *data warehousing* espacial sustentada e completa.

A análise inicial aos modelos propostos na literatura para a implementação de *data warehouses* espaciais facilitou bastante o processo, mas foi necessário efetuar investigações complementares, relativamente à geocodificação dos dados com características espaciais e ao armazenamento de dados espaciais na base de dados. A geocodificação dos dados é um processo complexo, mas existem no mercado várias soluções para a realização desta atividade, com custos diferentes, dependendo dos interesses do projeto. Para este trabalho, foram utilizadas duas soluções disponíveis na Internet, que convertem os dados com características espaciais em coordenadas geográficas. Para a conversão dos dados do cliente foi utilizado um *website*, enquanto para geocodificação dos dados relativos aos países e cidades foi utilizada a ferramenta *shape2sql*. Após a obtenção destes resultados foi necessário armazenar a informação geográfica na base de dados. A maioria das bases de dados presentes no mercado atualmente possui o tipo *geometry* e o tipo *geography* para o armazenamento de dados

espaciais, ambos utilizados neste trabalho. Este processo, em conjunto com a criação do processo de ETL, consumiu grande parte do tempo de realização deste trabalho.

No processo de ETL foi necessário tomar algumas decisões, com o objetivo de focar o trabalho apenas na transformação do *data warehouse* convencional num *data warehouse* espacial. Tendo em conta esta opção, tomou-se a decisão de extrair diretamente a informação do *data warehouse* convencional, em vez de usar as fontes de dados no processo de ETL. Contudo, foi necessário criar uma base de dados como fonte de dados externa, que possuía os dados espaciais necessários. Após a conclusão do processo de ETL, analisou-se o resultado do povoamento do *data warehouse* espacial, tendo especial atenção às novas dimensões espaciais e às dimensões que sofreram alterações.

Em seguida são apresentados, de forma resumida, as várias fases deste trabalho, desde do início dos trabalhos até à análise dos resultados:

- Recolha e análise dos vários modelos propostos na literatura até à data atual.
- O Levantamento de requisitos, com o principal foco nos dados espaciais e identificação da existência dos mesmos.
- O desenho do *data mart*, onde foram identificadas as dimensões espaciais, acrescentadas ao *data warehouse* convencional, bem como, a identificação das alterações a efetuar nas dimensões pré-existentes.
- Investigação e análise de vários métodos de geocodificação de dados.
- A criação de uma base de dados externa, com dados espaciais necessários para o data warehouse espacial, utilizando duas formas distintas de povoar as suas tabelas.
- A criação de um ETL, capaz de povoar as novas dimensões espaciais, bem como, atualizar as dimensões pré existentes, utilizando, como auxilio, uma *data stating area*.
- A validação dos dados presentes nas dimensões espaciais.

Após a realização deste trabalho, é possível concluir que os *data warehouses* espaciais, constituem uma mais-valia para os agentes de decisão e que possuem várias vantagens em relação aos *data warehouses* convencionais, como a representação de dados num mapa, o que facilita a visualização de padrões, entre outras. Tendo em conta o valor acrescido que este tipo ferramenta traz

para as organizações, acredito que devia ser uma área com mais investimento, por parte das organizações e dos investigadores por forma a criar uma metodologia capaz de reunir o consenso e de tornar este processo mais simplificado. Contudo, como ficou provado ao longo deste trabalho, a conversão de um *data warehouse* convencional em um *data warehouse* espacial, apesar de ser um processo complexo e tendo em conta as vantagens que traz para a organização, é algo que deveria ser estudado pelas organizações.

5.2 Linhas de Orientação para Trabalho Futuro

Este trabalho pode ser considerado útil para qualquer pessoa interessada na criação de um *data warehouse* espacial ou na conversão de um *data warehouse* convencional num *data warehouse* espacial, pois fornece indicações precisas sobre as várias fases de um projeto deste tipo. Contudo, a criação do *data warehouse* espacial é, apenas a base de um sistema de *data warehousing*. Por isso, seria interessante prosseguir os trabalhos deste projeto, analisando os restantes elementos que constituem um sistema de *data warehousing*.

Os agentes de decisão procuram efetuar as suas análises de forma rápida e simples. Com intuito de fornecer estas características aos utilizadores do *data warehouse*, para além da construção de um cubo, de forma melhorar a performance das queries efetuadas sobre os dados presentes no *data warehouse*. Após concluída a construção do cubo, seria interessante construir uma ferramenta SOLAP, que iria permitir aos utilizadores de forma simples, visualizar os dados do *data warehouse* sobre a forma de mapas.

Outra vertente que seria interessante adicionar, seria um processo de ETL para o povoamento regular do *data warehouse* espacial. O processo de ETL proposto neste trabalho apenas contempla o povoamento inicial. Para tal, seria necessário construir uma ferramenta capaz de automatizar o processo de geocodificação dos dados extraídos das fontes de dados, de forma a serem adicionados ao *data warehouse* espacial. Com a construção da ferramenta OLAP previamente referida e do processo de ETL para o povoamento regular, o sistema de *data warehousing* resultante dessas adições seria capaz de lidar com todas as necessidades dos agentes de decisão.

Bibliografia

Yvan Bédard, Tim Merrett and Jiawei Han. " Fundamentals of spatial data warehousing for geographic knowledge discovery". 2001, in Geographic Data Mining and Knowledge Discovery. H. Miller and J. Han, Taylor & Francis.

Inmon, W.H. "Building the Data Warehouse" 3rd ed. New York: John Wiley & Sons, Inc, 2002. Gray, P. and Israel, C. "The Data Warehouse Industry". 1999.

Rivest,S., Bédard,Y. and Marchand,P. (2001). "TOWARD BETTER SUPPORT FOR SPATIAL DECISION MAKING: DEFINING THE CHARACTERISTICS OF SPATIAL ON-‐LINE ANALYTICAL PROCESSING (SOLAP)". Geomatica. 55(4), 539-‐555.

Franklin, C. 1992. "An Introduction to Greographic Information Systems: Linking Maps to Databases". Database. April, 13-‐21.

Bédard, Y. 2003. "Integrating GIS componentes with knowledge discovery technology for environmental health decision support". International Journal of Medical Informatics, 70, 79-‐94.

Malinowski, E. and Zimányi, E. 2008. "Advanced Data Warehouse: Design From Conventional to Spatial and Temporal Applications. Springer.

Adrian Bridgwater. (2009). "IDC Study: Boom Time for Data Warehouse Vendors". Disponível: http://www.zdnet.com/idc-‐study-‐boom-‐time-‐for -‐data-‐warehouse-‐vendors-‐4010011788/. Visto pela última vez a 4 de Novembro 2012.

Stefanovic, N., Han,J. and Koperski,K. 1998. "Selective Materialization: An Efficietn Method for Spatial Data Cube Construction". Pacific-‐Asia Conference on Knowledge Discovery and Data Mining.

Kimball, R. and Ross,M. 2002. “The Data Warehouse Toolkit”. 2nd ed. NY:John Wiley & Sons, Inc..

Golfarelli,M. and Rizzi,S. 2009. “Data Warehouse Design: Modern Principles and Methodologies”. McGraw–-Hill Osborne Media.

Stefanovic, N., Han, J. and Koperski, K.. 2000. “Object–-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes”. IEEE Transactions On Knowledge and Data Engineering. 12(6), November/December.

Bédard, Y., Rivest, S., Proulx, M.J., Nadeau, M.. 2003. “SOLAP: A New Type of User Interface to Support Spatio–-Temporal Multidimensional Data Exploration and Analysis”. Laval University.

Shekhar, S. and Chawla, S.. 2003. “Spatial Databases: A Tour”. Prentice Hall.

Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, Panos Vassiliadis (2003). Fundamentals of Data Warehouses. 2nd ed. Springer.

Hajer Bâazaoui Zghal, Sami Faïz, Henda Ben Ghézala. (2003). “CASME : A CASE Tool for Spatial Data Marts Design and Generation”.

Robson N. Fidalgo, Valeria C. Times, Joel Silva, Fernando F. Souza, Ana C. Salgado. (2010). “Providing Multidimensional and Geographical Integration Based on a GDW and Metamodels”. Journal of information and data management.

Alastair Aitchison (2012). “Pro Spatial with SQL Server 2012”. Apress.

Maria Luisa Damiani ,Stefano Spaccapietra. "Spatial Data Warehouse Modelling.". 2009. IGI Global.

Menneck, B., Higgins, G.. (1999). “Spatial Data in the Data Warehouse: A Nomenclature for Design and Use”. Proceedings of the 5th Annual Americas Conference on Information Systems.

Gonzales, M.L.. (1999). “Spatial OLAP: conquering Geography”. DB2 magazine.

Yuan, M., Buttenfield, B., Gahegan, M. N., and Miller, H.. (2000). “Geospatial Data Mining and Knowledge Discovery”. Research Challenges in Geographic Information Science.

Indulska, M., Orlowska, M.. (2002). “On aggregation issues in spatial data management”. CM International Conference Proceeding Series, Proceedings of the thirteenth Australasian conference on Database technologies.

Thiago, L., Cristina D., Valéria, C., Anjolina G., Ricardo R.. (2009). “The impact of spatial data redundancy on SOLAP query performance”. Journal of the Brazilian Computer Society.

Margy Ross. (2008). “Snowflakes, Outriggers, and Bridges”. Available: http://www.kimballgroup.com/2008/09/03/design-tip-105-snowflakes-outriggers-and-bridges/. Last accessed 20 Outubro 2013.

Carr, W.. (2006). “Philosophy, Methodology and Action Research”.Journal of Philosophy of Education.

Malinowski, E. (2006). "Concepts and Methodological Framework For Spatio- Temporal Data Warehouse Design".

Michelle A. Poolet. 2007. “*Data Warehousing: Dimension Basics”*. http://sqlmag.com/business-intelligence/data-warehousing-dimension-basics. [Último acesso 24 Outubro 2013].

M. Demarest. 1997. “The Politics of Data Warehousing”. http://www.uncg.edu/ism/ism611/politics.pdf. [Último acesso 09 Outubro 2013].

João Ferreira, Miguel Miranda, António Abelha e José Machado, 2010. “O Processo ETL em Sistemas Data Warehouse”. INForum.