Sistema de Business Intelligence aplicado aos Grupos de Diagnósticos Homogéneos

por

Manuel Pedro Antunes Barrento

Dissertação apresentada como requisito

parcial para obtenção do grau de

Mestre em Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão da Informação

da

Universidade Nova de Lisboa

SISTEMA DE BUSINESS INTELLIGENCE

APLICADO AOS

GRUPOS DE DIAGNÓSTICOS HOMOGÉNEOS

Dissertação Orientada por:

Professor Doutor Miguel de Castro Simões Ferreira Neto

Professora Doutora Maria do Rosário Oliveira Martins

Novembro

2010

Índice

1 Agradecimentos

O percurso de elaboração da presente tese não foi fácil, exigiu bastante esforço, empenho, motivação e sobretudo coragem, no entanto, sem o apoio da minha família, aos quais desde já agradeço, não teria sido possível terminá-la.

Quero agradecer bastante ao meu orientador da tese de mestrado, o Professor Doutor Miguel de Castro Neto que sempre acreditou em mim, apoiou-me, potenciou-me para a investigação e deu-me total autonomia para seguir o meu caminho para a obtenção do grau de Mestre. Sinceramente, fiquei deveras satisfeito com as orientações que tive.

Também uma palavra de agradecimento para a minha co-orientadora, a Professora Doutora Maria do Rosário Oliveira Martins pela sua disponibilidade para o sucesso da presente tese.

Por último, um agradecimento muito especial à ACCS (Administração Central do Sistema de Saúde) pela disponibilização dos dados e, sobretudo, na pessoa da Dra. Manuela Rolim, pelos muitos esclarecimentos a dúvidas que foram surgindo durante o longo percurso que foi este projecto.

2 Resumo

A presente tese baseia-se na implementação de raiz de um Sistema de Bussiness Intelligence (BI), aplicado aos Grupos de Diagnósticos Homogéneos (GDH's).

São abordadas todas as fases de desenvolvimento, desde a modelação do Data Warehouse, desenho e construção dos procedimentos de ETL, até à elaboração de relatórios e "dashboards" para análise da informação.

A ideia da presente tese, surgiu com o intuito de contribuir para o melhoramento da monitorização da classificação dos pacientes em GDH's, de forma a funcionar como instrumento de suporte à tomada de decisão.

A principal meta deste sistema de BI, consiste em aproximar os utilizadores finais (que geralmente não são técnicos de informática) de dados que podem ser facilmente consultados e analisados, de forma a poder suportar a tomada de decisão. Para atingir esta meta, é necessário efectuar grandes transformações aos dados provenientes de sistemas operacionais, de forma a torná-los legíveis e de fácil acesso a quem tem de tomar decisões.

3 Abstract

This thesis is based on developing from scratch a Bussiness Intelligence (BI) system to Diagnosis Related Groups (DRG's). All phases of development are covered from the modeling of Data Warehouse design and construction of the ETL procedures to “reporting” and “dashboards” that are the main tools to analyze information. This idea arose in order to improve monitoring of the evolution of DRG's and how to support decision making.

The main goal of this BI system is to approximate the end-users (who are not computer technicians) of data that can be easily accessed and analyzed. To achieve this goal it is necessary to make major changes on data from operating systems to make it readable and accessible to anyone who has to make decisions.

4 Lista de Figuras

5 Siglas

ACSS – Administração Central do Sistema de Saúde

AS – Analysis Services

BD – Base de Dados

BI – Business Intelligence

BIDS – Business Intelligence Development Studio

CD-ROM – Compact Disc Read Only Memory

CID-9-MC – Classificação Internacional de Doenças, 9ª Revisão, Modificação Clínica

CMS – Centers for Medicare and Medicaid Services

CRM – Customer Relationship Management

ERP – Enterprise Resource Planning

DM – Data Mart

DRG - Diagnosis Related Groups

DW – Data Warehouse

HCFA - Health Care and Financing Administration

I/O – Input/Output

GCD – Grandes Categorias Diagnósticas

GDH – Grupos de Diagnósticos Homogéneos

JPEG – Joint Photographic Experts Group

ODBC – Open Data Base Connectivity

OLAP – On-Line Analytical Processing

MS – Management Studio

RAM – Random Access Memory

RFID – Radio-Frequency Identification

SCD – Slowly Changing Dimensions

SA – Staging Area

SAS – Statistical Analysis System

SK – Surrogate Keys

SMTP – Simple Mail Transfer Protocol

SNS – Serviço Nacional de Saúde

SPSS – Statistical Package for the Social Sciences

SQL – Structured Query Language

TI – Tecnologias de Informação

T-SQL – Transact SQL

6 Introdução

Os sistemas de saúde são parte integrante de um sistema social onde a adequação das suas respostas, em termos de saúde, às expectativas da população é um objectivo complexo só realizável com a participação de todos aqueles que se preocupam com o futuro das organizações e das boas práticas de saúde, numa altura em que a referência à tecnologia assume um papel de extrema relevância.

O termo da **Business Intelligence** (BI) não é tão recente como poderíamos imaginar, tomando em consideração a sua actual e crescente popularidade.

O seu conceito prático, já era usado na antiguidade. Algumas sociedades do Médio Oriente, utilizavam os princípios básicos da BI quando cruzavam informações, obtidas de diversas fontes, sobre o funcionamento da mãe natureza em prol dos benefícios directos para as comunidades aldeãs. Depois de recolhida a informação, esta era objecto de análise, como por exemplo, o comportamento das marés, os períodos chuvosos e de seca, a posição dos astros, entre outras, dais quais iriam resultar importantes decisões de forma a rentabilizar e melhorar a qualidade de vida das respectivas comunidades (Primak, 2009).

É evidente que o Mundo em que vivemos mudou desde então, porém o conceito permanece inalterado. A necessidade de cruzar informações para a realização de uma gestão empresarial eficiente e eficaz, é actualmente uma realidade tão presente na nossa sociedade como no passado.

O interesse actual pela BI tem crescido exponencialmente, na medida em que a sua aplicação possibilita a qualquer organização, seja ela de saúde ou outra, realizar uma série de análises e previsões, de forma a agilizar os processos relacionados com as tomadas de decisão. É o que defende Howard Dresner[1], vice-presidente da empresa Gartner[2], que é detentor da paternidade do termo BI.

[1] Howard Dresner foi vice-presidente e investigador da Gartner. Possui 22 anos de experiência na área de BI e já percorreu o mundo em conferências sobre esta temática. A sua biografia pode ser consultada em: http://www.gartner.com/research/fellows/asset_79427_1175.jsp

[2] A Gartner é uma empresa líder mundial em consultoria e investigação na área de tecnologias de informação. Na área de BI, a Gartner é conhecido por efectuar estudos comparativos sobre as tecnologias existentes no mercado e seus fornecedores. No capítulo "Ferramentas de Business Intelligence" da presente tese é apresentado um estudo de forma a compreender-se o posicionamento e potencial das ferramentas de BI e respectivos fornecedores. Mais informações e detalhes sobre a Gartner podem ser consultados em: http://www.gartner.com/technology/about.jsp

7 Enquadramento da necessidade do projecto

Este projecto tem a sua origem na necessidade de resposta há crescente exigência de informação, clara e concisa, sobretudo na área da saúde.

De forma a suportar as tomadas de decisão, surgiu a necessidade de desenvolver este projecto de raiz, ou seja, um sistema de BI completo para uma subárea da saúde, os GDH´s, que podemos definir como sendo um sistema de classificação de doentes internados, em grupos clinicamente coerentes e similares do ponto de vista do consumo de recursos.

Hoje em dia, os dados relacionados com os GDH´s são muitas vezes armazenados em diversos sistemas operacionais (base de dados e ficheiros) num formato deveras inacessível aos utilizadores finais (estes, muitas vezes, só possuem conhecimentos informáticos na óptica do utilizador, limitando-se a acompanhar o estado actual dos GDH´s, assim como a sua evolução segundo um grande conjunto de características como o diagnóstico, o hospital, o distrito, a sazonalidade, entre outras), o que os leva a uma total dependência do departamento de informática, para obterem os dados num formato e características que lhes respondam às suas necessidades. Se, por cada análise que os utilizadores pretendam fazer sobre dados dos GDH´s, corresponder um pedido para obtenção dos mesmos num aspecto amigável, ao departamento de informática, leva-nos a satisfazer tal necessidade através de um sistema que se “aproxime” mais dos utilizadores finais.

Uma outra limitação dos actuais sistemas operacionais que armazenam os dados sobre os GDH´s, tem como base a ausência de uma visão única dos dados, visto estes se encontrarem dispersos em vários sistemas, o que dificulta o acompanhamento do seu estado actual e futura evolução.

Por outro lado, se os dados sobre os GDH´s vão aumentando surge a necessidade de armazenar histórico de forma a possibilitar a comparação entre anos, meses homólogos, evoluções nos últimos cinco anos, entre outras análises. Estas análises, não são passíveis de realização, em tempo útil, com os sistemas operacionais, vocacionados apenas para registar transacções relativas aos GDH´s e não guardar histórico, visto que, o seu foco não é permitir análises.

Assim sendo, emergiu deste diagnóstico a oportunidade de realizar uma tese de mestrado em que o seu foco tem como base o desenvolvimento de raiz de um sistema de BI aplicado aos GDH´s.

O sistema de BI, criado para o efeito, é constituído por várias componentes que serão detalhadas nesta tese mais à frente, no entanto, é importante salientar que o desenvolvimento deste sistema é complexo, mas de fácil apresentação dos dados (forma simples e concisa) que são apresentados aos utilizadores, cuja finalidade se prende com cruzamentos de dados e correspondentes análises, elaboradas de uma forma intuitiva sem ser necessário grandes conhecimentos informáticos (apenas na óptica do utilizador final).

Por fim, os sistemas de BI, para além de colmatarem as lacunas dos sistemas operacionais, permitem o armazenamento de histórico, o que permite aos utilizadores efectuarem análises comparativas, de evolução, melhorando assim, a tomada de decisões, tanto a nível operacional como estratégico.

8 Objectivo

É com base nos conceitos de arquitectura de um Data Warehouse[3] (DW) que se pretende aprofundar/inovar a modelização do mesmo aplicado à Saúde, bem como, quais as técnicas usadas, o modelo em estrela[4] ("star-schema") e floco de neve[5] ("snow-flake"), e, quais as dimensões[6] e factos[7] que são importantes para a sua construção no sentido de permitir analisar dados clínicos de diversas perspectivas multidimensionais (que consistem em análises das métricas por uma ou mais dimensões definidas no modelo).

A grande vantagem dos sistemas de BI é a unificação da informação que se encontra dispersa por diversos sistemas fonte, aos quais é impossível aceder para efectuar consultas/análises multidimensionais aos dados. A partir deste ponto de partida, o objectivo da presente tese, consiste em desenvolver de raiz uma solução de BI completa ("end-to-end"), que inclui procedimentos de ETL para extracção, transformação e carregamentos dos dados fonte, uma base de dados intermédia para auxiliar o processamento dos dados, um modelo de DW e por último uma plataforma de "reporting" aplicada aos GDH's visto que o sistema operacional que guarda este tipo de informação não permite consultas/análises simples do ponto de vista do utilizador final.

Assim sendo, o modelo de DW desenvolvido, irá dar legibilidade ao modelo de dados, bem como colocá-lo num formato em que seja possível efectuar um grande número de consultas/análises, sem grandes dificuldades, através de uma aplicação de "reporting", que pretende melhorar o suporte à decisão.

Todos os conceitos mencionados nesta secção serão detalhados ao longo da presente tese.

[3] Grande repositório (base de dados) central de dados consolidados estruturado para realizar consultas analíticas complexas, e adequado para o apoio à tomada de decisão.
[4] Definição no capítulo de Conceitos de Modelação de Dados.
[5] Definição no capítulo de Conceitos de Modelação de Dados.
[6] Definição no capítulo de Conceitos de Modelação de Dados.
[7] Definição no capítulo de Conceitos de Modelação de Dados.

9 Business Intelligence

9.1 Arquitectura da Business Intelligence

Segundo vários analistas de informação, quase a totalidade dos desafios técnicos que se prendem com obtenção, tratamento e análise de grandes quantidades de dados, encontram-se resolvidos. De forma a constatar, o atrás exposto, veja-se alguns filmes de Hollywood, em um hacker, através de um computador portátil acede a um número ilimitado de dados que estão acessíveis na internet e que tenha conhecimento de algumas passwords (algumas delas obtidas em segundos, através de programas de "hacking"), resulte no roubo de identidade, etc. Felizmente que para a nossa privacidade não é assim tão fácil.

Certamente que se tornou fácil a captura e armazenamento de grandes quantidades de dados. A quantidade de dados em número são ainda hoje difíceis de absorver, no entanto, em termos de espaço, os volumes de dados têm crescido dos megabytes para os gigabytes, e depois para os terabytes. Algumas bases de dados corporativas estão rapidamente a chegar aos petabytes (1.000 terabytes). Enquanto que os computadores e servidores antigos já não têm capacidade e potência para lidar com os volumes de dados necessários para as aplicações analíticas, os recentes processadores de 64 bits conseguem-no.

No entanto, no dia-a-dia das organizações estas confrontam-se com a grande quantidade de dados que geram, e que raramente sabem o que fazer com eles. Os dados carregados nos sistemas são como uma caixa de fotografias que é guardada no sótão à espera que um dia seja encontrada uma solução para o armazenamento caótico. Além disso, a maioria dos departamentos de TI investem excessivamente no apoio e manutenção de capacidades básicas transaccionais. Ao contrário da vanguarda analítica, até mesmo em organizações com sistemas transaccionais de referência apresentam grandes limitações na limpeza de dados e na integração com aplicações analíticas (Davenport & Harris, 2007).

Em suma, as tecnologias de informação melhoraram a capacidade de armazenamento de grandes volumetrias de dados, e a maioria das organizações não acompanhou a sua evolução na capacidade de os gerir, analisar e aplicar. No entanto, as organizações que competem na componente analítica podem não ter resolvido todos estes problemas, mas ao apostarem em ferramentas analíticas apresentam um melhor desempenho do que os seus concorrentes.

9.2 Componente Analítica e Arquitectura da Business Intelligence

Todas as organizações com sérias aspirações analíticas têm um departamento ou equipa de TI envolvida no projecto de BI. Por exemplo, através da introdução da componente analítica nos processos de negócio, o departamento ou equipa de TI pode ajudar a desenvolver e a manter uma vantagem competitiva da organização.

Determinar as capacidades técnicas necessárias para a competitividade analítica requer uma colaboração próxima entre as TI e os gestores de negócio. Este é um princípio que as organizações como a Progressive Insurance (seguradora) compreendem plenamente. Glen Renwick, director executivo da Progressive Insurance e ex-chefe de TI, sabe como é crítico para alinhar as TI com a estratégia de negócio *"… Aqui na Progressive temos a equipa de TI a trabalhar lado a lado com a equipa organizacional no sentido de resolverem problemas técnicos que têm impacto no negócio. Do lado do negócio, é reconhecido a importância das TI no core-business. Os nossos planos de negócios e de TI estão intrinsecamente ligados, porque os seus objectivos de trabalho também o estão.".* Encontramos este mesmo nível de alinhamento de TI/negócios em muitas outras organizações.

Na componente analítica, as organizações, de forma a manterem-se competitivas, estabelecem um guia de princípios para assegurar quais os investimentos em TI que reflectem as suas prioridades corporativas. Os princípios devem incluir as seguintes afirmações:

- O risco associado com o conflito de fontes de informação deve ser reduzido.

- As aplicações devem ser integradas, visto que a componente analítica é transversal a toda a organização.

- A componente analítica deve estar habilitada para fazer parte da estratégia da organização.

A responsabilidade pela obtenção dos dados, tecnologia e correcção dos processos é da competência do denominado arquitecto de TI. Este arquitecto, em estreita colaboração com o CIO, deve determinar como os componentes da infraestrutura de TI (hardware, software e redes) vão trabalhar juntos para fornecer os dados, a tecnologia e o suporte necessários para o negócio. Esta tarefa é mais fácil para o start-ups como a Netflix (empresa líder mundial em serviços de subscrição pela internet de filmes e programas de televisão), que pode criar o seu

ambiente de TI com a componente analítica em mente desde o início. No entanto, em grandes organizações instituídas, a infra-estrutura de TI pode, por vezes, parecer que foi construída através de uma série de empregos (part-time) de fim-de-semana. O sistema faz o trabalho para o qual foi concebido, mas é susceptível de criar problemas quando for aplicado a uma outra finalidade.

Para assegurar que o ambiente de TI cobre todas as necessidades de uma organização em cada etapa da vertente analítica, essas mesmas organizações devem incorporar a componente analítica e outras tecnologias de BI na sua arquitectura de TI.

As organizações já estabelecidas seguem um processo evolutivo para desenvolver as suas capacidades analíticas em TI, baseado em etapas:

Etapa 1. - A organização é atormentada pela falta de dados ou da sua má qualidade, múltiplas definições dos seus dados, e sistemas mal integrados.

Etapa 2. - A organização recolhe dados transaccionais de forma eficiente, mas muitas das vezes faltam os dados correctos para uma melhor tomada de decisão.

Etapa 3. - A organização tem uma proliferação de ferramentas de BI e data marts (base de dados de pequenas dimensões que respondem apenas a unidades de negócio específicas), mas a maior parte dos dados mantêm-se desintegrados, desnormalizados e inacessíveis.

Etapa 4. - A organização possui dados de grande qualidade, um plano analítico empresarial vasto, processos e políticas de TI, e alguns automatismos analíticos.

Etapa 5. - A organização tem uma arquitectura analítica empresarial completa, toda ela automatizada e integrada em processos altamente sofisticados.

Os departamentos de TI usam o termo BI para abranger não só a componente analítica – a utilização de dados para análise, previsão, predição, optimização, entre outras – como também os processos e tecnologias usadas para agrupar, gerir e reportar dados orientados para a decisão. A arquitectura de BI (um subconjunto da arquitectura global de TI) engloba um vasto conjunto empresarial de sistemas, aplicações, e políticas de processos que habilitam a

sofisticação da componente analítica, permitindo que os dados, conteúdo e análises fluam em tempo útil a quem necessita da informação.

Os pontos importantes para que uma infra-estrutura de TI seja competitiva na componente analítica são:

- Os analistas têm acesso directo (quase instantâneo) aos dados, ou seja, se possível em tempo-real.

- Os profissionais que trabalham a informação passam o tempo a analisar os dados e a compreender as suas implicações ao invés de os consolidarem e formatarem.

- Os gestores dedicam-se a melhorar os processos e a performance do negócio com base nos dados, relatórios e sistemas transaccionais.

- Os gestores nunca discutem sobre a precisão/exactidão dos números apresentados.

- Os dados são geridos numa perspectiva empresarial e transversal através do seu ciclo de vida, que começa com a sua criação até ao seu arquivo ou destruição.

- Os processos críticos para a tomada de decisão são altamente automatizados e integrados.

- Os dados, por rotina, são automaticamente partilhados por toda a organização, bem como pelos seus clientes e fornecedores.

- Os relatórios e as análises sintetizam a informação de diversas fontes.

- As organizações gerem os dados como um recurso estratégico corporativo em todas as iniciativas do negócio.

A arquitectura de BI tem que ser construída de forma a ser capaz de fornecer a informação de uma forma rápida, fiável e precisa, para ajudar os estatísticos, analistas de informação, chefes funcionais e gestão de topo a tomar decisões de complexidade variada. Também disponibiliza

informação através de diversos canais de distribuição incluindo os relatórios tradicionais, ferramentas de análises "ad hoc", "dashboards" corporativos, folhas de cálculo, e-mail e alertas por SMS. A gestão de dados é complexa e dispendiosa, como exemplo a empresa americana Amazon.com, gastou mais de dez anos e mil milhões de dólares para construir, organizar e proteger os seus data warehouses.

O cumprimento dos requisitos legais e reguladores da informação é outra actividade que depende de uma arquitectura de BI robusta. Por exemplo, a lei de Sarbanes e Oxley de 2002 requer que executivos, auditores e outros utilizadores dos dados corporativos para demonstrem que as suas decisões são baseadas em dados oficiais, fiáveis, com significado e precisos. Também requer que os mesmos confirmem que os dados fornecem uma visão clara do negócio, das suas tendências, riscos e oportunidades.

Conceptualmente, é necessário dividir a arquitectura de BI em seis elementos (Figura 1):

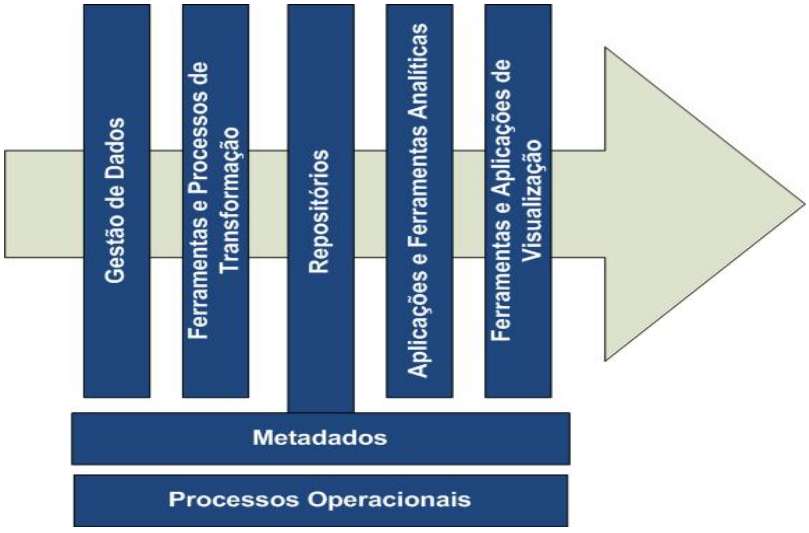


Figura 1 – Arquitectura de BI. Adaptado de: (Davenport & Harris, 2007).

- A gestão de dados é que define como é que os dados são obtidos e geridos.

- São as ferramentas e os processos de transformação que descrevem como é que os dados são extraídos, enriquecidos, transmitidos e carregados nas bases de dados.

- Os repositórios armazenam e organizam os dados e os metadados (informação sobre os dados) para seu uso.

- Aplicações e outras ferramentas analíticas são usadas para análise dos dados.

- Aplicações e ferramentas de visualização guiam os analistas de informação de como aceder, visualizar e manipular os dados.

- Os processos operacionais determinam o grau de importância das actividades administrativas que são endereçadas, como por exemplo, a segurança, o tratamento de erros, a auditoria, o arquivo e a privacidade.

Cada elemento da Figura 1 será analisado em detalhe com particular atenção aos dados, visto que são o foco de todas as decisões numa arquitectura de BI.

9.2.1 Gestão de dados

O objectivo de uma estratégia de gestão de dados bem desenhada, consiste em assegurar que a organização tem a informação correcta e que a usa de forma apropriada. Grandes organizações investem milhões de dólares em sistemas que captam dados de todas as fontes. Sistemas de ERP[8] (enterprise resource planning), CRM[9] (customer relationship management), entre outros, garantem que nenhuma transacção ocorra sem ficar registada. Muitas

[8] Os ERP´s (Enterprise Resource Planning) ou Sistemas Integrados de Gestão Empresarial são sistemas de informação que integram todos os dados e processos da uma organização num sistema único. A integração pode ser vista sob a perspectiva funcional (sistemas de: finanças, contabilidade, recursos humanos, manufactura, marketing, vendas, compras, entre outras). Os ERPs em termos gerais, são uma plataforma de software desenvolvida para integrar os diversos departamentos de uma organização, possibilitando a automatização e o armazenamento de dados do negócio.

[9] Em termos simples, o CRM pode ser entendido como uma estratégia que permite à organização como um todo ter uma visão única do seu cliente e, a partir daí, saber explorar as oportunidades de negócio. Para isso é necessário aproveitar todas as interacções que tem com o cliente no sentido de captar dados e transformá-los em informações que possam ser disseminadas pela organização, de forma a que todos os departamentos - call center, vendas, marketing, administração, entrem outros - vejam o cliente da mesma forma, ou seja, saibam quem ele é, seus gostos e preferências, quantas vezes ligou, reclamações que fez, sugestões que deu, quanto traz de valor para a empresa, entre outras. Actualmente, poucas organizações conhecem seus clientes com essa profundidade. Porque será este factor importante? Estudos feitos no mercado norte-americano concluíram que, num prazo de cinco anos, uma companhia perde metade dos seus clientes e gasta cinco vezes mais na conquista de um novo consumidor do que na retenção do antigo. Outro dado interessante, é que um comprador satisfeito comenta sua compra com cinco pessoas, enquanto que um insatisfeito queixa-se da organização com nove. Por esses motivos, os princípios básicos do CRM sustentam a necessidade de saber identificar, diferenciar (por seu valor e necessidades) e interagir com o cliente para estabelecer uma relação de aprendizagem contínua para que seja possível oferecer um atendimento personalizado e satisfatório para ambos, ou seja, consumidor e empresa (Spiner, 2006).

organizações também adquirem externamente dados, ou seja, adquirem-nos a organizações externas de forma a enriquecerem a informação sobre o seu negócio.

Neste ambiente, a sobrecarga de dados pode ser um problema para os executivos. No entanto, o grande desafio que as organizações enfrentam em relação aos dados está relacionado com a sua qualidade, nomeadamente a informação inconsistente, fragmentada e fora do contexto. Mesmo as melhores organizações têm muitas dificuldades em resolver os seus problemas com os dados. As organizações que competem na componente analítica dedicam uma atenção extraordinária aos processos e políticas de gestão de dados. Por exemplo, a "Capital One" (empresa de cartões de crédito) estima que 25% do seu departamento de TI trabalha na resolução de problemas com os dados – que representa uma grande percentagem em comparação com outras organizações.

Há um ganho significativo para aqueles que investem em controlar a gestão de dados. Por exemplo, a Continental Airlines integra no seu DW cerca de 10 terabytes de dados de vinte e cinco sistemas operacionais. Os dados são usados para aplicações analíticas que despoletam alertas em tempo-real e análises estratégicas de longo prazo. Os alertas notificam os agentes dos clientes sobre atrasos de voos e que alternativas existem. Os analistas de marketing usam outros dados, recolhidos pelo sistema para estudar as tendências do cliente e dos preços, os analistas de logística planeiam uma optimização dos recursos (aviões e tripulações). A empresa estima que já poupou mais de 250 dólares nos primeiros cinco anos das suas actividades de "data warehousing" e BI.

Para alcançar os benefícios de uma competitividade analítica, especialistas de negócio e em TI, devem resolver os problemas dos dados respondendo a cinco perguntas:

- Relevância dos dados: O que é necessário para competir na componente analítica?
- Fontes de dados: Onde é que estes dados podem ser obtidos?
- Volumetria dos dados: Quantos dados é que são necessários?
- Qualidade dos dados: Como é que os dados podem ser mais

	precisos e enriquecidos para análise?
• Controlo de dados:	Que regras e processos são necessários para gerir os dados desde a sua criação até ao seu destino final?

Vamos analisar mais em pormenor estas questões.

Que dados é que são necessários para competir na componente analítica?

A verdadeira pergunta que carece de resposta, é a que se prende com os dados necessários e que são fundamentais para a diferenciação competitiva e desempenho do negócio? Para responder a esta questão, os gestores necessitam de ter uma compreensão clara da capacidade da organização, que actividades é que a suportam essa capacidade, e que relação existe entre as métricas operacionais e estratégicas da empresa, assim como o desempenho do negócio.

Mas assegurar que os analistas de informação têm acessos aos dados necessários pode ser uma tarefa complicada. Por vezes, é necessário um número, por exemplo, qual o "score" de créditos que um banco atingiu. Mas nem tudo pode ser reduzido a um número. O rácio de desempenho dos funcionários não reflecte uma imagem completa do seu trabalho durante um ano como uma avaliação escrita da sua chefia. A situação torna-se mais complicada quando os recursos do negócio e das TI se culpam mutuamente quando são adquiridos dados errados ou pela indisponibilidade dos dados correctos. Estudos conduzidos pela "Gartner" mostram consecutivamente que os profissionais de TI acreditam que os gestores do negócio não compreendem que dados é que necessitam. Por outro lado, os gestores de negócio afirmam que os profissionais de TI não dominam o negócio ao ponto de dar significado aos dados disponíveis. Até ao presente momento ainda não existe uma solução para este problema, no entanto, o começo da sua resolução consiste em colocar recursos do negócio e de TI a trabalhar em equipa. Sem esta cooperação, a capacidade da organização em reunir os dados que necessita para competir na componente analítica está condenada.

Uma questão relacionada com a cooperação entre o negócio e as TI consiste em definir o relacionamento entre os dados usados para análises. É necessário possuir uma experiência considerável de negócio para ajudar as TI a compreender o potencial das relações que existem nos dados e que são essenciais para a organização. A importância desta tarefa pode ser vista num exemplo que envolve clientes de um seguro de saúde. Da perspectiva da companhia de

seguros, existe uma diversidade de clientes (corporativos, que possuem seguro através das suas empresas, assinantes individuais e familiares). Tanto a companhia de seguros como cada pessoa segurada têm ligações com diversas entidades de prestadoras de serviços de saúde, como hospitais e médicos. O médico é o especialista, que presta serviços para certos hospitais, clínicas e outras entidades. Qualquer indivíduo pode fazer um seguro saúde numa das companhias de seguro disponíveis para o efeito. Sem a compreensão da natureza das ligações entre médicos, seguradoras, segurados, e entidades prestadoras de serviços de saúde, a utilidade dos dados para a componente analítica é bastante limitada.

Onde é que os dados podem ser obtidos?

Os dados para os sistemas de BI são provenientes de várias fontes, sendo o ponto crucial a sua gestão, que necessita de ser realizada através de uma infra-estrutura transversal a toda a organização. Só desta forma é que é possível atingir a escalabilidade numa organização. A existência de uma fonte comum de dados para todas as aplicações permite ter uma visão única e consistente da realidade do negócio, que é essencial para todos os que utilizam a componente analítica. Visto que é possível criar este tipo de ambiente factual através da integração e transformação dos dados provenientes de diversos sistemas, as organizações são aconselhadas a actualizar e integrar os seus processos e sistemas transaccionais antes de embarcarem nesta tarefa.

Para informação interna, os sistemas são o início lógico do processo. Por exemplo, uma organização que pretende optimizar a sua cadeia de abastecimento pode começar por uma aplicação de planeamento de pedidos. No entanto, pode ser difícil analisar os dados dos sistemas transaccionais (como o controlo de inventário) visto que a sua finalidade não está direccionada para decisões de gestão. Os sistemas empresariais são aplicações de software integrado que automatizam, cruzam e gerem o fluxo de informação para os processos de negócio tais como o atendimento de pedidos, que muitas das vezes ajudam as organizações a traçar um caminho direccionado para a competitividade analítica, proporcionando dados consistentes, fiáveis e em tempo útil para tarefas de relatórios financeiros e de optimização da cadeia de abastecimento. Os fornecedores de software, estão cada vez a apostar na incorporação da componente analítica nos seus sistemas empresariais de forma a que os utilizadores possam desenvolver previsões de venda e modelos alternativos para solucionar os problemas do negócio.

Além dos sistemas corporativos, os computadores pessoais e os servidores de uma organização estão geralmente carregados de dados, que têm a sua origem em bases de dados, folhas de cálculo, apresentações e relatórios. Em alguns casos, estas fontes de dados são armazenadas numa aplicação central de gestão do conhecimento que geralmente não está disponível para toda a organização.

Para obtenção de informação externa, os gestores podem adquirir os dados pretendidos através de empresas especializadas que prestam informações financeiras de outras organizações e do mercado, consumo de crédito e avaliações do mercado. A todos os níveis, o governo é um dos grandes fornecedores de informação, assim como os "web sites" das organizações, poderoso recurso, visto que são acedidos por diversos clientes e fornecedores.

Os dados podem também ter origem em outras fontes, nomeadamente, e-mail's, aplicações de voz, imagens (mapas e fotos disponíveis através da internet) e biométrica (impressões digitais e identificação da íris). Quanto mais distante tiverem os dados (numéricos e cadeias de caracteres) do "standard", mais complexo será a integração com outros dados para análise.

Mais dados sobre o mundo físico, provenientes da tecnologia de sensores e identificação por rádio-frequência (RFID) começa também a estar disponível. Por exemplo, uma caixa de vinho pode ser monitorizada para validar se está a ser mantida à temperatura correcta.

É difícil e dispendioso obter dados bastante importantes e valiosos para o negócio, como por exemplo, informações sensíveis sobre clientes, lançamentos de novos produtos e estratégias de preço. Os concorrentes analíticos adoptam aproximações inovadoras para ganharem permissão para recolherem os dados que precisam. Por exemplo, a seguradora "Progressive" tinha um programa que consistia em oferecer descontos a clientes que acordassem em instalar um dispositivo que obtivesse dados sobre a forma de condução. O presidente da empresa Peter Lewis, via esta capacidade como a chave para implementar preços mais fiáveis e angariação de bons clientes, isto está reflectido na sua frase "… É sobre o ser capaz de cobrar-lhes por tudo o que acontece, ao invés do que eles [os clientes] dizem que está a acontecer. Então, o que vai acontecer? Nós vamos chegar a todas as pessoas que raramente conduzem, enquanto que os nossos concorrentes irão ficar com as que apresentam riscos mais elevados …".

Qual a volumetria de dados necessária?

As organizações para além de agregarem os dados correctos precisam de capturar grandes volumetrias de forma a calcular tendências e prever comportamentos. O que é uma grande volumetria de dados? Como perspectiva, a organização Wal-Mart (compras on-line) possui uma base de dados de 583 terabytes e em 2006, 20 terabytes seria o tamanho da colecção impressa da biblioteca Congresso nos Estados Unidos. Felizmente que a tecnologia e a as técnicas de "mining" e de gestão de grandes volumetrias de dados estão a fazer um enorme progresso.

Duas armadilhas devem ser tidas na necessidade de grandes quantidades de dados. A primeira está relacionada com a tentação das organizações possuírem ou tentarem reunir todos os dados possíveis, esta preocupação pode-se resumir na frase "caso um dia venha a ser necessário". Se os analistas de informação tiverem de percorrer montanhas digitais de dados irrelevantes irão certamente desistir das ferramentas que usam para analisar os dados. Outro aspecto a ter em atenção, prende-se com o que não pode ser feito, ou seja, os coleccionadores de dados rapidamente chegam à conclusão que não podem reunir todos os dados e mesmo que tentem, os custos para a obtenção dos mesmos, irão ultrapassar os benefícios. Desta forma regressa novamente a questão de saber o que impulsiona valor numa organização; esta compreensão irá prevenir as empresas de reunirem dados de forma indiscriminada (Davenport & Harris, 2007).

A segunda armadilha, tem a ver com a obtenção de dados, que não são fáceis de obter, e que não são necessariamente importantes. Muitos profissionais das TI defendem esta abordagem, porque os liberta da responsabilidade de determinar qual a informação importante e não importante para o negócio. Por exemplo, muitas organizações caem na armadilha de fornecer dados aos gestores que são um subproduto dos sistemas de transacção só porque é o que está disponível. Talvez um dia as tecnologias emergentes eliminem a necessidade de "separar o trigo do joio". Mas até que isso aconteça, é necessário aplicar inteligência humana ao processo para evitar a sobrecarga de dados.

Como é que se pode enriquecer os dados?

Quantidade sem qualidade é uma receita para o falhanço. Os gestores estão a par do problema: num estudo sobre os desafios que as organizações enfrentam no desenvolvimento de capacidades de BI, a qualidade de dados ficou em segundo, perdendo apenas para as limitações de orçamento.

As organizações têm a tendência a armazenar os seus dados em silos funcionais de difícil acesso. Como resultado, os dados estão geralmente desorganizados. Para a maioria das organizações, diferentes definições de dados fundamentais, tais como o cliente ou produto aumentam a confusão. Por exemplo, quando a empresa ”Corporação Canadiana dos Pneus” decidiu criar uma estrutura para os seus dados, descobriu que o seu data warehouse assegurava seis níveis diferentes de inventário. Outros dados não estavam de todo disponíveis, como a comparação dos valores de vendas de determinados produtos nas suas lojas em todo o Canadá. Durante vários anos, a empresa criou um plano para reunir novos dados que fossem de encontro às necessidades analíticas da empresa.
Resumindo e concluindo, podemos enumerar algumas características que potenciam o valor dos dados:

- **Estão correctos**. Embora algumas análises se traduzam em gráficos e outras necessitem de uma precisão à casa decimal, todas elas necessitam que os dados que as alimentam passam os testes de credibilidade dos revisores.

- **Estão completos**. A definição de completo varia consoante o negócio da organização, venda de cimento, cartões de crédito, seguros de saúde, entre outros. No entanto, a definição dos dados completos está dependente da capacidade da organização.

- **Estão actualizados**. Mais uma definição que varia; para algumas questões de negócio, como a emergência médica, os dados devem estar disponíveis instantaneamente para serem transmitidos em tempo-real às ambulâncias e profissionais de saúde. Para outras decisões de negócio, como a previsão de orçamento, apenas é necessária uma actualização periódica – diária, semanal ou mensal.

- **São consistentes**. Para ajudar os decisores, é necessário normalizar os dados e as definições comuns. Desta forma são eliminadas as redundâncias e as hipóteses de inconsistências ou desactualizações dos dados.

- **Estão contextualizados**. Quando os dados são enriquecidos com os metadados (geralmente definidos como a estrutura dos dados sobre os dados), o seu significado deve tornar-se claro.

- **Estão controlados**. De forma a cumprir com os requisitos do negócio, legais e reguladores da segurança, privacidade e auditoria, os dados devem ser rigorosamente supervisionados.

Que regras e processos é que são necessários para gerir os dados desde a sua aquisição até ao seu destino?

Cada etapa do ciclo de vida da gestão de dados apresenta desafios técnicos e de gestão que podem provocar um impacto significativo na capacidade da organização em competir na componente analítica.

- **Obtenção de dados**. Criar ao obter dados é o primeiro passo. Para a informação interna da organização, os profissionais de TI devem trabalhar em conjunto com os gestores de negócio. O objectivo da obtenção de dados é determinar quais os que dados necessários e qual a melhor forma de os integrar nos sistemas de TI, em conformidade com os processos de negócio. Este é o caminho para obter dados de qualidade na fonte.

- **Limpeza dos dados**. Detectar e remover dados desactualizados, incorrectos, incompletos ou redundantes é das tarefas mais importantes, dispendiosas e consumidoras de tempo em qualquer iniciativa de BI. Estima-se que entre 25% a 30% de uma iniciativa de BI vai para a limpeza de dados inicial. O papel das TI consiste em estabelecer métodos e sistemas para adquirir, organizar, processar e manter a informação, no entanto, a limpeza dos dados é da responsabilidade de todos os que geram e usam os dados.

- **Organização e armazenamento dos dados**. Após os dados terem sido adquiridos e limpos, os processos sistemáticos de extracção, integração e sintetização devem ser definidos. Os dados devem ser colocados no formato e no repositório correcto para que estejam prontos a usar.

- **Manutenção dos dados**. Após o repositório estar criado e populado com dados, os gestores devem decidir como e quando é que os dados devem ser actualizados. Estes mesmos gestores devem criar procedimentos para se assegurarem da privacidade dos

dados, segurança e integridade (protecção contra a corrupção, perca por erro humano, vírus e problemas de hardware). As políticas e os processos de manutenção dos dados devem também ser desenvolvidos numa óptica de quando é que os dados necessitam de serem armazenados, arquivados ou retirados. Alguns concorrentes analíticos estimaram que gastam 500.000 dólares em manutenção para 1 milhão gasto em desenvolvimento de novas capacidades técnicas de BI.

Uma vez que a organização abordou todas as questões de gestão de dados, o próximo passo consiste em determinar as tecnologias e os processos necessários para os extrair, transformar e carregar os dados no DW.

9.2.2 Ferramentas e Processos de Transformação

Para que os dados sejam utilizáveis pelos gestores, é preciso em primeiro lugar passarem por um processo conhecido nas TI como ETL, para extracção, transformação e carregamento. As tarefas de extracção dos dados fonte e carregamento num repositório são relativamente simples, no entanto, a limpeza e a sua transformação são uma questão mais complexa.

De forma a que os dados no DW sejam coerentes, é necessário primeiro que tudo limpá-los e validá-los, através de regras de negócio que usam ferramentas de limpeza de dados como o Trillium (por exemplo, uma regra simples, poderá ser o número de telefone que é composto por 9 algarismos). Os procedimentos de transformação definem a lógica de negócio que mapeiam os dados da fonte até ao seu destino. Os gestores de negócio e de TI devem despender um esforço significativo de forma a transformarem os dados em informação utilizável. Embora existam no mercado vários fornecedores com ferramentas para automatizarem o processo de transformação, será sempre necessário um esforço manual.

A transformação de dados, assenta numa padronização das definições dos dados de forma a que os conceitos de negócio se mantenham consistentes e que as suas definições sejam comparáveis em toda a organização. Por exemplo, um cliente pode ser definido como uma empresa num sistema, mas em outro pode ser definido como um indivíduo. Esta questão requer que os gestores decidam o que fazer sobre os dados que faltam. Algumas vezes é possível preencher os campos a branco com dados inferidos, projecções baseadas nos dados disponíveis, ou mantêm-nos a branco e não podem ser usados para análise. Estas tarefas requerem um esforço contínuo visto que novas questões estarão sempre a levantar-se.

9.2.3 Repositórios

As organizações têm diversas opções para organizarem e armazenarem os seus dados analíticos:

- Os DW's são bases de dados que contêm dados integrados de diferentes fontes que são regularmente actualizados. Contêm dados históricos para facilitar as análises de desempenho do negócio ao longo do tempo. Um DW pode ser um módulo de um sistema empresarial ou uma base de dados independente. Algumas organizações possuem uma base de dados designadas por SA que é usada para preparar os dados obtidos de diferentes fontes e carregá-los no DW.

- Um data mart (DM) é um repositório separado ou a uma secção particionada do DW global. Os DM's são geralmente usados para suporte a funções específicas do negócio ou processo, assim como, contêm análises pré-determinadas para que os gestores possam, de forma independente, trabalhar com um subconjunto de dados sem possuírem grandes conhecimentos estatísticos. Inicialmente, algumas organizações não viram a necessidade de para além do data warehouse, criarem também um conjunto independente de data marts ou modelos analíticos com dados directamente da fonte. Por exemplo, uma organização na área dos químicos possui 60 data marts. Hoje em dia, esta abordagem é rara porque cria problemas de manutenção ao departamento de TI. Os data marts devem ser usados se os seus arquitectos estão certos que não será necessário no futuro um conjunto mais amplo de dados para análise.

- O repositório de metadados deve conter informação técnica e definições dos dados, que incluem informação sobre a fonte, como é calculada, informação bibliográfica e a unidade de medida. Pode incluir também informação sobre fiabilidade, precisão e instruções de como é que os dados devem ser aplicados. Um repositório comum de metadados usado por todas as aplicações analíticas é fundamental para assegurar a consistência dos dados. Consolidando toda a informação necessária para limpeza de dados num repositório único, reduz significativamente o tempo necessário para manutenção.

Uma vez que os dados estão organizados e prontos, é tempo de determinar a tecnologia analítica aplicações necessárias.

9.2.4 Ferramentas e Aplicações Analíticas

Escolher as ferramentas/aplicações, ou seja software analítico, depende de vários factores.

A primeira tarefa consiste em determinar como é que o processo de tomada de decisão pode ser embutido no processo de negócio. Deverá existir um ser humano que revê os dados e a componente analítica e toma uma decisão, ou a decisão deve ser automatizada de forma a que seja alguma coisa que aconteça na natureza do fluxo de processo? Se a resposta é a segunda opção, existem tecnologias que estruturam o fluxo de processo e fornecem regras de decisão (qualitativas e quantitativas) para a tomada de decisão.

A próxima tarefa/decisão consiste em usar uma aplicação fornecida por uma organização externa ou criar uma solução à medida. Os fornecedores de sistemas empresariais como a Oracle e a SAP estão a desenvolver cada vez mais (e mais sofisticadas) aplicações analíticas que são incorporadas nos seus produtos. De acordo com o IDC (Corporação Internacional de Dados), os projectos que implementam um pacote de uma aplicação analítica tem um retorno de investimento de 140%, enquanto que um desenvolvimento à medida de ferramentas analíticas tem um retorno de investimento de 104%. Para uma organização, a decisão de "comprar feito ou fazer" assenta se já existe um pacote com a solução ou se existem níveis de competência suficiente para um desenvolvimento à medida.

No entanto, existem diversas ferramentas com potencial para análises de dados que permitem às organizações desenvolver as suas próprias análises (ver capítulo Tecnologias Analíticas). A Business Objects e SAS oferecem produtos integrados de ferramentas e aplicações. Algumas ferramentas são desenhadas para visualizar fatias ou detalhe dos dados, enquanto que existem outras mais sofisticadas na componente estatística. Algumas ferramentas comportam uma grande diversidade de tipos de dados, enquanto outras são mais limitadas (apenas para dados altamente estruturados ou análises textuais). Existem também ferramentas que extrapolam a partir dos dados históricos, enquanto outras procuram novas tendências ou relações.

Quer seja usada uma solução à medida ou uma aplicação comprada no mercado, a área de TI de uma organização deve acomodar uma diversidade de ferramentas para os diferentes tipos de análises de dados (ver capítulo Tecnologias Analíticas). Naturalmente, os colaboradores têm

preferência por produtos familiares, como folhas de cálculo, mesmo não sendo apropriados para realizar determinadas análises. Outro problema consiste na ausência de uma arquitectura global para escolher uma ferramenta, o que pode resultar numa excessiva proliferação tecnológica. Num questionário de 2005, as grandes organizações responderam que em média possuíam 30 ferramentas de BI, de, em média, 3.2 fornecedores. No passado, existia esta necessidade porque diferentes fornecedores cobriam ofertas diferentes – uns mais focados em "reporting" financeiro, outros em consultas ("queries") ad hoc ou análise estatística. Embora exista diversidade entre nos fornecedores, estes já começam a oferecer pacotes de BI cada vez mais integrados.

9.2.5 Tecnologias Analíticas

9.2.5.1 Tecnologias Analíticas Típicas

Os gestores das organizações que estão a planear tornarem-se competitivos na componente analítica devem estar familiarizados com as categorias chave das ferramentas de software analítico:

- As *folhas do cálculo,* do Microsoft Excel, são as ferramentas analíticas mais comuns, porque são fáceis de usar e reflectem os modelos mentais dos utilizadores. Analistas e gestores usam-nas como o último passo da componente analítica antes de os dados serem apresentados num formato de relatório ou gráfico aos decisores. No entanto, muitos utilizadores usam as folhas de cálculo para tarefas que não deviam, o que conduz a erros e conclusões incorrectas. Mesmo usando, de forma correcta, as folhas de cálculo estão sujeitas ao erro humano (mais de 20% das folhas de cálculo têm erros, e 5% têm células calculadas de forma incorrecta). Para minimizar estas falhas, os gestores insistem em começar com dados precisos, validados e com utilizadores que tenham as competências adequadas e experiência para desenvolverem as folhas de cálculo e modelos necessários para a monitorização e ou tomada de decisão.

- Os *processadores analíticos on-line* são geralmente conhecidos pela sua abreviação, OLAP, e são usados para decisões e análises semi-estruturadas. Numa base de dados relacional (RDBMS) os dados são armazenados em tabelas relacionais de forma a tornar eficiente a organização dos dados em sistemas transaccionais. No entanto, não são eficientes para analisar dados em formato de vector (como é o caso da forma, em célula, como os dados estão organizados numa folha de cálculo) ou linhas temporais.

As ferramentas OLAP foram especificamente desenhadas para problemas multidimensionais. Os dados são organizados em cubos (de dados) para permitir análises por tempo, pela geografia, por linhas de produto, entre outras. Os cubos de dados são conjuntos de dados que assentam em três ou mais variáveis e que estão preparados para "reporting" e análises, estes cubos podem ser idealizados como folhas de cálculo multidimensionais. Enquanto que nas folhas de cálculo do Excel existe um número finito de dimensões, os modelos OLAP não têm essa limitação. Como resultado, necessitam de competências especializadas para serem desenvolvidos, ou podem ser criados por "power users" que estejam familiarizados com as suas capacidades. Ao invés das tradicionais folhas de cálculo, as ferramentas OLAP têm de lidar com a proliferação de dados ou rapidamente os modelos se tornam inviáveis. Para consultas ("queries") complexas, as ferramentas OLAP têm a reputação de produzir uma resposta em 0,1% do tempo que seria necessário para ser respondida por um modelo relacional. A Business Objects e a Cognos são fornecedores que estão entre os líderes de mercado nesta categoria.

- Os *algoritmos quantitativos e estatísticos* permitem que os gestores ou profissionais na área da estatística possam analisar os dados de forma sofisticada. Os algoritmos processam dados quantitativos para chegarem ao objectivo ideal como o melhor preço ou uma quantidade exacta. Os algoritmos estatísticos também englobam modelação preditiva, optimização e simulação. Note-se que no inicio dos anos 70, organizações como a SAS e a SPSS introduziram no mercado aplicações informáticas que tornaram a estatística mais acessível a todos.

- Os *motores de regras* processam uma série de regras de negócio que usam instruções condicionais para endereçarem questões lógicas. Por exemplo, se uma pessoa com menos de 25 anos, sem casa própria e sem ser licenciado, pretender efectuar um seguro para uma mota, não será possível. Os motores de regras podem fazer parte de uma aplicação automatizada ou proporcionarem recomendações para quem precise de tipos de decisões muito particulares.

- As *ferramentas de data mining* assentam numa gama que vai desde a aritmética computacional à inteligência artificial, estatística, árvores de decisão, redes neuronais e teoria de Bayes. O seu objectivo consiste em identificar padrões em conjuntos de dados complexos e pouco estruturados. Por exemplo, a empresa Sprint usa a

tecnologia analítica de redes neuronais para prever quais os clientes que não estão dispostos a trocar o seu telefone da rede fixa por um telemóvel.

- As *ferramentas de text mining* podem ajudar os gestores a identificarem de forma rápida, praticamente em tempo real, padrões emergentes. A simples contagem da mesma palavra ou frases em websites é um exemplo simples de *text mining*. A monitorização de blogs técnicos pode ajudar um vendedor a reconhecer, em poucas horas, que um novo produto tem defeito antes de começar a receber reclamações dos clientes. Outros produtos de *text mining* podem reconhecer referências a pessoas, lugares, objectos ou tópicos e usarem esta informação para desenharem inferências acerca do comportamento da concorrência.

- As *ferramentas de simulação* modelam os processos de negócio através de um conjunto de funções matemáticas, cientificas, de engenharia e financeiras. Assim como os sistemas de desenho assistido por computador (CAD) que são usados por engenheiros para modelar e desenhar um novo produto, as ferramentas de simulação são usadas em bastantes áreas como a engenharia, a investigação e desenvolvimento (ID), entre outras. As simulações podem ser usadas como dispositivo de formação de forma a ajudar os utilizadores a compreenderem as implicações da alteração de um processo de negócio. Também podem ser usadas para ajudar a verificar o fluxo de informação ou produtos, como por exemplo, ajudar os funcionários das organizações de saúde a decidir para onde é que devem enviar os órgãos doados de acordo com certos critérios que englobam o tipo de sangue e as limitações geográficas.

9.2.5.2 Tecnologias Analíticas Emergentes

Existem algumas tecnologias analíticas que no futuro próximo irão ter um papel importante nas organizações, a ver:

- A *categorização de texto* é um processo que usa modelos ou regras de estatística para atribuir uma taxa à relevância de um documento para um certo tópico. Esta tecnologia pode ser usada para avaliar de forma dinâmica os produtos dos seus concorrentes nos seus websites.

- *Os algoritmos genéticos* são uma classe estocástica de métodos de optimização que usam os princípios da reprodução genética natural (“crossover” ou mutações de

estruturas de DNA). Uma aplicação comum consiste na optimização de rotas de distribuição de produtos alimentares.

- *Os sistemas inteligentes* ("expert systems") não são uma tecnologia nova, mas têm vindo a amadurecer com a idade. Aplicações especializadas de inteligência artificial são capazes de disponibilizar conhecimento a quem toma decisões.

- O *audio e video mining* são parecidos com as outras ferramentas de *text* ou *data mining*, no entanto, procuram padrões em áudio (som), vídeo ou imagens.

- A *swarm intelligence* esta é utilizada para aumentar o realismo das simulações e para compreender os efeitos dramáticos das alterações de baixo nível num sistema. Esta tecnologia pode ser observada nas sociedades complexas de formigas e abelhas.

- A *extracção de informação* tem como função extrair conceitos como nomes, entidades geográficas e relacionamentos de grandes quantidades de dados (normalmente textuais) não estruturados.

9.2.6 Aplicações e Ferramentas de Apresentação

As análises só são úteis se tiverem utilidade, desta redundância resulta que as organizações que apostam na componente analítica devem capacitar os seus colaboradores para transmitirem os seus conhecimentos a outros colaboradores através de ferramentas de "reporting", "scorecards" ou portais. As ferramentas de apresentação devem permitir que os utilizadores criem um denominado relatório "ad hoc" (no momento) e consigam visualizar, de forma interactiva, dados. Os utilizadores devem estar alertados para a diversidade de ferramentas de comunicação (como o e-mail, PDA's, telemóveis) de forma a partilharem dados.

Os fornecedores de soluções de BI, como a SAP Business Objects, a Cognos da IBM, a SAS, a Oracle Hyperion, a Microsoft e a MicroStrategy oferecem produtos que incluem a apresentação de dados e soluções de "reporting". Os sistemas empresariais tornaram-se mais analíticos, e os fornecedores como a SAP e a Oracle rapidamente incorporaram nos seus produtos essas capacidades. As aplicações analíticas, geralmente, possuem uma interface para ser usada pelos gestores, analistas e estatísticos.

Uma nova geração de ferramentas analíticas visuais desenvolvidas por novos fornecedores como a Spotfire e a Visual Sciences, e de fornecedores tradicionais como a SAS, permitem a manipulação de dados e análises através de uma interface visual muito intuitiva. Um gestor, por exemplo, pode analisar um gráfico de dados, excluir valores de outliers, e executar uma regressão linear que encaixe nos dados sem ter grandes conhecimentos de estatística. Visto que permitem a exploração de dados sem o risco de, acidentalmente, modificarem o modelo base, as ferramentas analíticas visuais aumentam o número de pessoas que podem realizar análises sofisticadas. Por exemplo, o CIO da farmacêutica Vertex estima que apenas 5% dos seus utilizadores podem fazer uso efectivo de ferramentas algorítmicas, mas outros 15% podem manipular a componente analítica visual.

9.2.7 Processos Operacionais

Este elemento da arquitectura de BI responde a questões sobre como é que a organização cria, gere e mantém as aplicações e os dados. Permite detalhar como é que um conjunto padrão de ferramentas e tecnologias são usadas para assegurar fiabilidade, escalabilidade e segurança do ambiente de TI. Normas, politicas e processos devem ser definidos e reforçados por toda a organização.

Questões como privacidade e segurança, assim como a capacidade de arquivar e auditar dados são factores críticos de extrema importância para assegurar a integridade dos dados. Esta preocupação revela-se tanto ao nível do negócio como das TI, porque falhas neste campo podem ter consequências desastrosas (como por exemplo, o roubo de um cartão de crédito de um cliente). Uma consequência da evolução dos requisitos legais e regulamentares pode levar a que os gestores sejam acusados criminalmente de negligência se estes ocultarem os procedimentos, a documentação e a validade dos dados que foram usados para as decisões de negócio (Davenport & Harris, 2007).

9.3 Conceito Técnico da Business Intelligence

Quando falamos de BI, estamos a falar de inteligência de negócio, ou seja, dos meios informacionais que permitem dotar qualquer organização de inteligência, flexibilidade e adaptabilidade face às mudanças que ocorram permanentemente no mercado onde actuam. Embora a BI seja nos dias de hoje alvo de algum mediatismo, o seu conceito foi definido em 1958 e o seu objectivo consiste em ajudar a melhorar os processos de tomada de decisão

(Hans, 1958), ou seja, um sistema de BI pode ser apelidado como um sistema de suporte à decisão (Daniel, 2007).

Os sistemas de BI actuais são interactivos e baseados em estruturas de sistemas e subsistemas de informação que auxiliam os decisores a usarem tecnologia, dados, documentos, conhecimento e modelos analíticos para identificarem, monitorizarem e resolverem problemas. A nova geração de sistemas BI, que inclui como principal componente tecnológica o Data Warehouse (DW), que consiste num grande repositório central de informação, e que apresenta um grande potencial em termos de análise da informação.

Durante a década de 90, a maioria das grandes organizações envolveram-se em projectos de Data Warehousing. O foco desses esforços centrou-se desde a combinação e unificação de múltiplos sistemas fonte (transaccionais) até ao desenvolvimento de ferramentas, mais propriamente interfaces, que permitissem aos utilizadores efectuarem análises e relatórios.

Transversal a qualquer área de análise, a implementação de um sistema de BI (neste caso aplicado à Saúde) deve seguir as seguintes etapas (Figura 2):

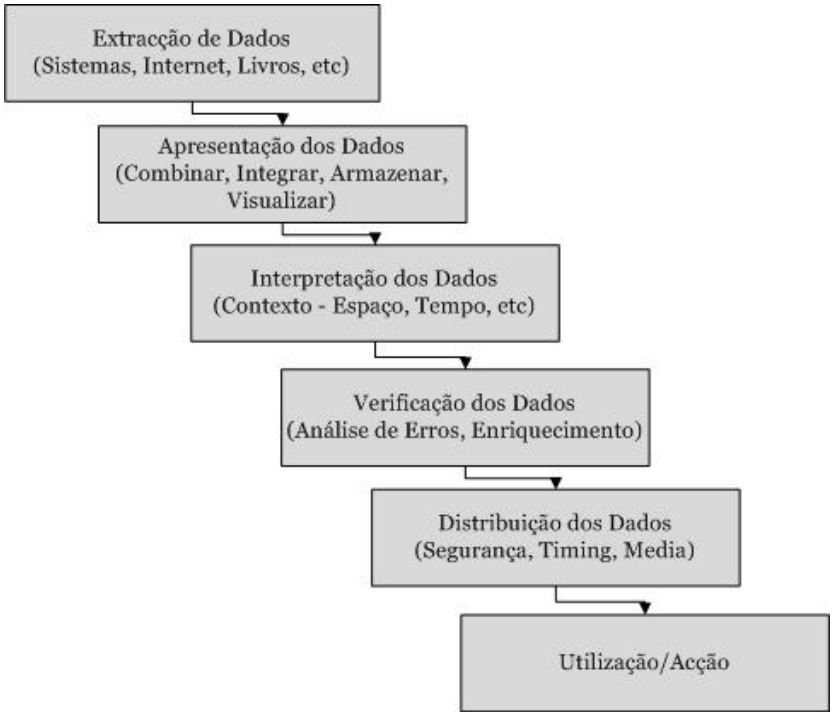


Figura 2 – Etapas de Implementação de um sistema de BI. Adaptado de: http://www.12manage.com/methods_business_intelligence_pt.html

Tipicamente, os sistemas de BI podem ser classificados em dois grandes tipos:

- os baseados em modelos (“model-driven”)

- os baseados em dados (“data-driven”)

Os sistemas baseados em modelos utilizam construções analíticas como a previsão (“forecasting”), algoritmos de optimização, simulações, árvores de decisão e motores de regras. Os sistemas baseados em dados lidam com DW, bases de dados e tecnologia OLAP[10] (On-Line Analytical Processing) que permitem o processamento analítico, assim como, a manipulação e a análise de grandes volumes de dados sob múltiplas perspectivas, ou seja, as denominadas análises multidimensionais da informação.

Um DW é uma base de dados construída de forma a suportar o processo de tomada de decisão numa organização. Os DW podem ser alimentados por diversas fontes de dados (bases de dados ou outras) ou Data Mart’s[11] (Owen, 2006). Uma das principais componentes de um sistema de BI é o DW e toda a sua envolvente. Assim sendo, é necessário começar a analisar de uma forma mais detalhada todo o processo de BI desde a extracção de dados dos sistemas fonte e sua manipulação, até à produção de relatórios por parte dos utilizadores. A Figura 3 ilustra a arquitectura típica de um DW que se encontra em diversas áreas de negócio como a Saúde.

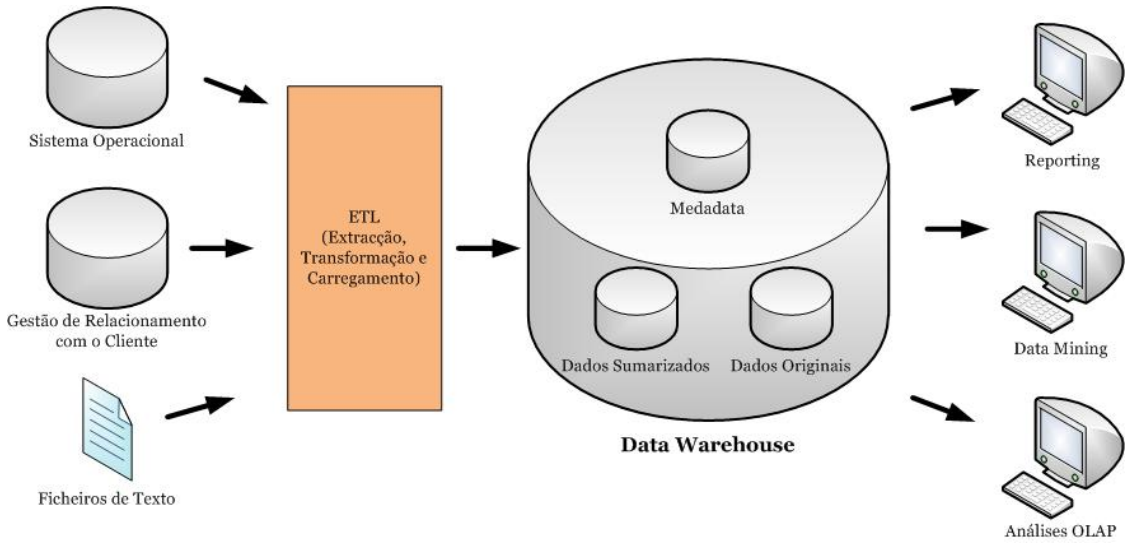


Figura 3 – Arquitectura de um DW. Adaptado de: www.datawarehouse4u.info/

A informação é extraída dos sistemas de fonte, transformada e carregada no DW através de processos ETL (Extracção, Transformação e Carregamento) ficando posteriormente disponível

[10] OLAP é a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas. As aplicações OLAP são usadas pelos gestores em qualquer nível da organização para lhes permitir análises comparativas que facilitem a sua tomada de decisões diária.

[11] Um data mart define-se como uma base de dados de pequena dimensão desenhada especificamente para um departamento (de marketing, financeiro, ou outro) de uma empresa.

para análise através de uma "reporting" que neste caso a que foi adoptada foi o MicroStrategy[12].

9.4 Modelação de um Data Warehouse

O desenho de um DW, que consiste na componente principal de um sistema de BI, assenta numa metodologia de modelação designada por modelo em estrela ("star schema"). Este modelo foi criado por Ralph Kimball, ao propor uma visão para a modelação de base de dados para sistemas de apoio a decisão. A sua principal característica é a presença de dados altamente redundantes, melhorando o desempenho. O modelo ou esquema em estrela é uma metodologia de modelação de dados utilizada para o desenho de um DW.

Os dados são modelados através de tabelas de dimensão ligadas a uma tabela de factos. As tabelas de dimensão contêm as características (descritivos) dos dados, enquanto que a tabela de factos contem as métricas (dados mensuráveis) e as "surrogate keys"[13] (SK) correspondentes às tabelas de dimensão.

O nome do modelo foi adoptado devido à semelhança com uma estrela. No centro da estrela encontra-se a tabela de factos e em ser redor encontram-se as tabelas de dimensão. A relação entre as tabelas é de 1 para N (1:N), ou seja, um registo numa tabela de dimensão origina N registos na tabela de factos (Figura 4).

[12] A MicroStrategy é uma empresa líder mundial na tecnologia de business intelligence, o software da MicroStrategy proporciona a criação de relatórios, análises e monitorização integrados, auxiliando as empresas a tomar melhores decisões para o seu negócio no quotidiano (MicroStrategy, 2010).

[13] O conceito de surrogate keys pertence surgiu com o DW no sentido de criar uma camada de abstracção entre as chaves operacionais (dos sistemas fonte) e as chaves internas do DW. As chaves internas do DW, designadas por SK's, são numéricas no sentido de aumentar a performance de cruzamento da informação entre as tabelas de dimensão e factos, ou seja, existem chaves operacionais que combinam dados numéricos com caracteres o que cria uma grande défice de performance quando se realizam consultas aos dados, logo para ultrapassar esta questão surgem as SK's que são chaves numéricas visto que o cruzamento entre chaves numéricas apresenta muito mais performance. Por estes motivos surgiu a criação de SK's visto que a volumetria de registos de um DW é elevada e é necessário bastante perforamnce na resolução das queries (consultas) aos dados.

Figura 4 – Exemplo de um modelo em estrela. Fonte: (Cincinnati Children's Hospital Medical Center, 2009).

Desta forma, as consultas iniciam-se com as tabelas de dimensão e posteriormente com a tabela de factos, assegurando a precisão dos dados através de uma estrutura de SK's onde não é necessário percorrer todas as tabelas, garantindo assim, um acesso aos dados mais eficiente e com melhor desempenho.

Das grandes vantagens deste modelo destacam-se a sua fácil compreensão para quem não é profissional das tecnologias de informação, visto que o número de ligações entre tabelas é reduzido.

Mais tarde, surge então um outro modelo designado por floco de neve ("snow-flake") que tem como base o modelo em estrela, e consiste nas dimensões que podem ser associadas a novas dimensões através de uma ligação entre elas, ficando emparelhadas. Este modelo é combinado com o modelo em estrela de forma a reduzir a redundância. A Figura 5 ilustra o exemplo relacionado com a geografia em que os dados poderiam estar todos concentrados numa tabela, no entanto, aplicando o conceito do modelo floco de neve, temos que um distrito tem N concelhos e que um concelho tem N freguesias.

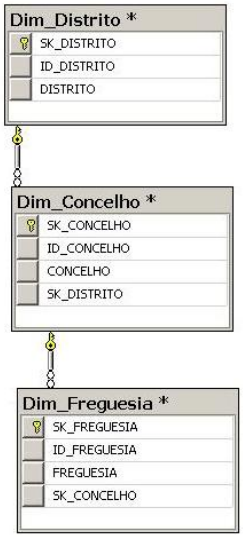


Figura 5 – Exemplo de um modelo em floco de neve para a dimensão localização.

A combinação do modelo em floco de neve com o modelo em estrela resulta na seguinte configuração (Figura 6):

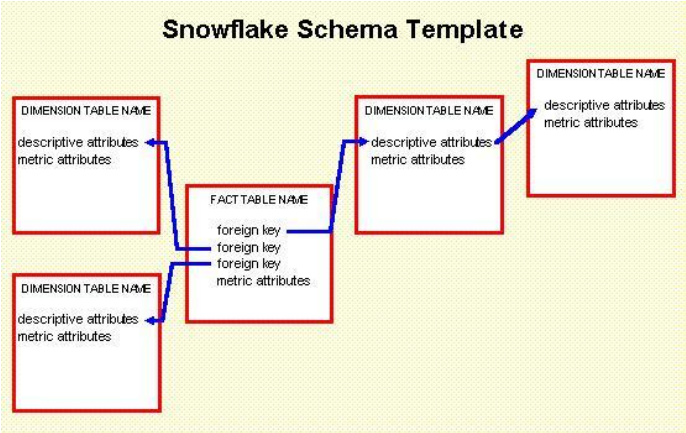


Figura 6 – Exemplo de um modelo em estrela combinado com o modelo floco de neve. Fonte: (Borysowich, 2007)

9.5 Slowly Changing Dimensions

As tabelas de dimensão caracterizam os dados, visto que são constituídas essencialmente por descritivos, como por exemplo, o nome completo de um paciente, o descritivo do diagnóstico, entre outros. Embora os atributos de uma tabela de dimensão são praticamente estáticos, não são fixo ao longo do tempo, por exemplo, o nome de um pessoa pode mudar após o

matrimónio e este tipo de mudanças raras têm de ser tidas em conta. Ou seja, os atributos de uma dimensão, embora de forma lenta, mudam ao longo do tempo. Assim sendo, os arquitectos de um modelo multidimensional têm de envolver, desde o início do seu desenho, os representantes do negócio para que, de forma pró-activa, ajudem a determinar a melhor estratégia de actualização das dimensões ao longo do tempo. Não se pode concluir que o negócio não se importa com as mudanças nas dimensões, mesmo que os representantes do negócio não o tenham mencionado durante o processo de levantamento de requisitos.

No caso da modelação dos GDH's, preservou-se a independência da estrutura dimensional com ajustes relativos à mudança do seu conteúdo. A estas dimensões, quase constantes, são designadas como "slowly changing dimensions". Foi Ralph Kimball o primeiro a introduzir em 1994 o conceito de "slowly changing dimensions" e desde então, os profissionais de IT colocaram SCDs como acrónimo do conceito inventado.

Para cada atributo das tabelas de dimensão, é necessário especificar a estratégia para assegurar a mudança. Por outras palavras, quando o valor de um atributo muda no mundo operacional, como é que essa mudança será assegurada no modelo dimensional? Existem 3 tipos de técnicas, que podem ser usadas para lidarem com as alterações de atributos.

9.5.1 Tipo 1: Sobreposição do Valor

Com a técnica do tipo 1, o valor antigo do atributo de cada linha da dimensão é sobreposto pelo valor corrente. Neste caso, o atributo reflecte sempre a atribuição mais recente.

Vamos assumir um exemplo de uma organização na área do retalho de equipamentos médicos electrónicos. Os compradores estão alinhados da mesma forma departamental (como por exemplo o departamento de fisioterapia, análises, cirurgia, radiologia, entre outros) que a organização de equipamentos médicos electrónicos, assim sendo, os produtos a serem adquiridos apresentam-se como departamentos. Um dos produtos procurados consiste numa "Máquina de raio X". A linha existente na dimensão produto para a "Máquina de raio X" é a seguinte (Figura 7):

Chave do Produto (SK)	Descrição do Produto	Departamento	Unidade de Stock (Chave Operacional)
12345	Máquina de raio X	Análises	ABC922-Z

Figura 7 – Exemplo de um registo original.

Logicamente que na dimensão produto poderia possuir mais atributos adicionais, no entanto, foram abreviados neste exemplo visto que não eram necessários. A chave primária é a SK em vez da unidade de stock (chave operacional ou natural). O único atributo inviolável do produto, ao contrário dos outros, consiste na unidade de stock. No estudo dos três tipos de técnicas SCD, assume-se que a chave natural da dimensão mantém-se constante.

A título de exemplo, suponha-se que, um responsável pelos stocks decide que a "Máquina de raio X" deve ser mudada do departamento de análises para o departamento de radiologia a 25 de Junho de 2010, no sentido de aumentar as vendas. Com a resposta do tipo 1, a linha existente da tabela de dimensão iria ser actualizada com o descritivo do novo departamento. O resultado da linha actualizada seria o seguinte (Figura 8):

Chave do Produto (SK)	Descrição do Produto	Departamento	Unidade de Stock (Chave Operacional)
12345	Máquina de raio X	Radiologia	ABC922-Z

Figura 8 – Exemplo de um registo após aplicação do tipo 1.

Neste caso, as chaves das tabelas de factos e dimensão não foram alteradas quando a "Máquina de raio X" mudou de departamento. As linhas na tabela de factos continuam a fazer referência à chave do produto 12345, independentemente, do departamento da "Máquina de raio X". Do momento em que o departamento de vendas definiu como estratégia para melhorar a performance, mover o produto de departamento, logo não se irá ter histórico, nem dados recentes para avaliar se foi ou não uma decisão acertada, visto que o produto em causa aparecerá (como se tivesse desde sempre) no departamento de Radiologia.

A vantagem do tipo 1 consiste na rapidez, ou seja, nas tabelas de dimensão, o valor pré-existente é sobreposto pelo valor corrente e as tabelas de factos mantêm-se sem qualquer tipo de alteração. O único problema consiste na manutenção do histórico de alterações dos atributos, ou seja, ao existir sobreposição de valores apenas são armazenados no DW os valores recentes visto que sobrepõem os anteriores. A resposta do tipo 1 é apropriada para alterações de atributos como correcções, assim como, se não houver valor em manter a descrição anterior. Para tomar a decisão de adoptar o tipo 1 é necessário ter o "input" do negócio para determinar se realmente é necessário (ou não) guardar histórico de alterações de atributos. Esta decisão nunca deverá ser tomada apenas pelo departamento de TI na medida em que pode não de ir encontro às necessidades do negócio.

9.5.2 Tipo 2: Adicionar uma linha na Dimensão

Uma resposta do tipo 2 já permite armazenar histórico. No momento que a "Máquina de raio X" muda de departamento é acrescentada uma nova linha na dimensão produto que reflecte o novo valor do atributo departamento. O resultado consiste em duas linhas na dimensão produto para a "Máquina de raio X" (Figura 9):

Chave do Produto (SK)	Descrição do Produto	Departamento	Unidade de Stock (Chave Operacional)
12345	Máquina de raio X	Análises	ABC922-Z
26789	Máquina de raio X	Radiologia	ABC922-Z

Figura 9 – Exemplo de um registo após aplicação do tipo 2.

Como podemos observar e compreender o porquê da chave operacional/natural unidade de stock não ser a chave primária da dimensão produto. São necessárias duas SK's para a mesma unidade de stock. Cada SK identifica unicamente um atributo do produto que pertencia à realidade do negócio durante um período de tempo. Com as alterações do tipo 2, a tabela de factos mantém-se inalterada. Não se percorre as linhas de histórico das tabelas de factos para alterar a chave do produto (SK), o que acontece é que as linhas da "Máquina de raio X" na tabela de factos antes de 15 de Fevereiro de 2010 fazem referência à chave de produto 12345 quando o produto é analisado sob a perspectiva do departamento (que neste caso é o de Análises). Após 15 de Fevereiro, as linhas da "Máquina de raio X" na tabela de factos têm como chave do produto 26789 que reflecte a mudança para o departamento de Radiologia até que surja outra alteração do tipo 2.

Se a restrição for apenas o atributo departamento, é perfeitamente preciso a diferenciação entre ambos os perfis dos produtos, no entanto, se a restrição consistir apenas na descrição do produto ("Máquina de raio X"), a consulta ("query") automaticamente retorna duas linhas da dimensão produto para o produto em questão e automaticamente liga-se ("join") à tabela de factos para o histórico completo de produtos. Neste cenário ter-se-á a sensação de registos duplicados nas tabelas. Por exemplo, no caso de se pretender contar, de forma correcta, o número de produtos será necessário usar a chave operacional/natural unidade de stock como base da contagem distinta em vez da SK. O campo da chave operacional/natural torna-se num campo fiável que assegura a separação dos registos do tipo 2 para um produto. Como alternativa, um indicador de linha recente poderá ser um atributo da dimensão bastante útil

para permitir aos utilizadores restringir de forma rápida as consultas (“queries”) aos perfis correntes.

Certamente que seria natural incluir uma data efectiva (“date stamp”) na linha da dimensão do tipo 2. A data efectiva refere-se ao momento em que os valores dos atributos se tornam válidos ou inválidos no caso de datas de expiração. Efectivamente, datas efectivas e de expiração são necessárias na SA para se determinar qual a SK que é válida quando se está a carregar dados nas tabelas de factos. Na tabela de dimensão, estas datas efectivas são extras extremamente úteis e não necessitam de particionamento de histórico.

Enquanto que a inclusão das datas efectivas e de expiração pode ser uma tarefa confortável para os arquitectos de base de dados, é necessário ter em atenção que a data efectiva está pouco relacionada com as datas na tabela de factos. Na tentativa de restringir na linha da dimensão a data efectiva pode levar a resultados incorrectos. Talvez se uma nova versão 2.0 do produto (“Máquina de raio X”) sair em Maio de 2010, uma nova chave operacional/natural (e respectiva SK) deve ser criada para este produto. Este cenário não é uma alteração do tipo 2, visto que o produto consiste numa nova entidade física. No entanto, se olharmos para a tabela de factos do revendedor, não é possível observar um particionamento abrupto do histórico. A versão antiga do produto continuará a ser vendida após o dia 1 de Maio de 2010, até que o inventário esteja esgotado. A nova versão 2.0 irá aparecer no dia 1 de Maio e gradualmente irá superar a versão antiga. Existirá um período de transição em que ambas as versões irão estar disponíveis. Ao se efectuar o registo da venda será reconhecido ambos os códigos de unidade de stock e não haverá dificuldade em assegurar as vendas de cada versão. Tendo uma data efectiva na linha da dimensão produto, não será possível usar esta data para restringir e particionar as vendas, visto que a data não tem relevância. No pior cenário, usando esta restrição pode levar a respostas erradas.

No entanto, as datas efectivas/expiração na dimensão podem ser úteis para análises mais avançadas. As datas permitem obter fatias de tempo precisas da dimensão por si só. A linha de data efectiva é a primeira data para a qual o perfil da descrição é válido. A linha de data de expiração terá um dia a menos da linha de data efectiva para a próxima atribuição, ou a data de quando o produto foi retirado do catálogo. Pode-se determinar como estaria o catálogo do produto em Dezembro de 2009 efectuando uma restrição à consulta (“query”) da tabela de produtos para devolver todas as linhas com data efectiva menor ou igual a 31 de Dezembro de 2009 e com data de expiração maior ou igual do que 31 de Dezembro de 2009.

A resposta do tipo 2 é uma técnica para sustentar análises utilizando atributos historicamente precisos. Esta resposta, segmenta perfeitamente o histórico da tabela de factos, porque as linhas de factos pré-alteradas possuem a SK pré-alterada. Outro aspecto consiste em detectar todas as alterações na dimensão.

Com uma resposta do tipo 2, é criada uma nova linha na dimensão com uma nova chave primária (SK) de forma a identificar unicamente o novo perfil de produto. Esta chave primária estabelece a ligação entre as tabelas de dimensão e de factos para um conjunto de características do produto. Não é totalmente necessário criar ligações ("joins") secundárias baseadas em datas efectivas e de expiração, assim como foi mencionado acima.

9.5.3 Tipo 3: Adicionar uma coluna na Dimensão

Enquanto que a resposta do tipo 2 particiona histórico, não permite associar o valor do novo atributo com o histórico antigo de factos e vice-versa. Com a resposta do tipo 2, quando se restringe em Departamento = "Radiologia", não se observa factos antes de 15 de Fevereiro de 2010. Em muitos dos casos, é esta informação que se pretende obter.

No entanto, ocasionalmente, poderá ser necessário observar dados de factos como se uma mudança nunca tivesse ocorrido. Esta perspectiva ocorre frequentemente na reorganização da força de vendas de uma organização. Embora tivessem ocorrido mudanças, alguns utilizadores pretendem ter a possibilidade de verificar as vendas no quotidiano em comparação com as vendas anteriores, no período homólogo, segundo a estrutura organizacional anterior. Por exemplo, para os meses de transição pode existir a necessidade de analisar o histórico dos novos nomes dos distritos em comparação com os nomes dos distritos anteriores. Uma resposta do tipo 2 não suporta este requisito, mas a do tipo 3 responde a este tipo de questões.

No exemplo da máquina de raio X, vamos assumir que é importante para as necessidades do negócio analisar ambos os valores (anterior e recente) do atributo departamento nos períodos anteriores e posteriores após a mudança de valor. Com uma resposta do tipo 2, não é adicionada mais uma linha na dimensão, mas sim uma nova coluna para capturar a alteração do atributo. No caso da "Máquina de raio X", a dimensão produto foi alterada para contemplar o atributo departamento anterior. Esta nova coluna é populada com o valor anterior do departamento (Análises). O departamento atributo é tratado da mesma forma do tipo 1 em que é sobreposto o valor para reflectir o departamento corrente (Radiologia). Todos os relatórios existentes e consultas ("queries") ficam a apontar automaticamente para a nova

descrição do departamento, visto que o nome da coluna não se alterou, apenas foi alterado o valor. Contudo, se os utilizadores do negócio pretenderem efectuar relatórios e consultas sobre o valor do departamento anterior (antigo) podem fazê-lo bastando apenas apontar as consultas para o atributo departamento anterior (Figura 10).

Chave do Produto (SK)	Descrição do Produto	Departamento	Departamento Anterior	Unidade de Stock (Chave Operacional)
12345	Máquina de raio X	Radiologia	Análises	ABC922-Z

Figura 10 – Exemplo de um registo após aplicação do tipo 3.

A reposta do tipo 3 é apropriada quanto existe a necessidade de suportar, simultaneamente, duas visões do negócio. Alguns arquitectos de base de dados apelidam o tipo 3 de realidade alternativa. Este cenário ocorre quando o atributo é uma etiqueta atribuída pelos humanos em vez de uma característica física. Embora a mudança tenha ocorrido, é logicamente possível agir como se não tivesse ocorrido. A resposta do tipo 3 diferencia-se do tipo 2 na medida em que dá resposta simultaneamente à descrição anterior e corrente ao mesmo tempo. No caso da reorganização das vendas, a gestão pode pretender sobrepor e analisar os resultados usando um mapa da organização de vendas durante um período de tempo.

A reposta do tipo 3 é raramente usada. Ao pensar-se que o número mais elevado do tipo de "slowly changing dimension" está associado ao tipo mais utilizado é um erro. Cada técnica tem a sua aplicação prática consoante as necessidades do negócio.

9.6 Índices

Um índice de uma base de dados é uma estrutura de dados que melhora a performance das operações de retorno de dados a um custo reduzido de escrita, mas com um aumento do espaço em disco. Os índices, podem ser criados usando uma ou mais colunas de uma tabela para proporcionarem "look-ups" (procura de valores/dados) de acesso eficiente e rápido. O espaço em disco necessário para armazenar um índice é, tipicamente, inferior a uma tabela (visto que os índices contêm apenas campos chave de acordo com a organização da tabela, excluindo todos os outros detalhes da tabela), permitindo assim, armazenar índices na

memória para uma tabela que exceda a capacidade de armazenamento em memória (Wikipedia, 2010).

No SQL Server (plataforma adoptada na presente tese para desenvolvimento do ETL e gestão do DW), os índices são organizados como B-trees (Microsoft, 2010) que são o principal tipo de índices usados nas base de dados, e são a base para os índices "clustered" e "nonclustered".

9.6.1 B-tree

Numa árvore, os registos são armazenados em folhas. O ponto de partida é a raiz. O número máximo de filhos por designa-se por "order" (ordem) da árvore. O número máximo de operações de acesso necessárias para atingir a folha desejada (dados armazenados na folha) designa-se por "depth" (nível). Na Figura 11 encontra-se a comparação entre uma árvore natural e uma B-tree. A figura seguinte (Figura 12) ilustra a estrutura de uma árvore binária que consiste na base do índice B-tree.

Árvore Natural	**B-tree**
Cresce de baixo para cima	Cresce de cima para baixo
Tronco principal	Raiz
Ramo	Nó
Folha	Folha

Figura 11 – Comparação entre árvore natural e binária.

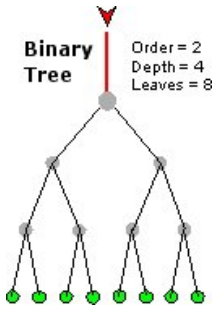


Figura 12 – Estrutura de uma árvore binária.

Quanto maior for a ordem, mais folhas e nós podem ser colocados a uma certa profundidade, o que significa que existem menos níveis para atravessar até à folha que contem os dados pretendidos. No exemplo da figura 12, o número de operações até a uma folha é igual ao nível.

Maior parte dos índices são demasiado grandes para a memória, o que significa que necessitam de ser armazenados em disco. Visto que a operação de I/O (Input/Output) nos sistemas de computadores apresenta um grande custo, os índices necessitam de ser armazenados de forma eficiente no I/O (Fleming, 2010). O índice B-tree é adequado para este cenário, nomeadamente, se forem construídos nós com o mesmo tamanho que um bloco físico de I/O, será necessário um I/O para ocorrer uma deslocação, em profundidade, ao nível da árvore. O exemplo da figura 13 representa um índice criado com base no primeiro nome de um campo.

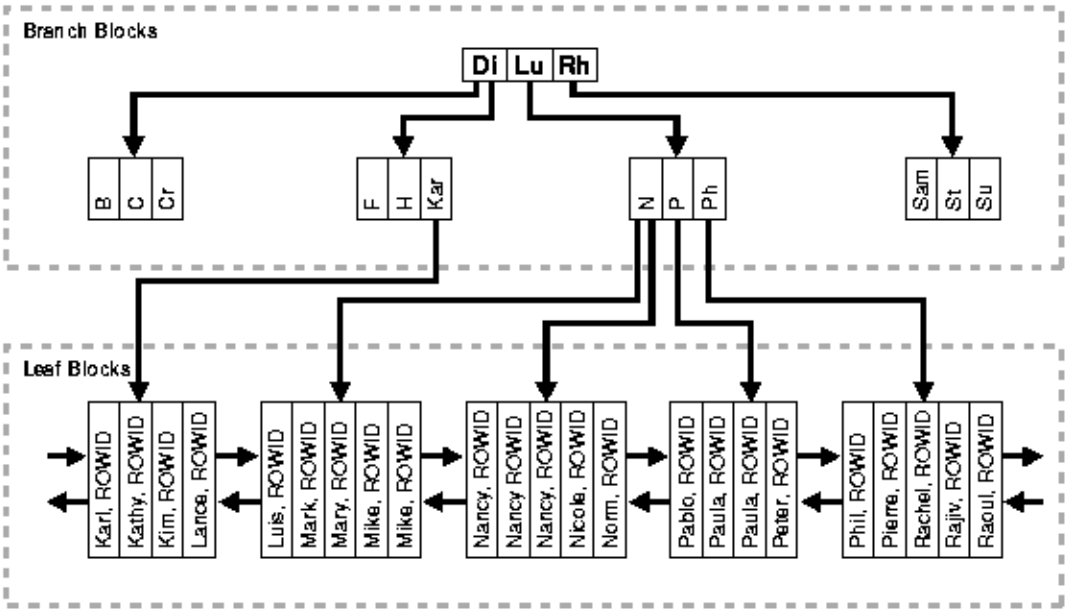


Figura 13 – Índice de base de dados num formato B-tree.

Se em cada nível existir um I/O, serão necessários três I/O para encontrar o nome “Mary” (ou outra folha).

Este é o conceito base dos índices de base de dados.

9.6.2 Clustered

Um “clustered” é um índice que determina a ordem física dos dados numa tabela. Ou seja, um índice “clustered” é análogo a uma lista telefónica, em que nos são apresentados os dados (números de telefone) ordenados pelo último nome. Visto que o índice “clustered” dita a ordem do armazenamento físico dos dados numa tabela, limita qualquer tabela a conter apenas um índice “clustered” (por cada tabela). No entanto, o índice permite múltiplas colunas

de uma tabela (índice composto), assim como a lista telefónica pode estar organizada pelo primeiro e último nome.

O índice "clustered" é particularmente eficiente em colunas que são usadas em procuras de intervalos de valores. Após, ser encontrada a linha com o primeiro valor, usando o índice "clustered", as linhas com os valores indexados subsequentes, estão garantidamente e fisicamente posicionados de uma forma adjacente. Por exemplo, se uma aplicação executa uma determinada consulta ("query") para retornar registos num intervalo de datas, um índice "clustered" pode rapidamente localizar as linhas que contêm a data de início e devolver as linhas adjacentes, na tabela, até atingir a data de fim. Este mecanismo pode ajudar bastante no aumento de performance de uma consulta ("query") deste tipo. Por outro lado, se existir uma coluna que é usada frequentemente para ordenar dados de uma tabela, é bastante vantajoso criar um índice "clustered" nessa coluna que implicitamente faz uma ordenação física dos dados da tabela por essa coluna, poupando assim, o custo de ordenação cada vez que essa coluna é consultada ("queried").

Do ponto de vista mais técnico, num índice "clustered" os nós folha contêm as páginas de dados da tabela subjacente. A raiz e os nós de nível intermédio contêm as páginas dos índices que encapsulam as linhas dos índices. Cada linha de índice contém um valor chave e um apontador para uma página de nível intermédio numa B-tree, ou para linha de dados no nível folha do índice. As páginas de cada nível do índice estão relacionadas numa lista duplamente ligada (Microsoft, 2010).

9.6.3 Nonclustered

Os índices "nonclustered" possuem a mesma estrutura que os B-tree, assim como os índices "clustered", excepto para seguintes diferenças:

- As linhas de dados da tabela subjacente não estão ordenadas e armazenadas com base nas chaves "nonclustered".

- O nível folha de um índice "nonclustered" é construído através de páginas de índices em vez de páginas de dados.

Os índices "nonclustered" podem ser definidos numa tabela em conjunto com um índice "clustered". Cada linha de índice, no índice "nonclustered", contém o valor da chave

"nonclustered" e um apontador de linha. Este apontador aponta para a linha, ou seja, o apontador é construído através do identificador do ficheiro, do número de página e do número da linha na página. Este tipo de apontador é designado como "Row ID", ou seja, identificador de linha.

9.6.4 Clustered vs. Nonclustered

De forma sucinta, as principais diferenças entre índices "clustered" e "nonclustered" resumem-se a (figura 14):

Índice Clustered Natural	Índice Nonclustered
Apresenta-se num formato de linhas e colunas	Apresenta-se num formato de relatório sobre as tabelas
Existe ao nível físico	Não são criados ao nível físico, mas sim ao nível lógico
Ordena os dados ao nível físico	Não ordena os dados ao nível físico
Funciona para a tabela completa	Uma tabela possui 255 índices "nonclustered"
Apenas existe uma tabela como no formato de dados armazenados	A tabelas tem bastantes índices "nonclustered"
Uma tabela contem apenas um índice "clustered"	Funciona em ordem dos dados

Figura 14 – Comparação entre índices "clustered" e "nonclustered".

9.7 Ferramentas de Business Intelligence

No mercado actual existem diversas ferramentas para a prática de BI. Em geral, esta funcionalidade faz parte do portfólio das ferramentas que efectuam integração de dados. Num recente relatório do Gartner Group sobre ferramentas de integração de dados, a IBM e a Informatica são classificadas como as organizações actualmente líderes do mercado de ferramentas de integração de dados. As duas organizações são classificadas como tendo maior número de clientes e maior flexibilidade nos seus produtos, nomeadamente o Informatica

Power Center e os produtos da família IBM Websphere. Seguindo de perto estas duas organizações líderes encontramos a SAP Business Objects e a Oracle. Estas últimas são classificadas como adversárias devido à falta de flexibilidade das suas implementações ou seja, demasiada especificidade a um determinado ambiente dos produtos oferecidos. A Microsoft tem evoluído bastante sobretudo com o software SQL Server 2008, em que o foco foi a performance do tratamento de dados. Por outro lado, a MicroStrategy é um "player" mundial visto que desde 1989 disponibiliza no mercado soluções de BI (apenas componente de "reporting") flexíveis, "user-friendly" e apelativas do ponto de vista gráfico. Por último, das organizações com o menor (ou mais fraco) posicionamento no mercado nesta área podem destacar-se a Sybase e a SAS que ainda se encontram numa fase bastante prematura de oferta de produtos desta natureza.

Em resumo, o posicionamento de algumas ferramentas de BI segundo um estudo da Gartner é o seguinte (Figura 15):

Figura 15 - Estado de artes das ferramentas de BI. Fonte: (Gartner, 2010)

Neste enquadramento, a ferramenta adoptada para todo o ETL foi o SQL Server 2008 Integration Services que permite uma grande flexibilidade de desenvolvimento e integração

com outras plataformas, nomeadamente o MicroStrategy que foi o software adoptado para a visualização dos dados e para a construção de reports e "dashboards". Contudo, a plataforma de "reporting" da MicroStrategy designada por "MicroStrategy Reporting Suite - Free Edition" que possui a limitação de usar apenas um CPU na resolução de reports (relatórios), prejudicando assim a performance das consultas aos dados. Esta limitação deve-se ao facto de a licença usada ser gratuita. A versão licenciada não possui esta limitação, no entanto, tem custos de licenciamento que para o desenvolvimento desta tese de mestrado não faria sentido suportar. A principal diferença entre ambas versões reside no número de CPU's que são usados para resolverem os relatórios, contudo, ao nível das funcionalidades, que é o mais importante para esta tese de mestrado, não existe qualquer diferença. A escolha da ferramenta de "reporting" da MicroStrategy deveu-se, principalmente, ao facto de esta organização ser um gigante na área de BI (componente de "reporting"), assim como, ao potencial da ferramenta em termos de flexibilidade, usabilidade e integração com a componente de ETL.

10 Business Intelligence na Saúde

Na área da saúde, o processamento de dados clínicos ainda se apresenta de uma forma antiquada (Hughes, 2004). Posteriormente, surgiram os CDW (Clinical Data Warehouse) e os sistemas de BI aplicados aos hospitais. Os CDW estão mais direccionados para as organizações de biomédica/farmacêutica, em que o seu objectivo consiste em reduzir custos e o ROI (Return of Investment).

As equipas de investigação e desenvolvimento pretendem com os CDW analisar informação clínica, como por exemplo, o historial patológico dos pacientes a quem administram determinados tipos de medicamentos de forma a compreenderem quais os mais eficientes, bem como que substâncias é que foram usadas, e se podem ser combinadas com outros medicamentos. Uma possível estrutura do CDW, para responder a estas questões, encontra-se na Figura 16.

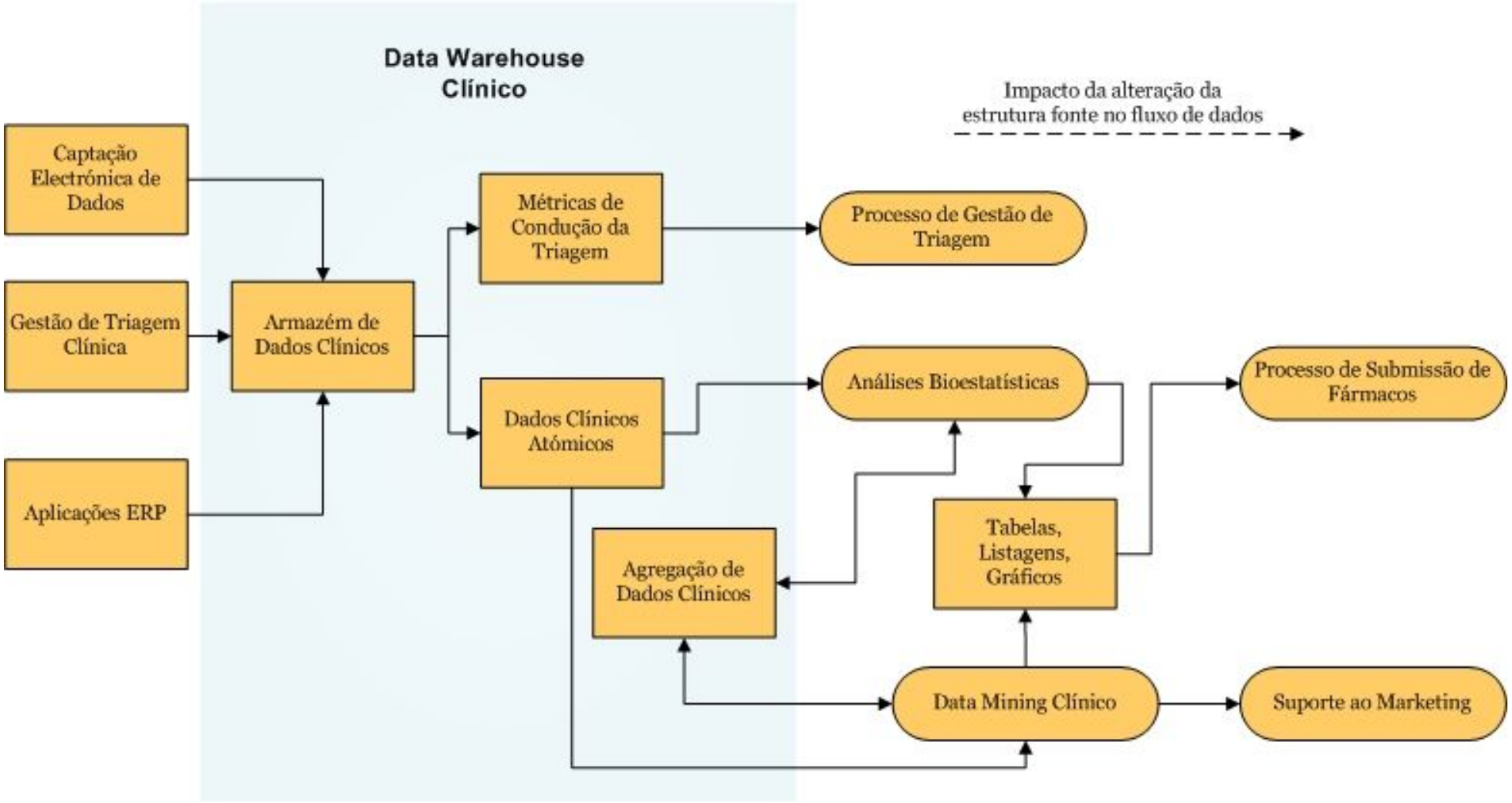


Figura 16 – Arquitectura de um CDW. Adaptado de: http://www.information-management.com/issues/20041101/1012400-1.html

Como se pode observar, para além dos objectivos referidos, este CDW permite ainda auxiliar o marketing na promoção de novos medicamentos.

Os sistemas de BI aplicados aos hospitais podem ter diversas finalidades, como a facturação e a gestão dos cuidados de saúde dos pacientes. Do ponto de vista financeiro, permitem a gestão da facturação dos pacientes e estão vocacionados para proporcionarem informação de optimização financeira, gestão de custos e coordenação dos fornecedores de seguros de saúde.

Tipicamente, os sistemas de BI aplicados à saúde contêm informação demográfica sobre o paciente, o tempo de internamento, taxas, custos e o detalhe por cada serviço facturado como cirurgia, exames radiológicos, tratamentos, etc. Outra vantagem destes sistemas de BI, consiste na optimização de esforços administrativos relacionados com a gestão de informação individual de cada paciente, tais como a sua localização, a medicação prescrita, entre outras.

Visto que em cada hospital a informação clínica e financeira se encontra dispersa em diversos sistemas transaccionais, que têm como função registar as ocorrências e que não têm capacidade de traduzir a quantidade de dados armazenada em qualidade, surgem os sistemas de BI que permitem tratar esses dados de forma a proporcionar conhecimento (qualidade de informação), criando um repositório central que suporta uma visão unificada dos dados clínicos e uma disponibilidade dos mesmos para análise, que vão de encontro às necessidades de informação dos utilizadores clínicos.

Os sistemas de BI, neste caso aplicados à saúde, podem ser implementados, como por exemplo, a gestão de camas nos hospitais. Esta gestão, que consiste num sistema de BI, cujo objectivo se baseia em optimizar ao máximo a utilização de camas. Neste caso, a monitorização da disponibilidade das camas é efectuada através de indicadores e "dashboards" que permitem identificar o número de camas disponíveis em tempo real, o que resulta na redução dos tempos de espera nas urgências, na previsão dos prazos de disponibilidade de camas, numa melhor gestão do período de indisponibilidade das camas, assim como numa melhor eficiência das transferências de pacientes para outras unidades de internamento ou mesmo para outro hospital. A título de exemplo, podemos referir a StatCom, como sendo uma organização que presta serviços informáticos na área da saúde, e, possui na sua carteira de produtos e serviços este tipo de sistema. A Figura 17 ilustra o tipo de análises e monitorização que é efectuada.

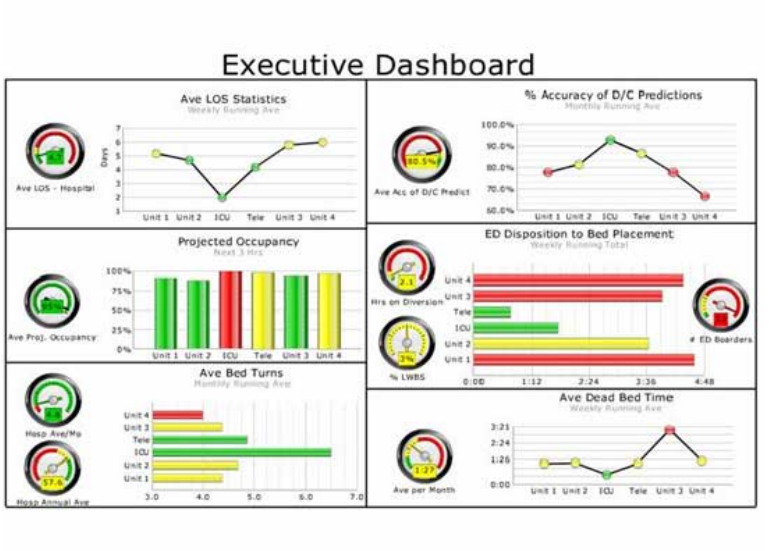


Figura 17 – Interface do sistema de gestão de camas. Fonte: http://www.statcom.com/healthcare-software-solutions/business-activity-monitoring.aspx

Existem muitos casos de sucesso de sistemas de BI aplicados à saúde e aos hospitais. Como exemplo, em Portugal, o Hospital Militar Principal implementou um sistema de BI para suprimir a falta de uniformização da informação em relação à sua inconsistência e flexibilidade, às dificuldades de responder as solicitações externas, à inexistência de um ambiente dedicado e a uma inadequada informação de gestão (Cunha, 2008). O sistema implementado teve como objectivo armazenar o histórico de consultas externas, movimento de internamentos, intervenções cirúrgicas, meios complementares diagnóstico e terapêutica e facturação. O acesso aos dados é efectuado através de relatórios elaborados pelos utilizadores.

Outro caso de sucesso, na implementação de um sistema de BI encontra-se no Hospital Samaritano, no Brasil. Esta instituição necessitava de uma solução que integrasse dados e fornecesse informações concretas sobre os primeiros socorros, pacientes internados e externos, assim como permitisse analisar, em termos de histórico, os diversos procedimentos cirúrgicos (custos das cirurgias). Nenhum software que o hospital possui permitia fornecer esta informação, logo os resultados obtidos com a implementação deste projecto de BI foram um enorme sucesso. A possibilidade de efectuar o cruzamento de dados e observar/visualizar o resultado desse cruzamento de informação trouxe benefícios para área financeira e para outros departamentos do hospital. Segundo Sérgio Lopez Bento, director do hospital, antes da implementação do sistema de BI, a área de controlo não conseguia analisar os resultados por serviços, especialidade médica ou procedimento. Actualmente, estas informações são disponibilizadas pelo sistema de BI e têm como origem a compilação de diversos relatórios (Meta Análise, 2008).

Existem outros exemplos de sucesso na implementação de sistemas de BI na Saúde, como o caso dos hospitais de Vancouver, no Canadá, e o de BayCare na Flórida, nos Estados Unidos da América (Himmelsbach, 2005).

Os benefícios destas implementações traduzem-se na unificação da informação (visão única da realidade), flexibilidade de análise, sofisticação e abrangência das análises e relatórios. Por outro lado, a nível de custos e em termos hospitalares, oferece a vantagem de reduções dos custos e melhor gestão de internamentos, cirurgias e transferências de pacientes para outras unidades de saúde.

10.1 Modelação de Data Warehouses na Saúde

A área da saúde apresenta características próprias que a tornam um caso interessante de desenho de DW. No trabalho de um dos principais autores na área do DW (Kimball & Ross, 2002), é apresentado uma revisão técnica da literatura que reflecte as melhores práticas de modelação de DW aplicados à saúde, e com base nesta revisão, apresenta-se uma proposta de revisão de um modelo utilizado como a base de partida para a construção de outros modelos mais detalhados, segundo as especificações próprias de cada área de negócio. Um outro modelo que será analisado mais à frente, será dos GDH.

Importa salientar que antes de elaborar qualquer modelo de dados, é necessário compreender toda a envolvente da área em análise, que neste caso é a saúde. O círculo de valor da saúde (Figura 18) no seu todo abrange uma rede de clínicas, hospitais, farmácias, laboratórios, companhias de seguros e o Governo, mais propriamente o Ministério da Saúde.

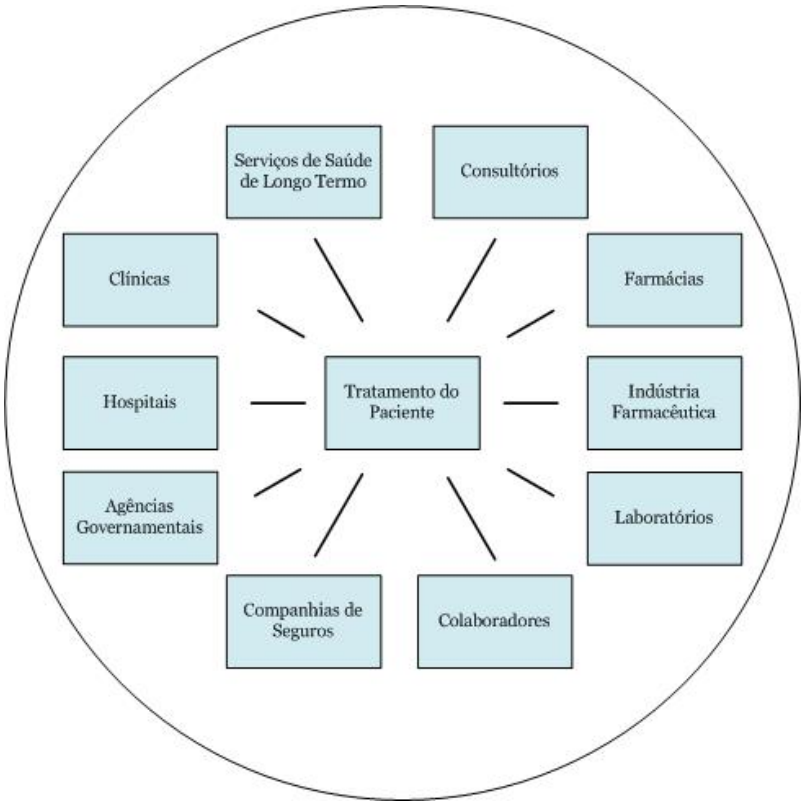


Figura 18 – Círculo de valor da saúde. Adaptado de: (Kimball & Ross, 2002).

O que se pode observar e concluir do círculo de valor da saúde é que qualquer tipo de organização de saúde tem como factor crítico a organização e uniformização dos registos de tratamentos dos pacientes. Existem dois tipos principais de registos para o tratamento de pacientes: os registos de facturação de tratamentos (que correspondem ao número de linhas na factura que o hospital apresenta ao paciente para liquidar); os registos médicos representam testes de laboratórios, descobertas e informações sobre o percurso de tratamento.

10.2 Exemplo de um Modelo de Dados aplicado à Gestão de Despesas

As principais dimensões (tabelas que descrevem os factos) que podem estar incluídas num modelo do DW aplicado à saúde são as seguintes:

- Colaborador Clínico (Médico, Enfermeiro).
- Entidade Pagadora (Paciente ou terceiros).
- Fornecedor de Tratamentos (Hospitais ou Clínicas).

- Tipos de Tratamento.
- Tipos de Medicamentos.
- Diagnósticos.
- Localização (do Hospital ou Clínica).

Após definidas as dimensões, é importante não esquecer os objectivos em termos de negócio do DW aplicado à saúde, ou seja, pretende-se analisar o número de facturas por cada dimensão já descrita (por paciente, diagnóstico, tratamento, data e outras combinações). Pretende-se também analisar a percentagem de facturas pagas e não pagas, quanto tempo é que demoram a ser pagas, qual o status corrente das facturas que ainda não foram pagas, etc. Toda esta informação necessita de ser actualizada em intervalos de 24 horas (Kimball & Ross, 2002).

Uma das vantagens do DW é a disponibilização de grandes volumes de informação consolidada numa janela de tempo bastante reduzida e que tem vindo a ser cada vez mais próxima do tempo real).

Colocado este objectivo e após identificação das dimensões do DW, é necessário construir uma tabela de factos que irá conter as métricas de facturação que serão cruzadas com as dimensões.

A tabela de factos irá representar o histórico de cada linha de factura, logo necessita da seguinte estrutura (Figura 19):

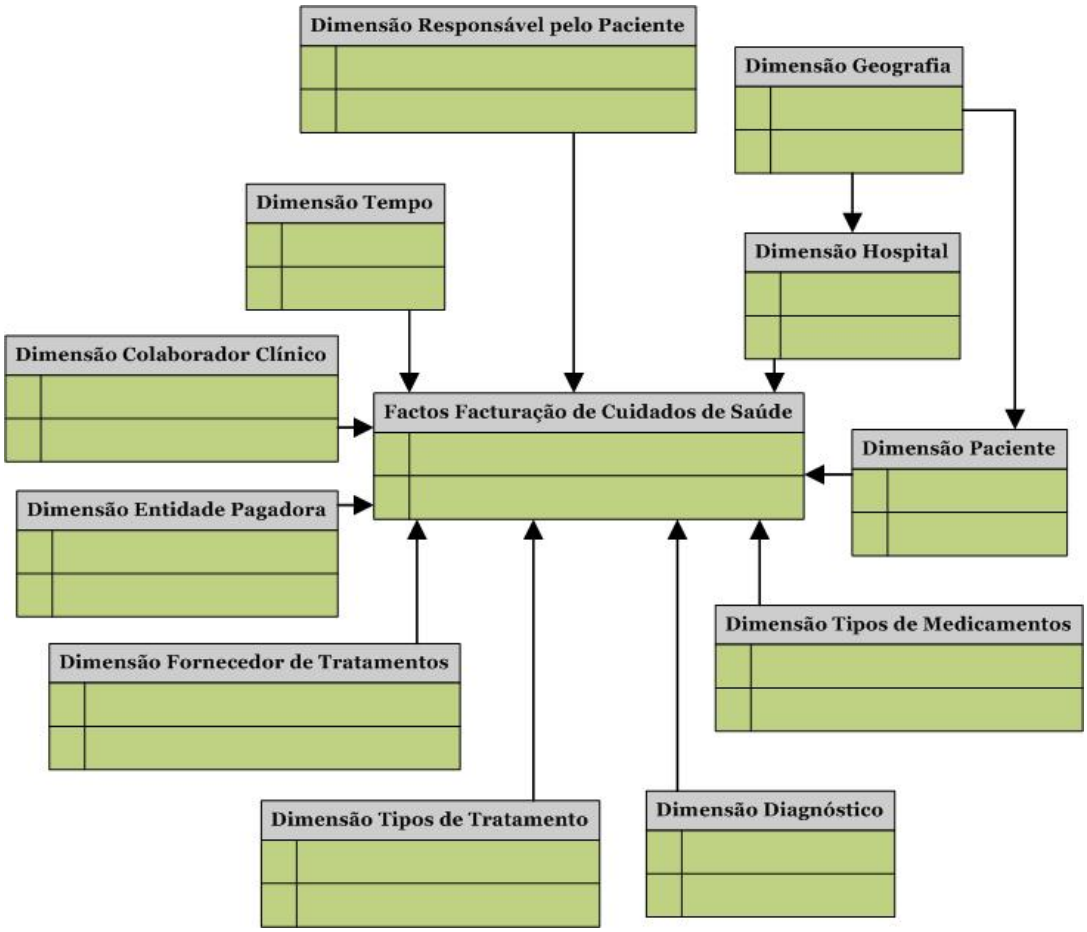


Figura 19 – Tabela de factos para a facturação de cuidados de saúde. Adaptado de: (Kimball & Ross, 2002).

O conjunto das dimensões e das tabelas de factos forma o modelo de dados que permite análises multidimensionais de acordo com os objectivos pretendidos, ou seja, este tipo de modelo permite responder a todas as análises propostas.

10.3 Exemplo de um Modelo de Dados aplicado aos Registos Médicos

Os registos médicos são um desafio para a modelação de um DW devido à sua extrema variedade. Uma ficha de paciente possui dados clínicos em diversos formatos, como por exemplo, dados numéricos (de exames médicos), comentários, notas escritas pelos profissionais de saúde (principalmente médicos), gráficos, fotografias e radiografias. Dada esta extensa variedade, não é esperado efectuar, simultaneamente, "queries" e "reports" que permitam analisar cada tipo de dados. No entanto, um possível modelo de DW pretende criar um "standard"/"framework" para todos os registos de cada paciente. Assumindo que a granularidade dos dados pode ser definida por um registo por paciente, é assim possível contemplar a maioria dos registos médicos numa única tabela de factos. Contudo, esta

primeira abordagem tem como desvantagens o grande número de campos (colunas) para cada linha (que representa um paciente), visto que os registos médicos podem assumir diversos tipos. Outra desvantagem, prende-se com a inflexibilidade de alterações à tabela de factos na medida em que caso surja um novo campo (atributo ou métrica) implicaria alterar fisicamente a tabela para adicioná-lo. Assim sendo, surge uma segunda abordagem ao nível da modelação de dados, que consiste em construir uma tabela de factos contendo uma "fact dimension", ou seja, uma "dimensão de factos" designada por tipo de entrada, ver (Figura 20), onde é descrito o que as linhas da tabela de factos significam, ou por outras palavras, o que os factos representam. A dimensão tipo de entrada determina quais dos 4 tipos de campos dos factos (valor "flag", comentário e nome do ficheiro JPEG) são válidos para a entrada específica, e como devem ser interpretados cada campo. Por exemplo, a coluna valor genérico é usada para cada valor numérico de entrada, e encontra-se anexada à linha de dimensão tipo de entrada. Se a entrada for uma "flag" (por exemplo Sim/Não ou Elevado/Médio/Reduzido), os valores correspondentes podem ser encontrados na dimensão tipo de entrada. Se a entrada é um comentário (texto livre) ou um elemento multimédia (como uma imagem/gráfico/fotografia JPEG), a dimensão tipo de entrada alerta a aplicação que requisita a informação (ferramenta de "reporting") para se focar nestes campos da tabela de factos.

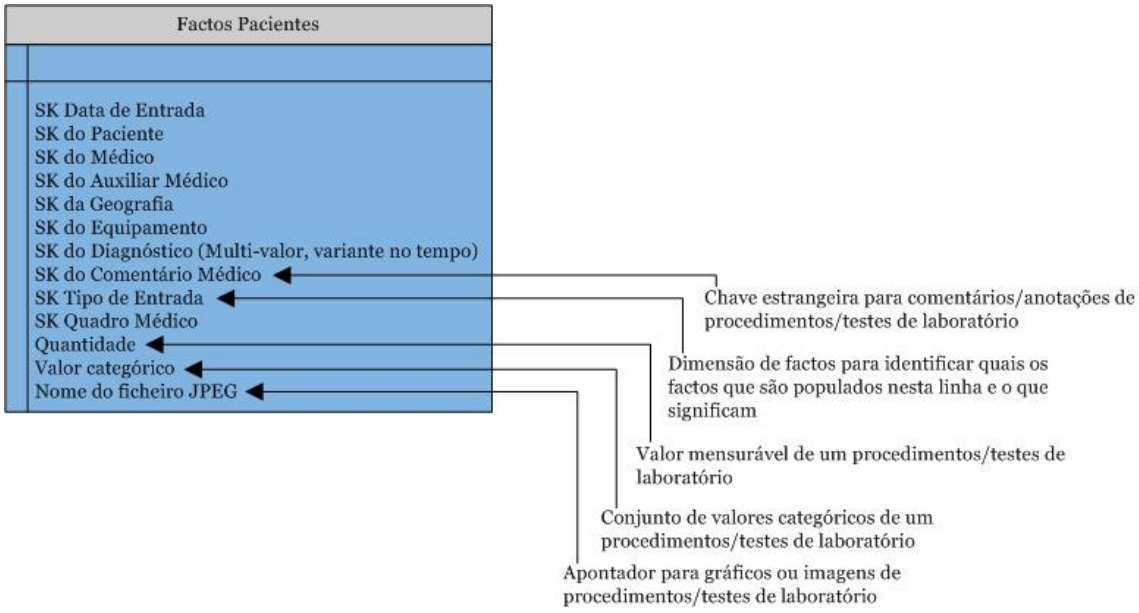


Figura 20 – Tabela de factos para armazenamento de registos médicos. Adaptado de: (Kimball & Ross, 2002).

Esta aproximação é extremamente flexível. Podem, simplesmente, ser adicionados novos tipos de métricas (medidas) através da inserção de novas linhas na "dimensão de factos" sem alterar a estrutura da tabela de factos. No entanto, esta abordagem implica algum "trade-off", na medida em que a utilização de uma "dimensão de factos" pode originar um grande número de

linhas (registos), ou seja, se um evento resultar em 10 valores de medida, na tabela de factos existirão 10 linhas na vez de uma, como da primeira abordagem (desenho clássico).

Para ambientes em que a informação é dispersa, como é o caso do clínico/laboratório, esta abordagem é bastante razoável. No entanto, à medida que a densidade dos factos cresce, o número de linhas na tabela de factos irá aumentar de forma bastante abrupta, o que pode levar ao regresso da primeira abordagem (tabela de factos clássica).

As restantes dimensões apresentadas na Figura 20 são auto-explicativas. A dimensão quadro médico consiste em agrupar com a mesma chave ("Quadro Médico") um número de registos que pertençam a um determinado quadro.

Os comentários/anotações (texto livre) não devem ser armazenados directamente na tabela de factos, visto que ocupam bastante espaço e raramente são usados nas consultas ("queries") efectuadas aos dados. Assim sendo, a tabela de factos deve conter uma chave estrangeira que aponta para uma dimensão "Comentário", como se pode observar na Figura 20.

A referência a uma imagem (nome do ficheiro JPEG) evita a inserção da imagem propriamente dita na base de dados. A vantagem de armazenar apenas o nome do ficheiro JPEG consiste na redução de espaço na base de dados visto que existe software gratuito para visualizar/criar/editar as imagens. A única desvantagem consiste em manter o sincronismo entre a base de dados de ficheiros gráficos (que se define como o servidor onde são guardados os ficheiros JPEG físicos) e a tabela de factos.

11 Grupos de Diagnósticos Homogéneos (GDH´s)

11.1 Enquadramento

A forma como cada país, através dos seus sistemas de saúde, define o sistema de financiamento, seja na obtenção dos seus recursos, seja na distribuição desses mesmos recursos, influencia fortemente o comportamento dos diferentes actores e instituições que participam no sistema: hospitais, médicos, enfermeiros, doentes, gestores, farmacêuticos ou seguradoras (Escoval, 1997)(Vertrees, Incentivos globais e competição nos serviços, 1998), transformando-se mesmo em autênticos "motores" da sua performance (Bentes, O financiamento dos hospitais, 1998). O seu domínio de actuação poderá ser estimulado ou constrangido pela estrutura do sistema e pelos seus incentivos, resultando as suas reacções na prossecução de um dos objectivos *major* dos sistemas de saúde e das instituições que dele fazem parte: os ganhos em saúde (Escoval, 1997).

Sabendo que a estrutura dos sistemas de financiamento mais comuns envolve duas componentes básicas, a definição da quantidade produzida e os preços que valorizam essa produção, deve-se reconhecer que a capacidade de criar incentivos por parte de um sistema de financiamento resulta, em grande parte, do sistema de preços que constitui essa forma de remuneração. De acordo com (Costa, Financiamento de serviços de saúde : a definição de preços, 1990), as organizações de saúde fazem mesmo depender a sua reestruturação produtiva e as suas decisões estratégicas do sistema de preços vigente. Uma vez que o actual sistema de preços pretende representar de uma forma fiel os custos médios por produto hospitalar, e num contexto em que os prestadores de cuidados de saúde não detêm a capacidade de determinar os seus próprios preços, existe a possibilidade de se assistir a uma prática tendenciosa decisional no sentido de tentar maximizar a eficiência técnica, pois essa situação é sinónimo de um maior montante de financiamento, que poderá perigosamente conduzir a casos de especialização ou de troca de produtos. Na verdade, ao serem utilizados os GDH´s como base do pagamento da actividade produtiva hospitalar, os preços relativos estabelecidos pelo Ministério da Saúde terão implicações importantes ao nível do financiamento se os custos médios por GDH´s diferirem sistematicamente dos preços estabelecidos e, se estes estão expostos a uma variação arbitrária e aleatória ao longo dos anos, aumenta o risco financeiro dos contraentes que fixam os preços nesta base. Assim sendo, emerge a lacuna do acompanhamento da evolução dos GDH´s.

11.2 Conceito

O sistema de classificação de doentes mais popular e aplicado a nível internacional é o sistema de Grupos de Diagnósticos Homogéneos (Casas, 1991), (Vertrees, Using DRGs for contracting in Romania, 1998b). Para além dos GDH´s, de entre os mais divulgados sistemas de classificação de doentes, são de salientar o CSI (Computerized Severity Index), o PMC (Patient Management Categories), o DS (Disease Staging) ou o AIM (Acuity Index Method)[14]. Estes sistemas apresentam diferenças quanto à sua definição, âmbito de aplicação, momento e escala de medição e quanto ao seu desempenho e possuem um grau de adequação diferente ao nível da análise de utilização de recursos, da revisão individual de casos ou da previsão do risco de morte (Pinto, 2002).

Os Diagnosis Related Groups (DRGs), sistema que actualmente vigora em Portugal, ficou conhecido por Grupos de Diagnósticos Homogéneos e podem definir-se como um sistema de classificação de doentes[15] internados em hospitais de agudos, em grupos clinicamente coerentes e homogéneos do ponto de vista do consumo de recursos, construídos a partir das características diagnósticas e dos perfis terapêuticos dos doentes[16], que explicam o seu consumo de recursos no hospital (Bentes, A utilização dos GDHs como instrumento de financiamento hospitalar, 1996). Os DRGs foram originalmente idealizados e operacionalizados nos EUA, em finais da década de 60, com objectivos relativamente afastados daqueles que hoje norteiam a sua utilização. Inerentes à sua criação na Yale University, estiveram motivações correspondentes às necessidades de revisão de utilização e de avaliação qualitativa dos cuidados de saúde em hospitais de agudos (Willems, 1989). Ao longo dos anos, outras gerações de DRGs foram desenvolvidas, dando lugar a novas versões comerciais. São elas os Medicare DRGs, os Refined DRGs (RDRGs), os All Patient DRGs (AP-DRGs), os Severity DRGs (SDRGs), os All Patient Refined DRGs (APR-DRGs) e os International-Refined DRGs (IR-DRGs)[17]. Desde 1 de Setembro de 1983, data a partir da qual os DRGs passaram a ser utilizados como base do sistema de pagamento prospectivo do internamento hospitalar para a Medicare (EUA), o interesse internacional aumentou, derivado principalmente da sua aparente capacidade de suster os custos hospitalares (Thorpe, 1987). Neste sentido, para além dos EUA, actualmente outros países utilizam os DRGs nos seus sistemas de saúde, também com versões modificadas ou revistas. Dos 16 países europeus que actualmente os utilizam, entre os quais

[14] Os estudos de (Charbonneau, 1988) ou (Iezzoni, 1989) comparam alguns destes sistemas de classificação de doentes.

[15] Um sistema de classificação de doentes é aquele em que os objectos que se pretendem agrupar são doentes, ou episódios de doença, e em que o objectivo é tornar compreensíveis as suas semelhanças e diferenças e permitir que os que pertençam à mesma classe sejam tratados de forma semelhante (Bentes, A utilização dos GDHs como instrumento de financiamento hospitalar, 1996).

[16] Características identificadas como diferenciadoras ao nível do consumo de recursos, como o sexo, a idade ou o destino após a alta.

[17] O estudo de (Averill, 1998) possui um maior desenvolvimento desta matéria.

estão a Espanha, a Dinamarca, Itália, Bélgica, França, Inglaterra, França, Noruega, País de Gales, Suécia ou a Irlanda (Vertrees, Using DRGs for contracting in Romania, 1998b), foi Portugal o país pioneiro na implementação dos GDH´s como mecanismo de financiamento dos hospitais públicos e de controlo de gestão dos hospitais (Dismuke, 1996). Fora da Europa destaca-se o grande desenvolvimento de novas versões de DRGs no Canadá e na Austrália. A Alemanha e o Japão encontram-se a avaliar a adequação dos DRGs às suas circunstâncias particulares (Vertrees, Using DRGs for contracting in Romania, 1998b).

11.3 Desenvolvimento em Portugal

O processo de implementação dos GDH´s em Portugal começou em 1984, quando foi estabelecido um acordo entre o Ministério da Saúde e a Universidade de Yale, donde resultou um projecto de trabalho liderado pelo Prof. Fetter (principal responsável pelo desenvolvimento dos DRGs) que conteve subjacentes os seguintes objectivos:

- Testar a possibilidade técnica de formar GDH´s a partir da informação contida nos resumos de alta hospitalar, bem como a sua consistência técnica;

- Desenvolver um sistema (operacional) de informação e de custeio por GDH´s.

Os resultados deste projecto foram encorajadores, de tal forma que em 1987 foram iniciados os estudos referentes ao processo de utilização dos GDH´s, como base de pagamento dos hospitais do Serviço Nacional de Saúde (SNS), e, em Janeiro de 1989 iniciou-se um período de transição para a sua implementação (Bentes, A utilização dos GDHs como instrumento de financiamento hospitalar, 1996). A sua operacionalização foi conseguida em 1990. Apesar de a intenção inicial de todo este processo se ter baseado num sistema de pagamento pela produção, para todos os sectores hospitalares relacionados com tratamento de doentes, foi dada primazia ao internamento na sua dupla vertente de facturação a terceiros pagadores e de pagamentos de serviços no âmbito do SNS (Bentes, Formas de pagamento de serviços hospitalares: resumo da comunicação, 1997). Até ao momento, foram utilizadas três das versões produzidas pela Health Care Financing Administration (HCFA) dos Medicare norte-americanos a 6ª, a 10ª e a 16ª[18], respectivamente introduzidas nos anos de 1990, 1996 e

[18] A HCFA 16 foi a versão usada no desenvolvimento da presente tese.

2001[19]. Embora em todos os anos se tenha procedido a reajustamentos nas versões norte-americanas, sobretudo por motivos comerciais, Portugal apenas alterou a versão utilizada quando se verificaram modificações significativas. Ainda que seja manifesta a importância que esta nova forma de financiamento da produção do internamento hospitalar detém na prossecução dos objectivos macro do sistema de saúde português, (Lima, 2000) refere a surpreendente falta de estudos publicados em Portugal acerca destas matérias, que certamente contribuiriam para um melhor conhecimento e esclarecimento do processo de implementação dos GDH´s. Apenas são identificados os estudos de (Dismuke, 1996) e de (Costa, Os DRGs e a gestão do hospital, 1994). Salienta-se este último por ser o único, apesar de ser não empírico, que questiona a validade dos GDH´s ao nível da sua efectividade e adequação.

11.4 Implementação Prática

Neste capítulo são detalhados ao pormenor todos os passos efectuados para a implementação prática do sistema de BI aplicado à Saúde, mais concretamente, aos GDH´s.

11.4.1Work-Flow de Implementação

Todo o desenvolvimento da presente tese pode traduzir-se numa sequência de passos que traduzem todas as etapas ultrapassadas. O primeiro passo, consistiu na análise dos dados fornecidos pela ACSS, que incluiu o estudo da sua estrutura, a forma como estão organizados, o seu tipo e o seu significado. Seguidamente passou-se para a modelação do DW que consistiu numa primeira fase em definir:

- o número de dimensões do modelo,
- quais os campos/dados fonte que iriam corresponder a cada dimensão,
- quais os campos/dados que iriam fazer parte dos factos (métricas).

A fase seguinte da modelação consistiu em desenhar o modelo do DW propriamente dito, que resultou na combinação de duas abordagens clássicas que são o modelo em estrela combinado com o modelo “snow-flake”. Ainda na fase de desenho e após fechado o modelo do DW, desenharam-se os processos ETL em que se definiram as estratégias de extracção dos dados fonte, quais as transformações necessárias a efectuar e a forma de carregamento nas tabelas finais (dimensões e factos).

[19] Nos EUA estas versões foram implementadas em 1988, 1992 e 1996.

Após o desenho e a construção do modelo do DW, assim como a implementação e afinação, em termos de performance, dos processos de ETL, iniciou-se a última etapa de um sistema de BI que consiste na componente de "reporting". Começou-se por construir um modelo lógico de "reporting" com hierarquias tendo como base o modelo físico do DW e seguidamente desenhou-se e implementou-se um conjunto de relatórios e "dashboards" dinâmicos. Contudo, também foi disponibilizado a opção de ser o utilizador a construir o seu próprio relatório. A Figura 21 retrata as etapas deste desenvolvimento em termos de work-flow.

Figura 21– Work-Flow de implementação.

11.4.2Primeira Fase – Análise dos Requisitos

Os dados foram gentilmente cedidos pela ACSS (em formato de CD-ROM) e dizem respeito aos GDH´s relativos ao ano de 2008. No sentido de reforçar e relembrar o conceito de GDH explicitado no capítulo sobre os GDH's, define-se em linhas gerais como sendo um sistema de classificação de doentes internados, em grupos clinicamente coerentes e similares do ponto de vista do consumo de recursos. O agrupamento tem por base as características dos doentes que recebem conjuntos similares de cuidados. A partir deste conceito resumido de GDH´s, iniciou-se a análise de conteúdo do CD-ROM que consistia em quatro ficheiros (Figura 22) com a extensão *.dbf que significa "Data Base File"(Maurer, 2000).

Name	Size	Packed	Type	Modified
..			Folder	
rc1_2008.DBF	1.057.898.970	33.122.534	DBF Viewer 2000 Document	06-05-2009 12:36
rc2_2008.DBF	484.381.580	16.210.898	DBF Viewer 2000 Document	06-05-2009 12:45
rc3_2008.DBF	1.049.206.310	32.011.516	DBF Viewer 2000 Document	06-05-2009 13:05
rc4_2008.DBF	101.433.700	3.125.705	DBF Viewer 2000 Document	06-05-2009 13:16
rc5_2008.DBF	103.367.720	3.084.851	DBF Viewer 2000 Document	06-05-2009 13:16

Figura 22 – Ficheiros de dados disponibilizados pela ACSS.

É nestes ficheiros que se encontra (em relação ao ano de 2008) a classificação de pacientes de todos os hospitais nacionais em GDH´s.

Para analisar-se o conteúdo/estrutura dos ficheiros (e para ter maior poder de manipulação do que no Microsoft Excel) criou-se uma base de dados (BD) temporária, designada “teste” na plataforma Microsoft SQL Server 2008, que foi a adoptada para o desenvolvimento das base de dados e respectivos processos de ETL. Assim sendo, através do Microsoft Management Studio (ferramenta que está incluída no pacote Microsoft SQL Server 2008) criou-se a base de dados “teste” e uma tabela “tbl_teste” como destino de um dos ficheiros disponibilizados.

A base de dados “teste” foi criada inicialmente com o intuito de extrair apenas um dos ficheiros fonte fornecidos, por forma a compreender a sua estrutura dos dados e metadados. Só após uma análise cuidada dos dados fonte é que é possível passar para uma fase de implementação.

No MS, começou-se por analisar o número e tipo de colunas do ficheiro importado para uma tabela teste (“tbl_teste”) da base de dados teste (Figura 23 e Figura 24).

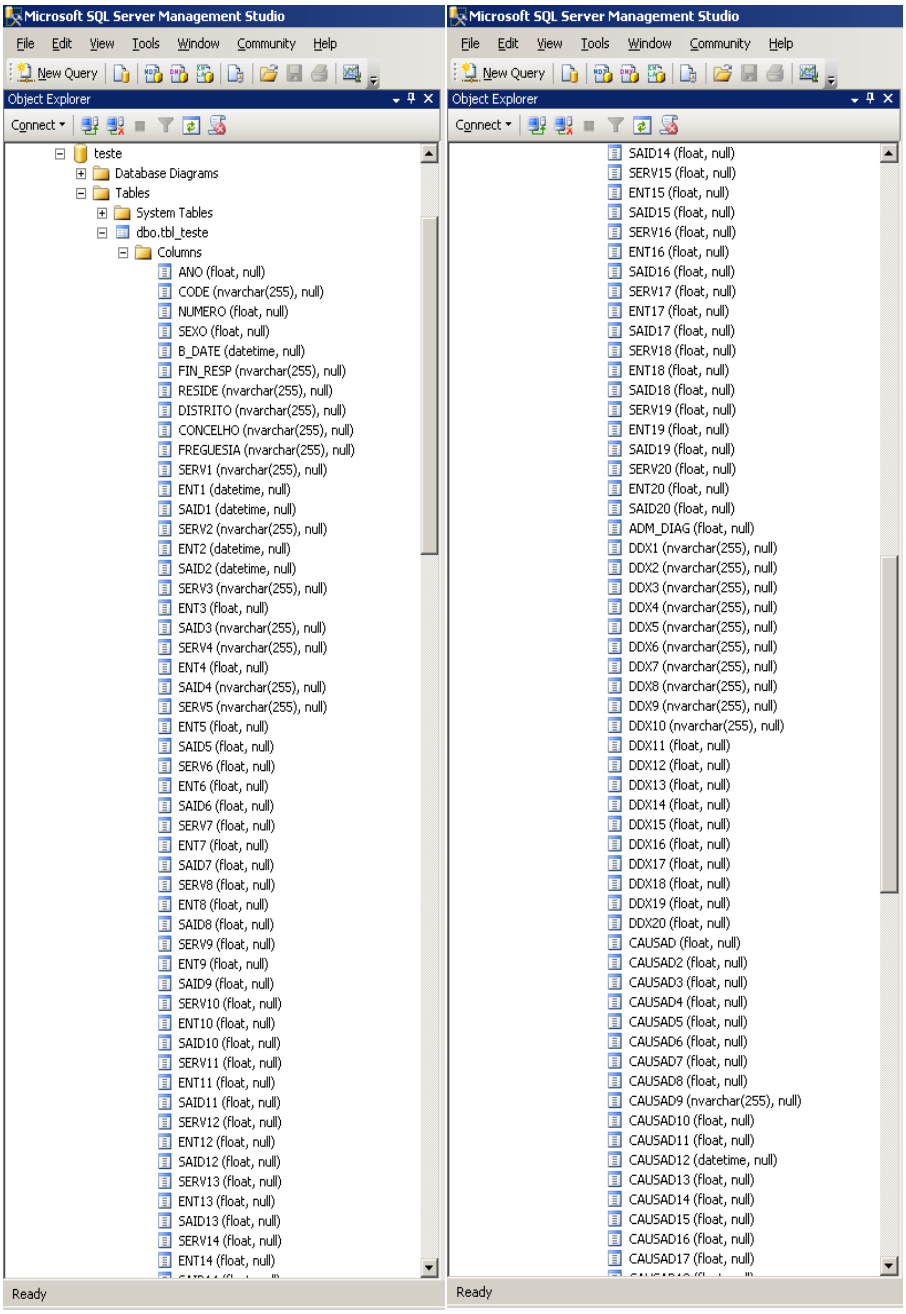


Figura 23 – Estrutura e tipo de dados (parte 1).

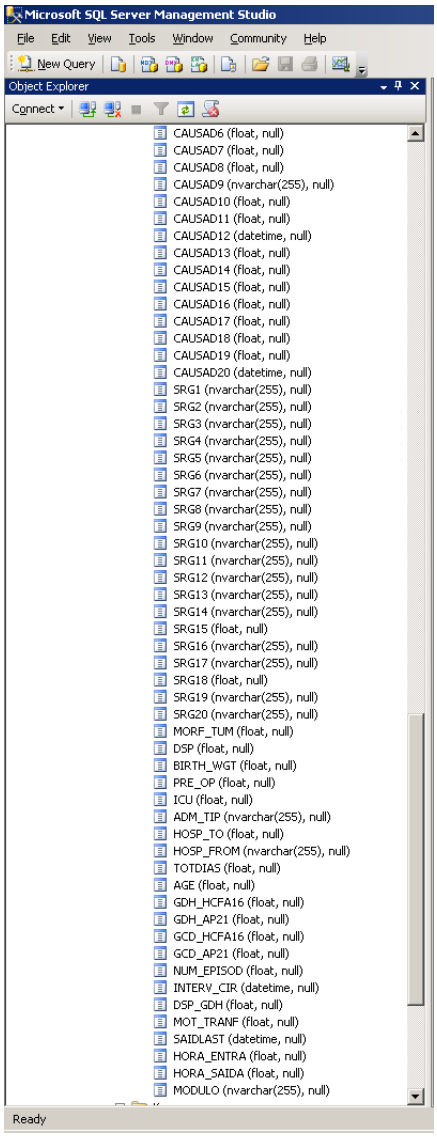


Figura 24 – Estrutura e tipo de dados (parte 2).

Verificou-se que a tabela era constituída por 153 colunas, ou seja, o que revela alguma inflexibilidade e complexidade do sistema para efectuar análises aos dados. Para se prosseguir para o próximo passo de criação do modelo de dados, tornou-se extremamente necessário dissecar o significado de cada campo da tabela. Para isso, solicitou-se à ACSS (nomeadamente à Dr. Manuela Rolim) informação sobre a codificação dos GDH´s e por outro lado, encontrou-se numa das inúmeras pesquisas na internet um documento designado “Auditoria às Bases de Dados dos GDH´s” que contem alguns dos significados dos campos contidos nos ficheiros fonte. A junção das duas fontes informação resultou numa tabela (Figura 25) constituída pelo nome do campo e respectivo significado.

Nome do Campo Fonte	Significado
ANO	Ano da codificação dos GDH´s
CODE	Sem descrição
NUMERO	Número de identificação do paciente
SEXO	Sexo do paciente
B_DATE	Data de nascimento do paciente
FIN_RESP	Entidade financeira responsável
RESIDE	Código interno da Freguesia onde reside o paciente
DISTRITO	Nome do Distrito de residência do Paciente
CONCELHO	Nome do Concelho de residência do Paciente
FREGUESIA	Nome da Freguesia de residência do Paciente
SERV1..20	Serviços
ENT1..20	Serviços – Datas de entrada
SAID1..20	Serviços – Datas de saída
ADM_DIAG	Diagnósticos de admissão
DDX1..20	Diagnósticos principal e secundários
CAUSAD..20	Causa externa
SRG1..SRG20	Procedimentos
MORF_TUM	Morfologia Tumoral
DSP	Destino após alta

BIRTH_WGT	Peso à nascença (Kg)
PRE_OP	Número de dias de pré-operatório
ICU	Número de dias em unidade de cuidados intensivos
ADM_TIP	Tipo de admissão
HOSP_TO	Hospital destino
HOSP_FROM	Hospital de proveniência
TOTDIAS	Número de dias de internamento
AGE	Idade do paciente
GDH_HCFA16	Classificação GDH segundo a HCFA[20]
GDH_AP21	Mesma classificação GDH (mas de outra portaria)
GCD_HCFA16	Classificação GCD segundo a HCFA
GCD_AP21	Mesma classificação GCD (mas de outra portaria)
NUM_EPISOD	Número do episódio
INTERV_CIR	Data de intervenção cirúrgica
DSP_GDH	Destino após alta
MOT_TRANF	Motivo da transferência
SAIDLAST	Data da alta
HORA_ENTRA	Hora de entrada
HORA_SAIDA	Hora da alta

[20] A Health Care and Financing Administration (HCFA) foi a organização percursora dos Centers for Medicare and Medicaid Services (CMS) que corresponde ao sistema de saúde americano, o qual utiliza a Classificação Internacional de Doenças (CID-9-MC) e os GDHs.

MODULO	Unidade (Ex: ambulatório)

Figura 25 – Significado das estruturas.

Após o levantamento do significado da estrutura de dados iniciou-se a interpretação dos dados em si. Para isso, efectuaram-se consultas (“queries”) à tabela de teste de forma a compreender o seu conteúdo. O resultado obtido para cada campo descrito acima foi um conjunto de códigos (Figura 26 e Figura 27), ou seja, identificadores aos quais é necessário complementar com os descritivos.

ANO	CODE	NUMERO	SEXO	B_DATE	FIN_RESP	RESIDE	DISTRITO	CONCELHO	FREGUESIA	SERV1	ENT1	SAID1	SERV2	ENT2	SAID2	SERV3	ENT3	SAID3
2008	P139	77187128	1	1959-03-31 00:00:00.000	971010	120709	Portalegre	Elvas	Terrugem	36005	2008-10-19 00:00:00.000	2008-10-20 00:00:00.000	36001	2008-10-20 00:00:00.000	2008-10-23 00:00:00.000	NULL	NULL	NULL
2008	P139	77187207	1	1974-12-21 00:00:00.000	912001	120703	Portalegre	Elvas	AssunÇÃO	36003	2008-10-24 00:00:00.000	2008-10-31 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77187288	1	1953-04-20 00:00:00.000	920012	070101	gvora	Alandroal	Alandroal (N Sra da ConceiÇÃO)	36002	2008-10-31 00:00:00.000	2008-11-04 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186954	1	1930-01-02 00:00:00.000	971010	120403	Portalegre	Campo Maior	SÃO JoÃO Baptista	36006	2008-10-13 00:00:00.000	2008-10-31 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186955	1	1953-11-24 00:00:00.000	9991	070301	gvora	Borba	Borba (Matriz)	36002	2008-10-03 00:00:00.000	2008-10-05 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186962	2	1947-07-26 00:00:00.000	971010	121404	Portalegre	Portalegre	Fortios	31010	2008-10-03 00:00:00.000	2008-10-04 00:00:00.000	31024	2008-10-04 00:00:00.000	2008-10-06 00:00:00.000	NULL	NULL	NULL
2008	P139	77186968	2	1958-10-11 00:00:00.000	971010	120403	Portalegre	Campo Maior	SÃO JoÃO Baptista	36006	2008-10-04 00:00:00.000	2008-10-08 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186991	2	1921-04-15 00:00:00.000	971010	120504	Portalegre	Castelo de Vide	SÃO JoÃO Baptista	31010	2008-10-06 00:00:00.000	2008-10-08 00:00:00.000	31020	2008-10-08 00:00:00.000	2008-10-15 00:00:00.000	NULL	NULL	NULL
2008	P139	77187045	2	1932-02-02 00:00:00.000	971010	070301	gvora	Borba	Borba (Matriz)	36005	2008-10-11 00:00:00.000	2008-10-12 00:00:00.000	36006	2008-10-12 00:00:00.000	2008-10-31 00:00:00.000	NULL	NULL	NULL
2008	P139	77187052	2	1919-05-05 00:00:00.000	914001	120401	Portalegre	Campo Maior	Nossa Senhora da ExpectaÇÃO	36006	2008-10-12 00:00:00.000	2008-10-14 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77187063	1	1996-10-15 00:00:00.000	913001	120706	Portalegre	Elvas	Santa EulÁlia	36002	2008-11-12 00:00:00.000	2008-11-12 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77187092	2	1913-11-11 00:00:00.000	971010	071402	gvora	Vila Viçosa	Ciladas	36006	2008-10-15 00:00:00.000	2008-10-17 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77187103	1	1937-07-06 00:00:00.000	971010	120603	Portalegre	Crato	Flor da Rosa	31010	2008-10-16 00:00:00.000	2008-10-17 00:00:00.000	31023	2008-10-17 00:00:00.000	2008-10-22 00:00:00.000	NULL	NULL	NULL
2008	P139	77186766	2	1980-05-23 00:00:00.000	971010	120802	Portalegre	Fronteira	Fronteira	31002	2008-09-22 00:00:00.000	2008-09-23 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186767	2	1977-10-25 00:00:00.000	911001	121409	Portalegre	Portalegre	SÚ	31002	2008-09-22 00:00:00.000	2008-09-24 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186771	1	2008-09-22 00:00:00.000	911001	121408	Portalegre	Portalegre	SÕo Lourenço	31017	2008-09-22 00:00:00.000	2008-09-24 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186773	2	1922-07-03 00:00:00.000	971010	120702	Portalegre	Elvas	AlcÕbova	36006	2008-09-22 00:00:00.000	2008-09-29 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186782	1	1924-05-25 00:00:00.000	971010	120705	Portalegre	Elvas	Caia e SÃO Pedro	36005	2008-09-23 00:00:00.000	2008-09-23 00:00:00.000	36006	2008-09-23 00:00:00.000	2008-09-30 00:00:00.000	NULL	NULL	NULL
2008	P139	77186787	2	1924-12-01 00:00:00.000	971010	120303	Portalegre	Avis	Avis	36003	2008-09-23 00:00:00.000	2008-09-26 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186787	2	1924-12-01 00:00:00.000	971010	120303	Portalegre	Avis	Avis	36003	2008-10-14 00:00:00.000	2008-10-24 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186795	2	1992-03-13 00:00:00.000	971010	120605	Portalegre	Crato	Monte da Pedra	31009	2008-09-24 00:00:00.000	2008-09-28 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186818	1	2008-07-07 00:00:00.000	916002	120703	Portalegre	Elvas	AssunÇÃO	31013	2008-09-25 00:00:00.000	2008-09-25 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186825	1	1940-06-23 00:00:00.000	911001	121408	Portalegre	Portalegre	SÕo Lourenço	31018	2008-09-25 00:00:00.000	2008-09-30 00:00:00.000	31023	2008-09-30 00:00:00.000	2008-10-03 00:00:00.000	NULL	NULL	NULL
2008	P139	77186830	1	2003-01-07 00:00:00.000	913001	121409	Portalegre	Portalegre	SÚ	31013	2008-09-26 00:00:00.000	2008-10-01 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186843	2	1926-08-03 00:00:00.000	911001	121404	Portalegre	Portalegre	Fortios	31002	2008-09-26 00:00:00.000	2008-10-10 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186862	2	1933-07-16 00:00:00.000	971010	120202	Portalegre	Arronches	Esperança	31002	2008-09-27 00:00:00.000	2008-10-08 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186867	1	1975-08-08 00:00:00.000	971010	120201	Portalegre	Arronches	AssunÇÃO	31022	2008-09-29 00:00:00.000	2008-10-03 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186870	1	1940-02-16 00:00:00.000	971010	120501	Portalegre	Castelo de Vide	N Sra da Graça de PÑvoa Me...	31022	2008-10-29 00:00:00.000	2008-10-31 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186893	2	1932-08-24 00:00:00.000	971010	120702	Portalegre	Elvas	AlcÕbova	36003	2008-09-29 00:00:00.000	2008-10-07 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186904	2	1981-12-21 00:00:00.000	971010	121502	Portalegre	Sousel	Casa Branca	31009	2008-10-03 00:00:00.000	2008-10-07 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186931	1	2006-10-14 00:00:00.000	971010	120703	Portalegre	Elvas	AssunÇÃO	31013	2008-10-01 00:00:00.000	2008-10-03 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL
2008	P139	77186587	2	1926-05-31 00:00:00.000	971010	121408	Portalegre	Portalegre	SÕo Lourenço	31002	2008-10-06 00:00:00.000	2008-10-10 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL

Figura 26 – Excerto dos dados (parte 1).

DSP	BIRTH_WGT	PRE_OP	ICU	ADM_TIP	HOSP_TO	HOSP_FROM	TOTDIAS	AGE	GDH_HCFA16	GDH_AP21	GCD_HCFA16	GCD_AP21	NUM_EPISOD	INTERV_CIR	DSP_GDH	MOT_TRANF	SAIDLAST	HORA_ENTRA	HORA_SAIDA
1	0	0	0	2	NULL	NULL	4	49	122	122	5	5	28005056	NULL	1	0	2008-10-23 00:00:00.000	28463	40260
1	0	0	0	2	NULL	NULL	7	33	90	90	4	4	28005221	NULL	1	0	2008-10-31 00:00:00.000	19331	50400
1	0	0	0	2	NULL	NULL	4	55	444	445	21	21	28005420	NULL	1	0	2008-11-04 00:00:00.000	57056	52920
1	0	0	0	2	NULL	NULL	18	78	17	832	1	1	28004899	NULL	1	0	2008-10-31 00:00:00.000	43616	56280
1	0	0	0	2	NULL	NULL	2	54	399	399	16	16	28004612	NULL	1	0	2008-10-05 00:00:00.000	2001	54000
20	0	0	0	2	NULL	NULL	3	61	127	544	5	5	28004627	NULL	20	0	2008-10-06 00:00:00.000	43088	40938
2	0	0	0	2	NULL	NULL	4	50	205	205	7	7	28004640	NULL	2	2	2008-10-08 00:00:00.000	48770	62400
20	0	1	7	2	NULL	NULL	9	67	148	505	6	6	28004711	2008-10-07 00:00:00.000	20	0	2008-10-15 00:00:00.000	30172	31200
1	0	0	0	2	NULL	NULL	20	76	321	320	11	11	28004840	NULL	1	0	2008-10-31 00:00:00.000	51371	56340
20	0	0	0	2	NULL	NULL	2	89	79	79	4	4	28004855	NULL	20	0	2008-10-14 00:00:00.000	23432	32400
1	0	0	0	1	NULL	NULL	0	12	163	163	6	6	28005755	2008-11-12 00:00:00.000	1	0	2008-11-12 00:00:00.000	53805	71880
20	0	0	0	2	NULL	NULL	2	88	89	89	4	4	28004965	NULL	20	0	2008-10-17 00:00:00.000	54117	12600
1	0	0	0	2	NULL	NULL	6	71	14	14	1	1	28004988	NULL	1	0	2008-10-22 00:00:00.000	21603	60400
1	0	0	0	2	NULL	NULL	1	28	183	814	6	6	28004274	NULL	1	0	2008-09-23 00:00:00.000	47014	43980
1	0	0	0	2	NULL	NULL	2	30	321	321	11	11	28004275	NULL	1	0	2008-09-24 00:00:00.000	50887	36900
1	3420	0	0	2	NULL	NULL	2	0	391	629	15	15	28004280	NULL	1	0	2008-09-24 00:00:00.000	51129	53700
1	0	0	0	2	NULL	NULL	7	86	17	16	1	1	28004284	NULL	1	0	2008-09-29 00:00:00.000	54630	59460
1	0	0	0	2	NULL	NULL	7	84	14	14	1	1	28004309	NULL	1	0	2008-09-30 00:00:00.000	42099	58320
1	0	0	0	2	NULL	NULL	3	83	240	240	8	8	28004320	NULL	1	0	2008-09-26 00:00:00.000	74489	62580
1	0	0	0	2	NULL	NULL	10	83	233	233	8	8	28004932	2008-10-14 00:00:00.000	1	0	2008-10-24 00:00:00.000	58788	59340
1	0	1	0	2	NULL	NULL	4	16	371	371	14	14	28004334	2008-09-25 00:00:00.000	1	0	2008-09-28 00:00:00.000	35371	34800
2	0	0	0	2	NULL	NULL	0	0	137	137	5	5	28004375	NULL	2	3	2008-09-25 00:00:00.000	58673	61200
1	0	0	0	2	NULL	NULL	8	68	139	139	5	5	28004396	NULL	1	0	2008-10-03 00:00:00.000	69934	51300
1	0	0	0	2	NULL	NULL	5	5	322	322	11	11	28004402	NULL	1	0	2008-10-01 00:00:00.000	897	80100
1	0	0	0	2	NULL	NULL	14	82	211	211	8	8	28004415	2008-09-26 00:00:00.000	1	0	2008-10-10 00:00:00.000	33679	35220
1	0	0	0	2	NULL	NULL	11	75	208	208	7	7	28004439	NULL	1	0	2008-10-08 00:00:00.000	39368	35020
1	0	0	0	2	NULL	NULL	5	33	84	84	4	4	28004452	NULL	1	0	2008-10-03 00:00:00.000	29750	50940
1	0	1	0	6	NULL	NULL	2	68	119	119	5	5	28005374	2008-10-30 00:00:00.000	1	0	2008-10-31 00:00:00.000	35239	42000
1	0	0	0	2	NULL	NULL	8	76	127	127	5	5	28004491	NULL	1	0	2008-10-07 00:00:00.000	67077	57780
1	0	1	0	2	NULL	NULL	4	26	371	371	14	14	28004622	2008-10-04 00:00:00.000	1	0	2008-10-07 00:00:00.000	70264	78720
1	0	0	0	2	NULL	NULL	2	1	184	777	6	6	28004572	NULL	1	0	2008-10-03 00:00:00.000	50073	64800
1	0	0	0	2	NULL	NULL	4	82	189	189	6	6	28004704	NULL	1	0	2008-10-10 00:00:00.000	42635	59580

Figura 27 – Excerto dos dados (parte 2).

Como exemplo, temos a coluna serviço (SERV1), tipo de admissão (ADM_TIP) e classificação do GDH (GDH_HCFA16) apenas com identificadores que terão correspondência com os respectivos descritivos. Outro facto a salientar consiste na qualidade dos dados que é bastante reduzida, o que se pode concluir pelo número de valores a nulo (NULL). Como exemplo (que se pode verificar nas figuras acima), verifica-se registos de pacientes que não têm Hospital atribuído, ou seja, os campos HOSP_TO e HOSP_FROM encontram-se a nulo. No entanto, o tema da qualidade de dados será retratado mais à frente visto que o mais importante nesta fase centra-se na interpretação/identificação dos códigos que se encontram nos dados. Só após esta parte analisada é que se pode iniciar o desenho do modelo do DW. Regressando ao conteúdo do CD-ROM cedido pela ACSS, verificou-se que este apenas continha o tipo de dados que foi apresentado nas figuras anteriores. Assim sendo, interagiu-se novamente com esta entidade, de forma a que nos fossem fornecidos os descritivos para os códigos que nos enviaram. O resultado da interacção foi bastante positivo e consistiu no envio, por parte da ACSS, de ficheiros Excel (Figura 28) com alguns dos mapeamentos entre os códigos e os descritivos dos campos que foram observados anteriormente.

COD_MOTIVO_TRANSF	DES_MOTIVO_TRANSF
-1	Não Definido
0	Sem transferência
1	Realização de Exames
2	Para Seguimento
3	Por Falta de Recursos
4	Para Tratamento de condição associada

Figura 28 – Exemplo de um ficheiro Excel com o mapeamento do campo motivo de transferência (MOT_TRANF).

Devido à qualidade dos dados e às boas práticas do desenho de DW's, acrescentou-se uma linha a todos os metadados no sentido de classificar os valores a nulo como "não definidos" e o respectivo código a "-1". No futuro, os dados classificados a "-1" poderão ser reclassificados com o valor correcto. Desta forma, é também mais perceptível para o utilizador final, detectar

falhas provenientes do sistema operacional (fonte de dados) ou erros de preenchimento na classificação dos GDH's.

Após esta análise dos dados e de estruturas fonte, iniciou-se o segundo passo que consistiu na criação do modelo do DW de forma a permitir análises multidimensionais.

11.4.3Segunda Fase – Desenho do modelo do DW

Para desenhar o modelo do DW idealizou-se um modelo em estrela, este modelo é muito versátil no sentido de acrescentar mais dimensões/métricas de análise, e de fácil entendimento para os utilizadores finais. Por último, este tipo de modelo clássico permite maior performance na consulta dos dados assim como multidimensionalidade.

Os modelos em estrela são constituídos por tabelas de dimensão e de factos, logo defini uma tabela de factos que contém as chaves estrangeiras (SK's) e as métricas total de dias de internamento (campo fonte TOTDIAS) e número de dias de pré-operatório (campo fonte PRE_OP). A tabela de factos será alimentada através dos ficheiros *.dbf.

Em relação à definição das dimensões, criaram-se as seguintes (Figura 29):

Fonte	Dimensão
Ficheiros Excel	Tipo de Admissão
	Causa Externa
	Diagnósticos
	Destino após alta
	Freguesia
	Concelho
	Distrito
	GCD
	GDH
	Hospital
	Morfologia Tumoral
	Motivo de Transferência

	Serviços
	Procedimento
Criada a partir dos factos	Paciente
Manual	Sazonalidade
Criada a partir dos Analysis Services	Time

Figura 29 – Dimensões do modelo do DW.

Como parte integrante do desenho, definiu-se SK's em todas as dimensões para criar uma camada de abstracção no sentido de armazenar o identificador operacional, que vem da fonte de dados. Outro motivo pelo qual se definiu as SK's consistiu no aumento de performance, visto que algumas chaves operacionais (ID's) eram formadas por caracteres, o que implica ligações entre tabelas ("joins"), consultas ("queries") mais lentas. Como a tabela de factos irá ter uma volumetria considerável (na ordem dos milhões de registos) é necessário aplicar todas as técnicas de aumento de performance, de forma a que o utilizador final não seja prejudicado no tempo que o sistema demora a resolver as suas consultas. Desta forma, ao usar-se as SK's, que são chaves inteiras, obtém-se uma melhor performance nos acessos aos dados.

O passo seguinte originou a criação de uma tabela de mapeamentos entre as fontes de dados e o DW (que se encontra no capítulo de anexo – tabela de mapeamento entre tabelas/colunas fonte e dimensões/colunas do DW) para servir como base à construção do modelo do DW e futuros processos de ETL que irão alimentar com dados as dimensões.

Após a modelação das dimensões, seguiu-se a construção da tabela factos que irá conter as métricas de análise, nomeadamente, o número de dias de internamento e pré-operatório. As dimensões contêm os atributos (descritivos) que irão caracterizar os factos (métricas) sob múltiplas perspectivas. A tabela de factos para permitir a multidimensionalidade será constituída por chaves estrangeiras (SK's das dimensões) e pelas métricas acima referidas. Neste cenário optou-se apenas por uma tabela de factos visto que o contexto das métricas de análise é comum a todas as dimensões.

E com base nesta tabela de mapeamentos definiu-se o modelo do DW (Figura 30).

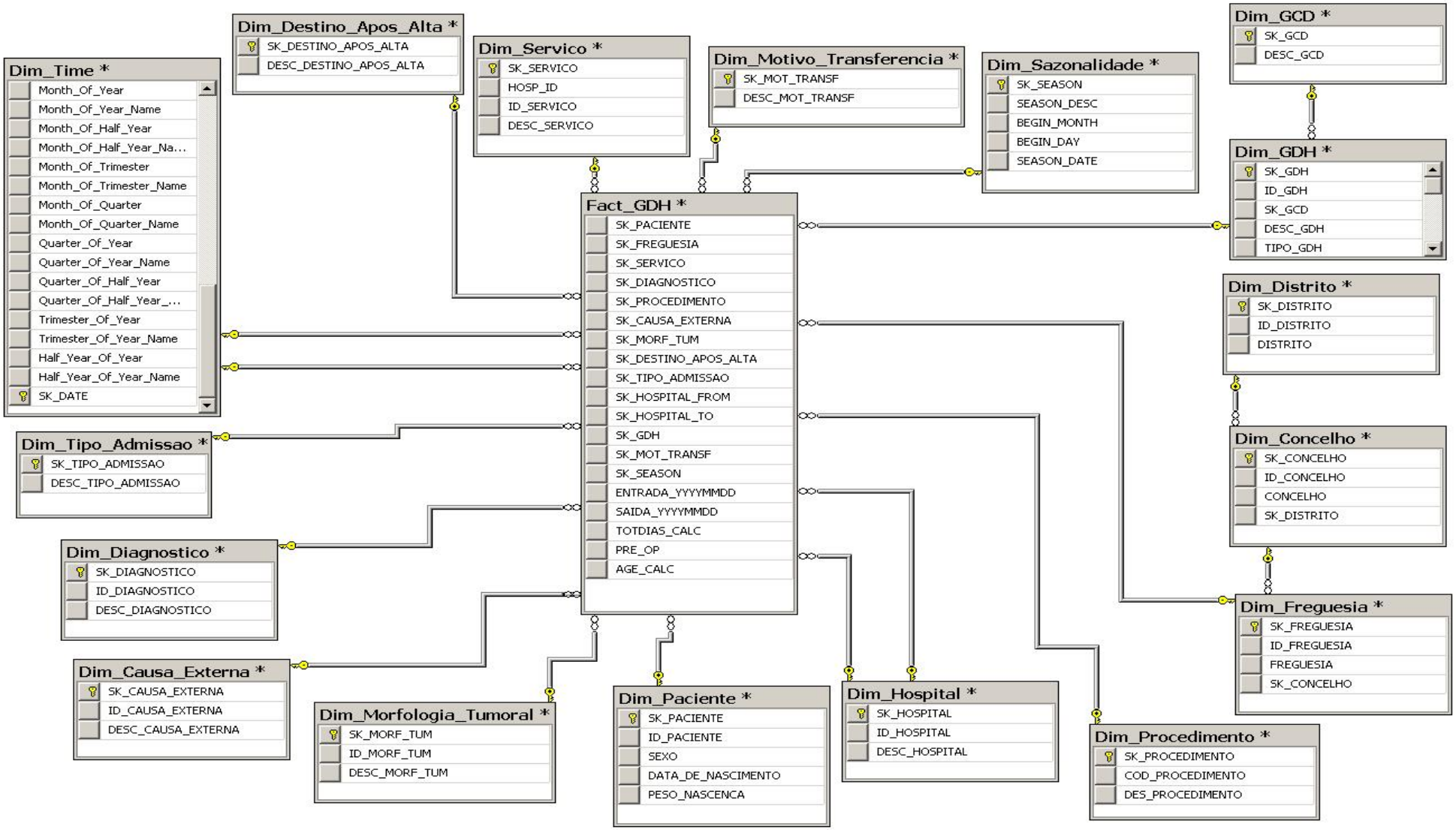


Figura 30 - Modelo do DW.

É de salientar que foi acrescentado ao modelo em estrela uma variante (também ela clássica) que consiste no modelo floco de neve (snow-flake). A componente floco de neve foi construída a através da localidade do paciente em que um distrito tem N concelhos, e um concelho tem N freguesias (Figura 5).

Através deste modelo está bem patente o conceito de multidimensionalidade, que nos permite em analisar os dados por mais do que uma dimensão. A título de exemplo, o utilizador final poderá ter a necessidade de analisar o total de dias de internamento (métrica) em 2008 (dimensão tempo), no Hospital Garcia de Horta (dimensão Hospital) em que o diagnóstico seja "ENCEFALITES VIRAIS TRANSMITIDAS POR CARRACAS" (dimensão diagnóstico).

11.4.4Terceira Fase – Set-up do ambiente de desenvolvimento do ETL

Após o desenho do modelo, iniciou-se a terceira fase da construção do sistema de BI que teve como base o "set-up" do ambiente de desenvolvimento dos processos de ETL, onde os dados serão carregados do sistema fonte, e que posteriormente serão transformados e carregados no DW de forma a estarem disponíveis para consulta.

O primeiro passo para construir os processos de ETL (tanto para as dimensões, como posteriormente para os factos) consistiu em criar um projecto de Integration Services na ferramenta, que vem incluída no SQL Server, designada por Business Intelligence Development Studio (BIDS), Figura 31.

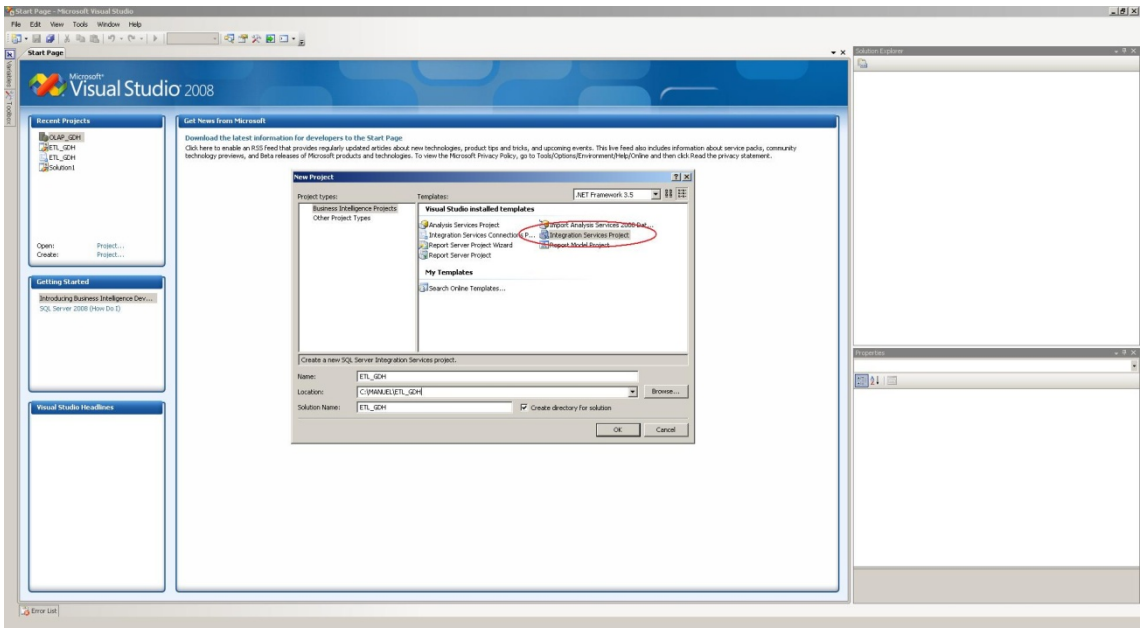


Figura 31 – Criação do projecto de Integration Services no BIDS.

O nome do projecto escolhido foi “ETL_GDH” e será neste projecto onde irão ficar todos os “packages” (processos de ETL) em Integration Services, ou seja, é neste projecto que todo o ETL será desenvolvido.

No desenvolvimento dos “packages” serão usadas algumas das componentes disponíveis na “toolbox” de controlo de fluxo (“control flow”) e de fluxo de dados (“data flow”) do BIDS. Na Figura 32 encontram-se algumas dessas componentes.

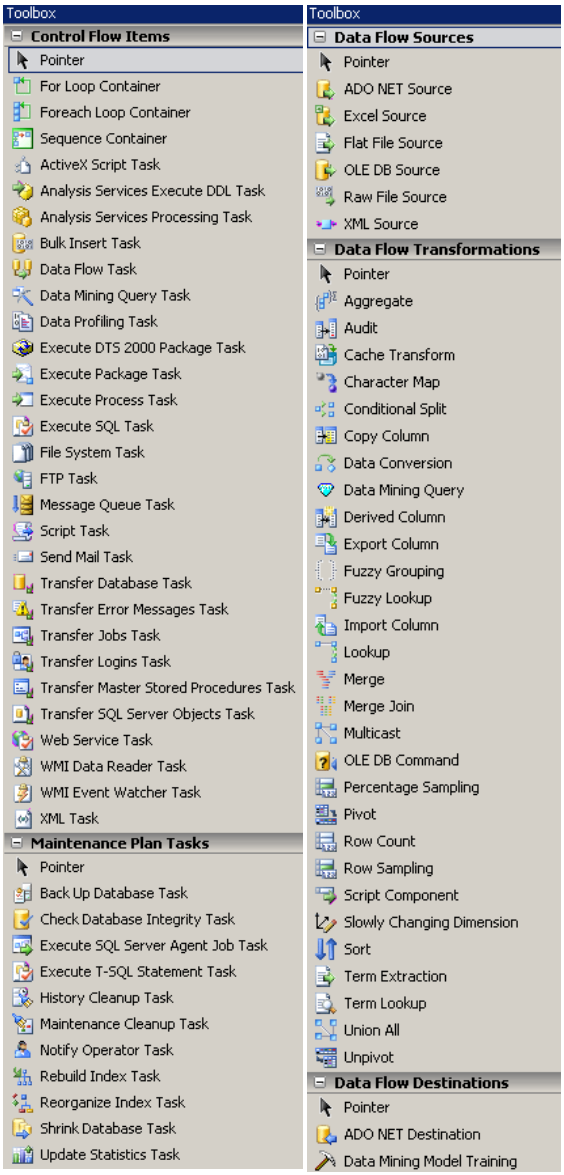


Figura 32 – Toolbox’s de componentes de controlo de fluxo e de fluxo de dados do BIDS.

A parametrização destas componentes inclui em grande parte desenvolvimento à medida. Para evitar concorrência e degradação de performance entre as análises dos utilizadores finais e as operações de transformações de dados realizadas pelos processos de ETL, criou-se uma base de dados intermédia designada “Staging Area” onde serão realizados todos os passos

intermédios de tratamentos de dados. Após os dados encontrarem-se no formato final, serão copiados para o DW. A base de dados criada, que funciona como DW, é designada por "DW_GDH".

Depois de criadas as bases de dados, configurou-se no projecto "ETL_GDH" as ligações ("connection strings") à SA e ao DW (Figura 33). As ligações ao sistema fonte serão abordadas mais à frente, visto que variam consoante o ficheiro Excel a carregar. Em relação à ligação aos factos, existem algumas especificidades que serão detalhadas mais à frente no capítulo de construção e carregamento dos factos.

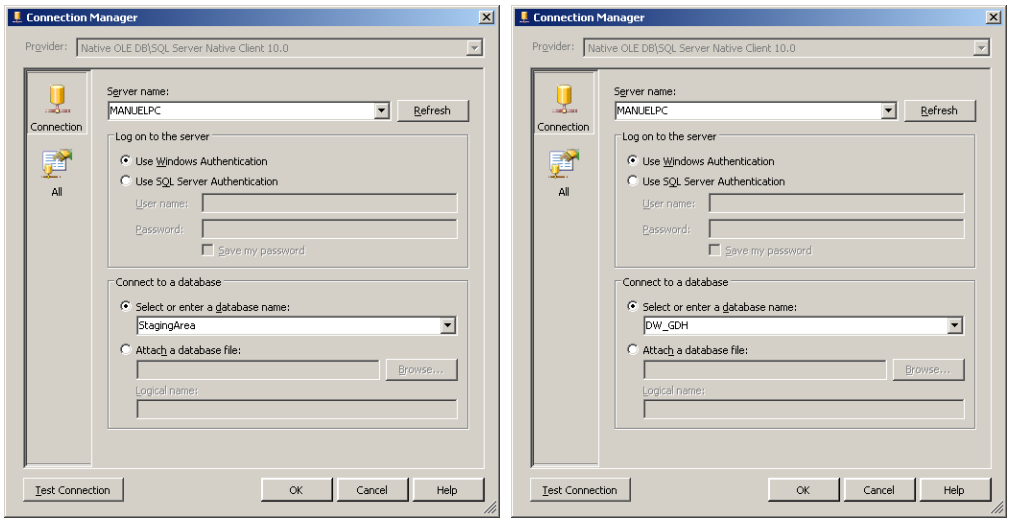


Figura 33 – Criação das ligações à base de dados Staging Area e DW_GDH.

Por último, foram desenvolvidos 14 "packages", um para cada dimensão excepto a dimensão tempo que seguiu outro processo, que será detalhado neste trabalho, um pouco mais à frente. Em relação aos factos, foi criado um "package" específico para este carregamento (Figura 34).

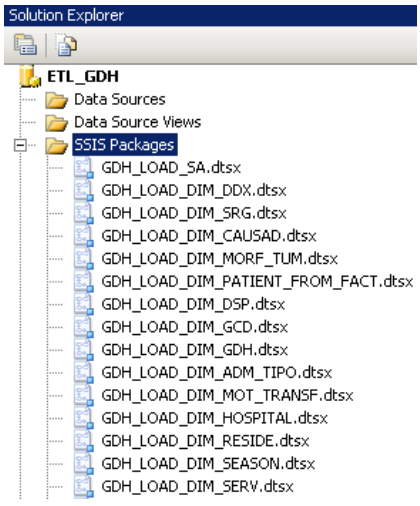


Figura 34 – Lista de "packages" (processos de ETL) desenvolvidos para carregamento do DW.

Estes 15 “packages”, que consistem nos processos de ETL para o carregamento do DW, serão detalhados, de forma minuciosa, nos capítulos seguintes.

11.4.5Quarta Fase – Construção e carregamento das Dimensões

Terminado o “set-up” do ambiente de desenvolvimento e tendo por base o desenho do modelo do DW, iniciou-se a implementação dos processos de ETL para alimentar com dados as dimensões.

As dimensões tiveram como fonte principal os ficheiros em Excel, Analysis Services e apenas uma delas se irá manter de forma manual.

11.4.5.1 Carregamento da Dimensão Diagnósticos

Para a dimensão diagnóstico foi criada, através do MS, a seguinte tabela no DW (Figura 35):

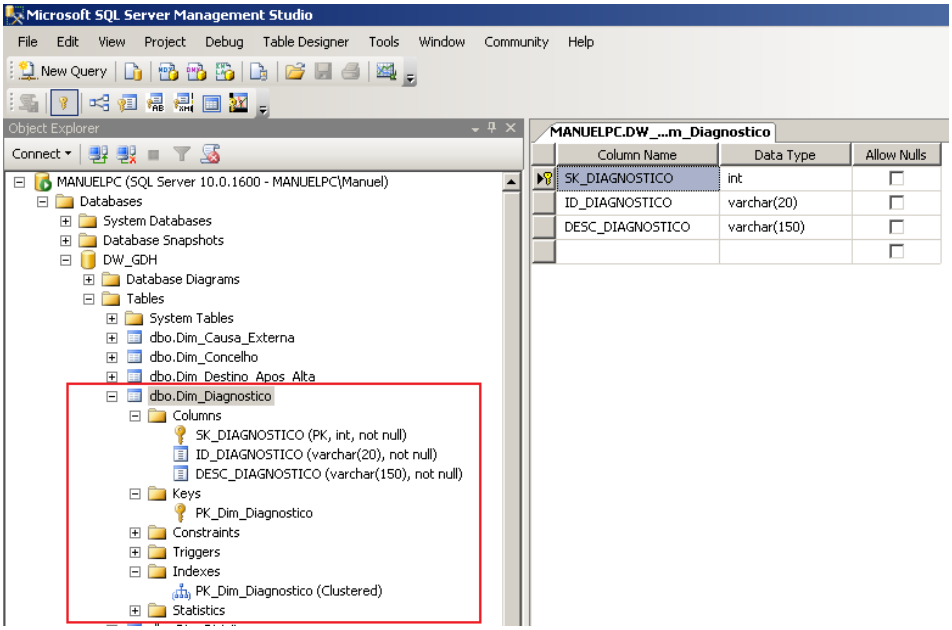


Figura 35 – Estrutura da tabela de dimensão diagnóstico.

Nesta tabela de dimensão diagnóstico as colunas são a chave interna (SK) do DW, ou seja, a SK_DIAGNOSTICO, a chave operacional que vem do sistema fonte, o ID_DIAGNOSTICO e a descrição do diagnóstico, proveniente também do sistema fonte, DESC_DIAGNOSTICO.

De salientar que se definiu como chave primária da tabela a SK, o que por sua vez, leva à criação automática de índice “clustered”, ou seja, os dados contidos nesta tabela estão ordenados pela coluna SK_DIAGNOSTICO para que o acesso aos dados e o seu cruzamento com a tabela de factos seja efectuado o mais rápido possível.

A fonte de dados de acordo com o capítulo de anexo "Tabela de mapeamento entre tabelas/colunas fonte e dimensões/colunas do DW" é um ficheiro Excel (SRC_DDX_DIAGNOSTICO.xlsx).

Criada a tabela destino, dimensão diagnóstico, e identificada a fonte de dados, criou-se um novo "package" (Figura 36) no projecto ETL_GDH, que consistirá num processo de ETL para carregar os dados da fonte para o destino que neste caso é a dimensão diagnóstico.

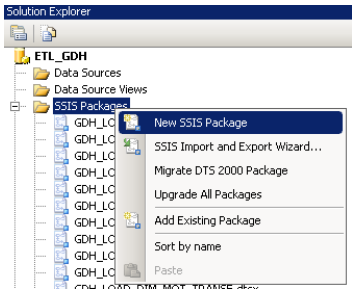


Figura 36 – Criação de um novo "package" (processo de ETL) para carregamento de dados.

O "package" de Integration Services criado, designado por "GDH_LOAD_DIM_DDX.dtsx" é constituído por uma componente de controlo de fluxo, "control flow" (Figura 37).

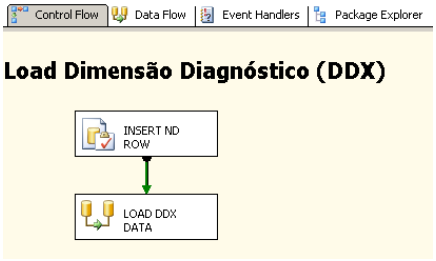


Figura 37 – Componente controlo de fluxo do package "GDH_LOAD_DIM_DDX.dtsx".

A primeira tarefa ("task") consiste numa "Execute SQL Task", designada "INSERT ND ROW" que insere na tabela "Dim_Diagnostico" um registo com o descritivo "NÃO DEFINIDO", SK com o valor -1 e ID com o valor "ND" (não definido). Caso o registo já exista na dimensão, este não será inserido novamente de forma a possibilitar o reprocessamento das dimensões sem inserção de duplicados. Como a coluna SK_DIAGNOSTICO é do tipo "identity", ou seja, é incrementada automaticamente, é necessário desligar o identity para inserir um registo com outro valor numérico diferente do último inserido + 1. Após o registo inserido volta-se a ligar o identity para a dimensão ao receber os dados fonte, a coluna SK_DIAGNOSTICO voltar a ser incrementada automaticamente de forma a que chave seja unívoca. O código implementado na primeira tarefa de carregamento da dimensão encontra-se na Figura 38.

```
If (select COUNT(1) from Dim_Diagnostico where SK_DIAGNOSTICO=-1)=0
```

```
Begin

SET IDENTITY_INSERT Dim_Diagnostico ON

INSERT INTO Dim_Diagnostico
([SK_DIAGNOSTICO]
,[ID_DIAGNOSTICO]
,[DESC_DIAGNOSTICO])
VALUES
(-1
,'ND'
,'NÃO DEFINIDO')

SET IDENTITY_INSERT Dim_Diagnostico OFF
```

Figura 38 – Tarefa de inserção de um registo com o descritivo "não definido".

Este registo é inserido na dimensão no caso do carregamento dos factos existir um ID de diagnóstico que não exista na dimensão ou que venha a nulo da fonte, o seu valor de SK será "-1" de forma a que o utilizador final perceba que existem registos com diagnósticos não definidos que posteriormente terão de ser carregados para a dimensão e os factos terão de ser reprocessados.

Após esta tarefa, surge outra do tipo fluxo de dados ("Data Flow Task"), designada por "LOAD DDX DATA" que consiste em carregar os dados do ficheiro fonte para o destino (tabela Dim_Diagnostico). Como se trata de processamento de dados, este tratamento é elaborado na secção de fluxo de dados ("data flow") do package (Figura 39).

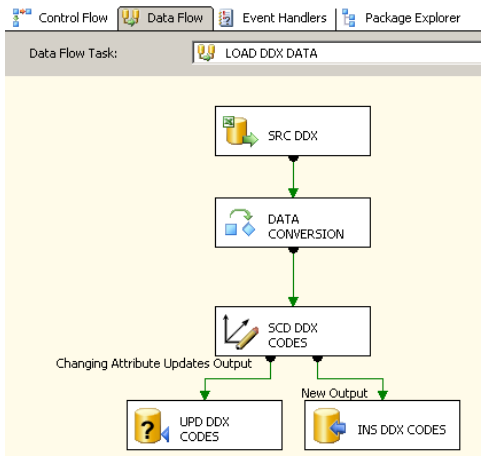


Figura 39 – Fluxo de dados para o carregamento da dimensão diagnóstico.

Começou-se por criar uma ligação ao ficheiro de Excel (Figura 40) que contém os dados fonte dos diagnósticos.

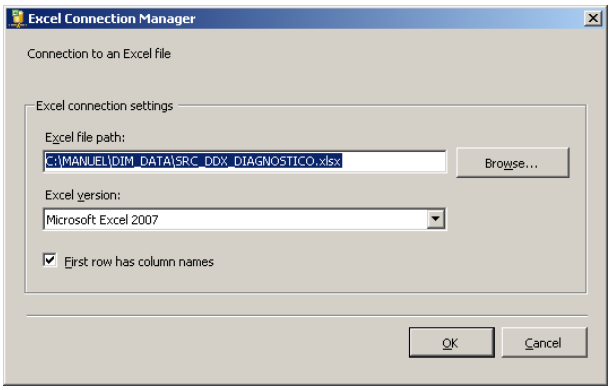


Figura 40 – Criação da ligação ao ficheiro Excel.

Definiu-se no fluxo de dados a fonte de dados, neste caso escolheu-se a fonte "Excel Source", renomeou-se para "SRC DDX" e configurou-se qual a "sheet" de Excel que iria ser importada assim como as respectiva colunas (Figura 41).

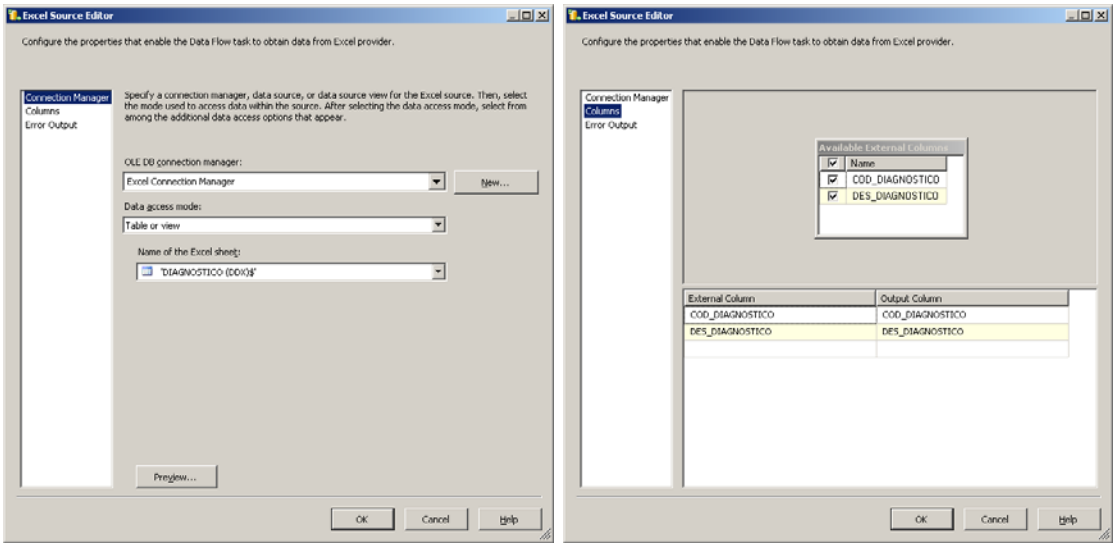


Figura 41 – Configuração da fonte de dados.

O passo seguinte consistiu na conversão de dados fonte visto que o tipo de dados da fonte não é igual ao tipo de dados do destino (Figura 42). A fonte de dados é do tipo "unicode string" e o destino "non-unicode string", logo com a conversão de tipo de dados esta questão ficou completamente resolvida.

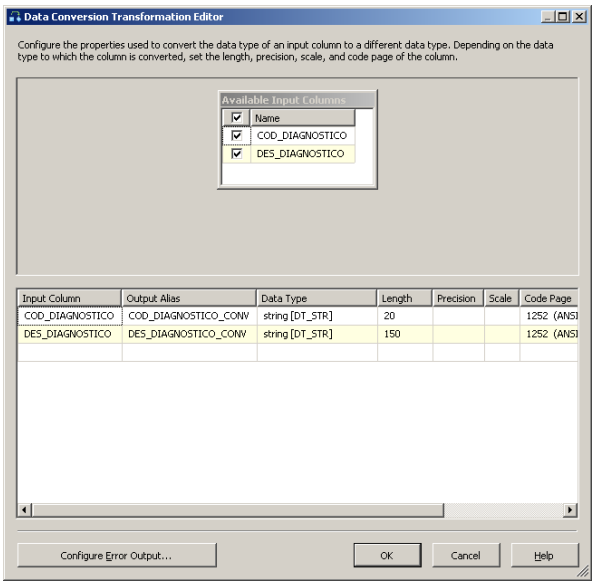


Figura 42 – Parametrização da conversão do tipo de dados.

Seguidamente, utilizou-se a tarefa (“Slowly Changing Dimension”), designada por “SCD DDX CODES”, para escolher o tipo de carregamento a efectuar, conforme foi descrito no capítulo dedicado ao Slowly Changing Dimensions.

Para parametrizar esta tarefa é necessário efectuar alguns passos. O primeiro passo consiste em definir os mapeamentos entre as colunas de input (provenientes da ficheiro Excel fonte) e as colunas da tabelas de dimensão (Dim_Diagnostico), assim como atribuir qual a chave (“business key”) do processo de carregamento da tabela que neste caso será o código operacional que vem da fonte (COD_DIAGNOSTICO), que mapeia para a coluna ID_DIAGNOSTICO da tabela de dimensão (Figura 43).

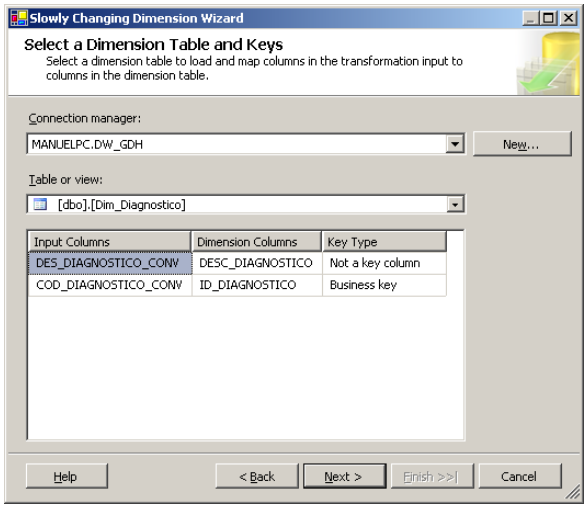


Figura 43 – Mapeamento entre as colunas fonte e destino, assim como definição da chave do negócio.

O passo seguinte, consiste em definir qual o tipo de alterações na dimensão que se pretende implementar. Visto que os descritivos, sobretudo os de diagnósticos, ao longo do ano não sofrem alterações/actualizações, mas podem surgir novos diagnósticos, optou-se pelo tipo 1 (sobreposição do valor), ou seja, isto significa que se o diagnóstico ainda não existir na dimensão é inserido, caso já exista, é actualizado o seu descritivo (Figura 44). Esta opção foi propagada para as restantes dimensões que têm como fonte o Excel.

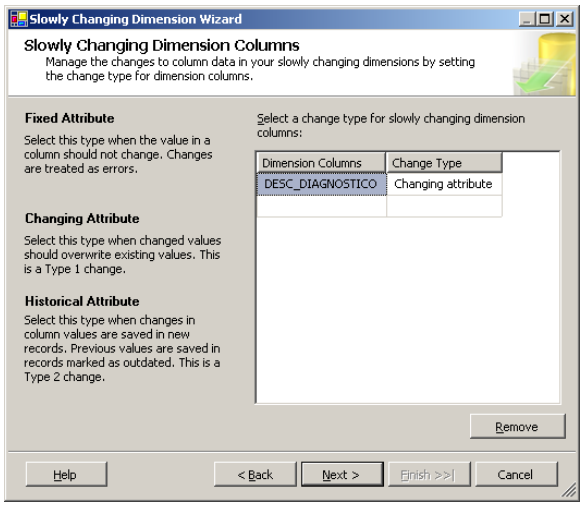


Figura 44 – Definição do tipo de "slowly changing dimension" a aplicar.

O último passo da parametrização do tarefa "slowly changing dimension" consiste em confirmar que todos os registos da dimensão devem ser actualizados sempre que exista uma alteração no seu descritivo fonte (Figura 45).

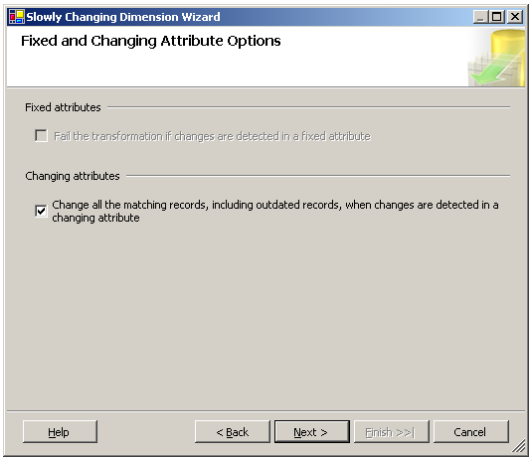


Figura 45 – Confirmação de que todos os registos que já existam na dimensão e que surjam com novos descritivos são actualizados.

Terminada a parametrização da tarefa “slowly changing dimension”, surgem outras duas tarefas, uma de inserção e outra de actualização (“update”) de valores.

A tarefa de inserção permite inserir novos registos na tabela Dim_Diagnostico, enquanto que a tarefa de actualização permite actualizar os descritivos dos códigos de diagnósticos já existentes. É de referir que os apenas os registos em que o seu descritivo foi alterado é que são actualizados, os restantes registos mantêm-se como foram carregados anteriormente. Sempre que existir um diagnóstico novo, esse registo será sempre inserido na Dim_Diagnostico com uma SK_DIAGNÓSTICO única. A SK_DIAGNÓSTICO, como foi referido anteriormente, consiste numa chave interna do tipo inteiro que identifica univocamente os registos da dimensão diagnóstico.

11.4.5.2 Carregamento da Dimensão Procedimento

O processo de carregamento da dimensão procedimento é idêntico ao carregamento da dimensão diagnóstico. Criou-se uma tabela de dimensão com a seguinte estrutura (Figura 46):

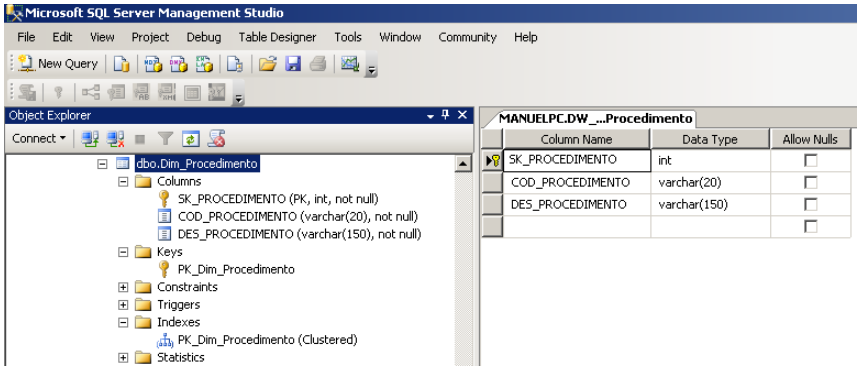


Figura 46 – Estrutura da tabela de dimensão procedimento.

A fonte de dados é o ficheiro Excel “SRC_SRG_PROCEDIMENTO.xlsx”. O processo de ETL (“package”) que efectua o carregamento da dimensão procedimento designa-se por “GDH_LOAD_DIM_SRG.dtsx” e a sua estrutura (Figura 47), tanto ao nível de controlo de fluxo como de fluxo de dados (Figura 48), é idêntica ao “package” de carregamento da dimensão diagnóstico.

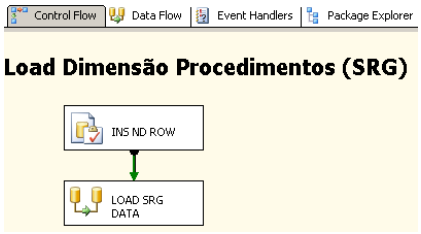


Figura 47 – Componente controlo de fluxo do package “GDH_LOAD_DIM_SRG.dtsx”.

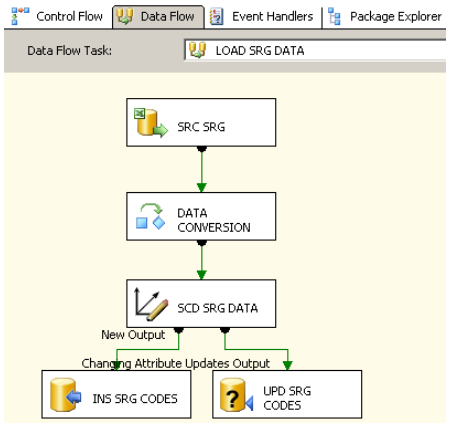


Figura 48 – Componente de fluxo de dados do package "GDH_LOAD_DIM_SRG.dtsx".

11.4.5.3 Carregamento da Dimensão Causa Externa

O processo de carregamento da dimensão causa externa, é idêntico ao carregamento da dimensão procedimento. Criou-se uma tabela de dimensão causa externa com a seguinte estrutura (Figura 49):

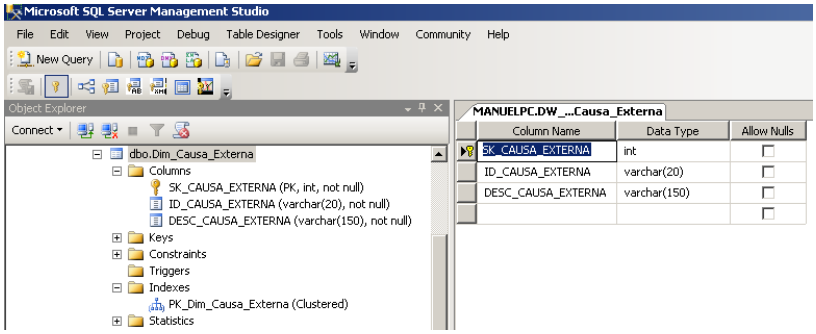


Figura 49 – Estrutura da tabela de dimensão causa externa.

A fonte de dados é o ficheiro Excel "SRC_CAUSAD_CAUSA_EXTERNA.xlsx" e o processo de ETL ("package") que efectua o carregamento da dimensão causa externa designa-se por "GDH_LOAD_DIM_CAUSAD.dtsx" e a sua estrutura, tanto ao nível de controlo de fluxo (Figura 50) como de fluxo de dados (Figura 51), é idêntica ao "package" de carregamento da dimensão procedimento.

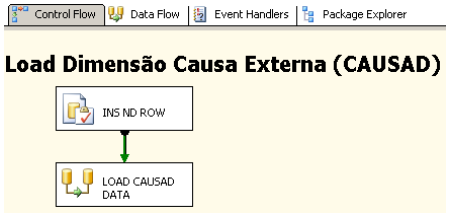


Figura 50 – Componente controlo de fluxo do package "SRC_CAUSAD_CAUSA_EXTERNA.xlsx".

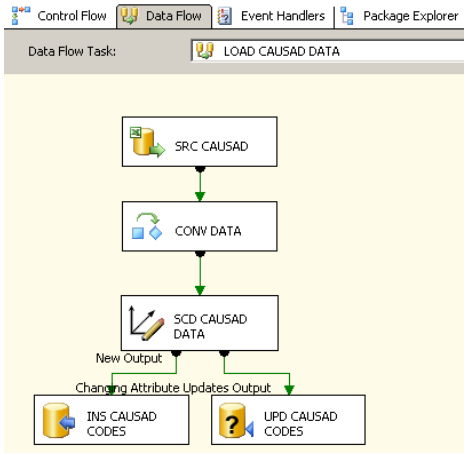


Figura 51 – Componente de fluxo de dados do package "SRC_CAUSAD_CAUSA_EXTERNA.xlsx".

11.4.5.4 Carregamento da Dimensão Morfologia Tumoral

O processo de carregamento da dimensão morfologia tumoral é idêntico ao carregamento da dimensão causa externa. Criou-se uma tabela de dimensão morfologia tumoral com a seguinte estrutura (Figura 52):

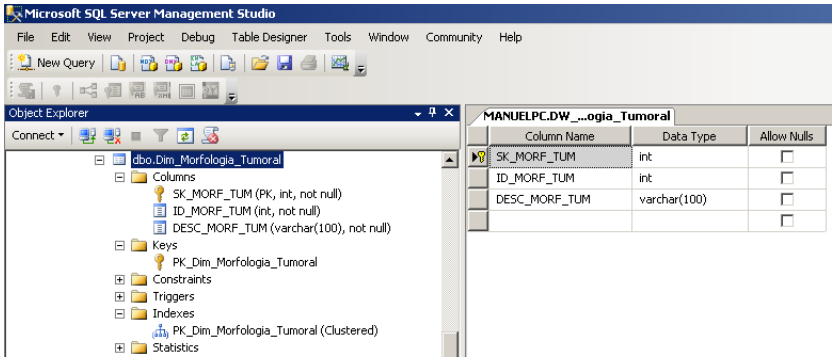


Figura 52 – Estrutura da tabela de dimensão morfologia tumoral.

A fonte de dados é o ficheiro Excel "SRC_MORF_TUM_MORFOLOGIA_TUMORAL.xlsx" e o processo de ETL ("package") que efectua o carregamento da dimensão morfologia tumoral designa-se por "GDH_LOAD_DIM_MORF_TUM.dtsx" e a sua estrutura, tanto ao nível de controlo de fluxo (Figura 53) como de fluxo de dados (Figura 54), é idêntica ao "package" de carregamento da dimensão causa externa.

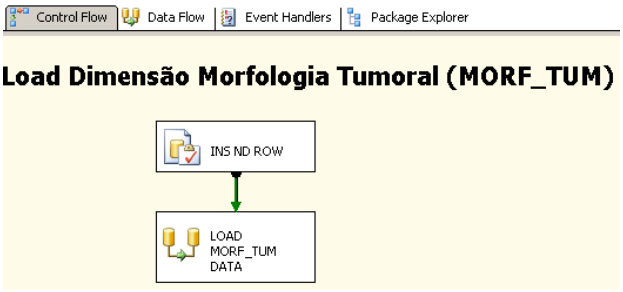


Figura 53 – Componente controlo de fluxo do package "GDH_LOAD_DIM_MORF_TUM.dtsx".

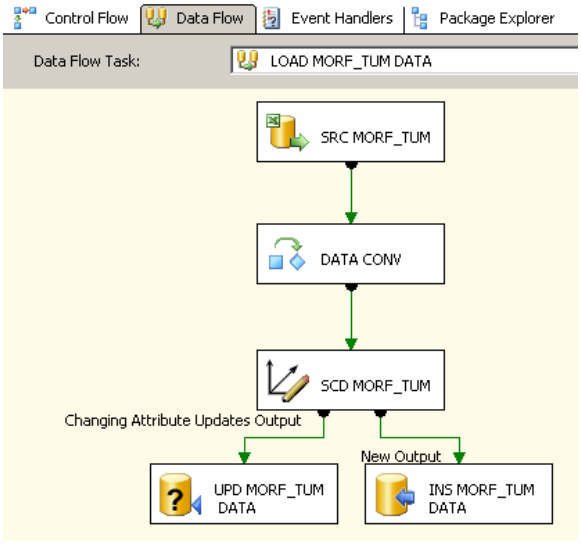


Figura 54 – Componente de fluxo de dados do package "GDH_LOAD_DIM_MORF_TUM.dtsx".

11.4.5.5 Carregamento da Dimensão Paciente

O carregamento da dimensão paciente ("GDH_LOAD_DIM_PATIENT_FROM_FACT.dtsx") é diferente de todos os outros, visto que é único que tem como base os factos, ou seja, esta dimensão é construída com base na tabela de factos.

Assim sendo, começou-se por definir a tabela destino que consiste na dimensão paciente (Figura 55).

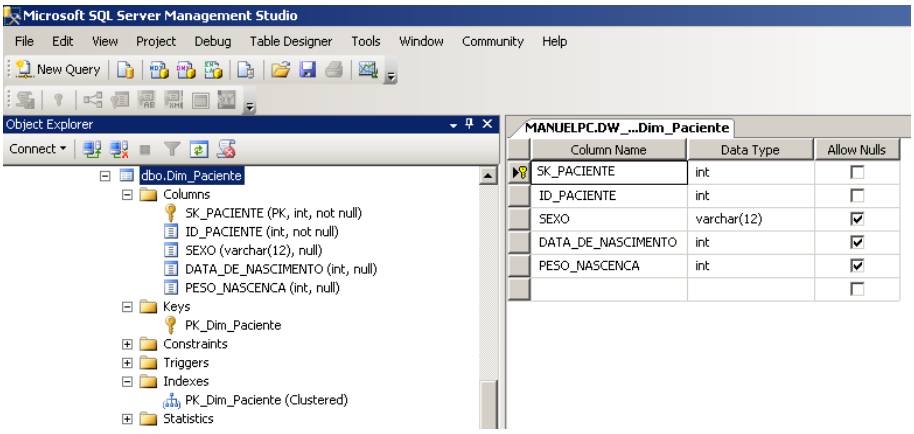


Figura 55 – Estrutura da tabela de dimensão paciente.

Os atributos que compõem a dimensão paciente e que são provenientes da tabela de factos são o identificador do paciente (coluna ID_PACIENTE), o seu sexo (coluna SEXO), a sua data de nascimento (DATA_DE_NASCIMENTO) e o seu peso à nascença (coluna PESO_NASCENCA). Criada a tabela destino, desenvolveu-se o fluxo de controlo da seguinte forma (Figura 56):

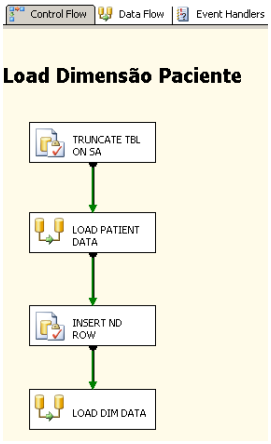


Figura 56 – Fluxo de controlo da dimensão paciente.

Em primeiro lugar criou-se uma tabela intermédia na SA, designada por "TBL_PRE_DIM_PATIENT" (Figura 57) e na primeira tarefa do fluxo de controlo utilizou-se a tarefa "execute SQL task", denominada "TRUNCATE TBL ON SA" para apagar todos os registos desta tabela (caso existam ou não). Este procedimento é realizado no início de forma a contemplar o reprocessamento dos dados.

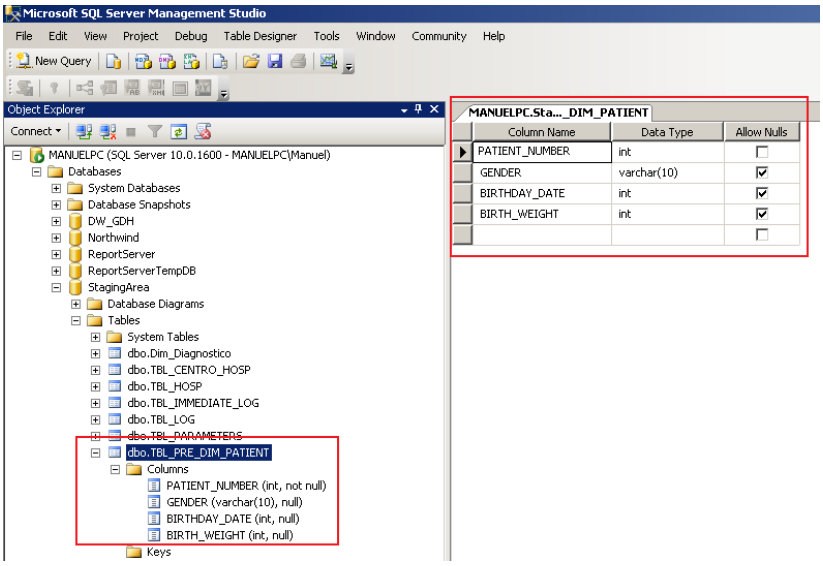


Figura 57 – Tabela intermédia de carregamento da dimensão paciente.

A tarefa seguinte é uma "data flow task" renomeada para "LOAD PATIENT DATA" que tem como fluxo de dados a seguinte estrutura (Figura 58):

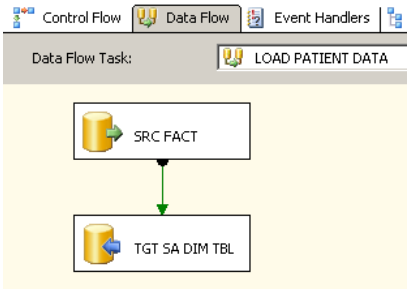


Figura 58 – Fluxo de dados intermédio de dados na Staging Area.

Nesta estrutura, utilizou-se a opção "OLE DB Source" renomeada para "SRC FACT" que contem a "query" de extracção dos factos (Figura 59).

```
select distinct cast(NUMERO as int) as PATIENT_NUMBER,
GENDER=CASE SEXO
            WHEN 1 THEN 'Masculino'
          ELSE 'Feminino'
      END,
  cast(CONVERT(varchar(8),B_DATE,112) as int) as BIRTHDAY_DATE,
      cast(BIRTH_WGT as int) as BIRTH_WEIGHT
from dbo.TBL_SRC_GDH_FACT
```

Figura 59 – Query de extracção dos factos.

A “query” de extracção dos dados efectua conversões aos tipos de dados, assim como traduz os valores numéricos “0” ou “1” para “Masculino” ou “Feminino”, visto que para o utilizador final os códigos não têm qualquer significado, enquanto que os descritivos são totalmente intuitivos e facilmente se percebe o sexo do paciente. A tabela fonte de extracção dos dados em questão (“TBL_SRC_GDH_FACT”) consiste no primeiro passo de importação dos factos (que se encontram nos ficheiros *.dbf) para a Staging Area. Mais detalhes sobre o carregamento dos factos são dados no capítulo “carregamento dos factos”.

Um pormenor importante da “query” de extracção consiste no uso da instrução “distinct” que como o próprio nome indica, apenas retorna valores distintos, ou seja, que não se repetem, visto que na dimensão não é possível ter dados duplicados, caso contrário iria gerar produtos cartesianos quando fosse cruzada informação de factos com esta dimensão. Com a utilização da instrução “distinct” elimina-se a possibilidade de surgirem duplicados.

O resultado desta “query” é guardado na tabela (que anteriormente no fluxo de controlo tinha sido apagada) “TBL_PRE_DIM_PATIENT” através da tarefa “OLE DB Destination”, renomeada para “TGT SA DIM TBL”.

Regressando ao fluxo de controlo, as próximas tarefas de inserção da linha de paciente não definido (sem código atribuído) através da tarefa “execute SQL task” (“INSERT ND ROW”), e o carregamento, através da “data flow task” renomeada para “LOAD DIM DATA” (Figura 60) são praticamente semelhantes ao carregamento das dimensões anteriores.

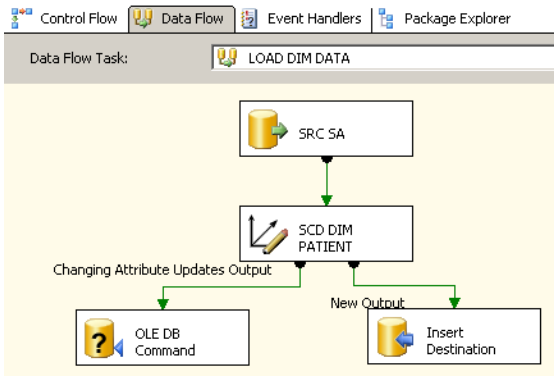


Figura 60 – Fluxo de dados para carregamento na dimensão paciente.

Uma das diferenças no carregamento da dimensão paciente consiste na fonte de dados para a tarefa “slowly changing dimension”. Na vez de um ficheiro Excel, é uma tabela (“TBL_PRE_DIM_PATIENT”) da SA. A outra diferença consiste nos campos/atributos a ter em linha de conta para serem actualizados, ou seja, neste caso como não se tem um descritivo

(nome) do paciente. As características (colunas) passíveis de actualização são o seu sexo, data de nascimento e peso à nascença (Figura 61).

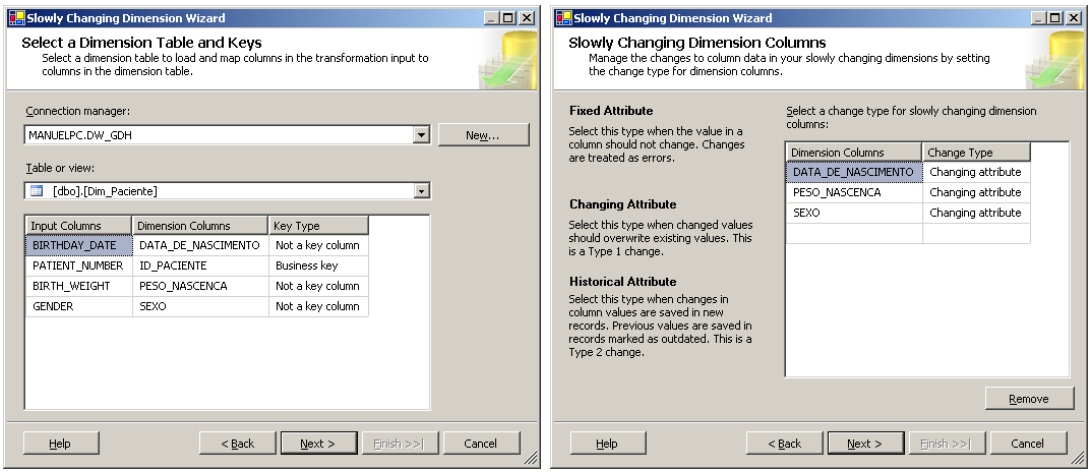


Figura 61 – Parametrizações da "slowly changing dimension" da dimensão paciente.

Também se pode dar o caso de o profissional de saúde se ter enganado ao preencher os dados do paciente, por exemplo, a sua data de nascimento. Caso esse cenário aconteça, a dimensão ao ser reprocessada, o registo do paciente que já existia será actualizado para a sua data de nascimento correcta. Este é mais um motivo pelo qual se optou por uma "slowly changing dimension" do tipo 1, visto que um paciente apenas pode ter uma data de nascimento.

11.4.5.6 Carregamento da Dimensão Time

Para o carregamento da dimensão time (tempo) houve a necessidade de criar um projecto de Analysis Services, que é uma das componentes do BIDS para criação de cubos. Neste caso específico, o projecto de Analysis Services será usado para criar a dimensão time na base de dados DW_GDH, ou seja, no DW.

O primeiro passo para a criação da dimensão time consistiu em criar um projecto do tipo de Analysis Services (Figura 62 e Figura 63):

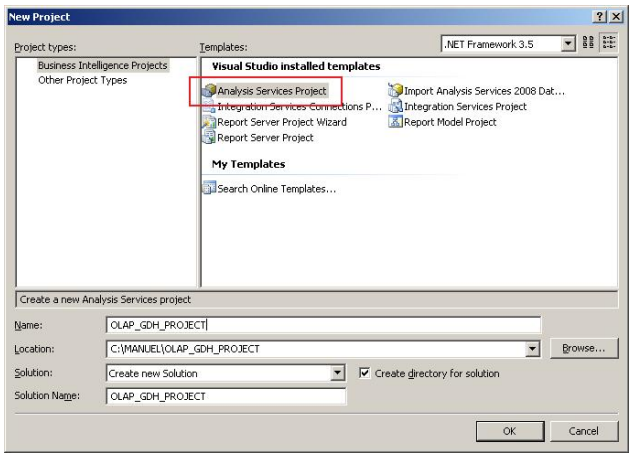


Figura 62 – Criação de um projecto de Analysis Services (parte 1).

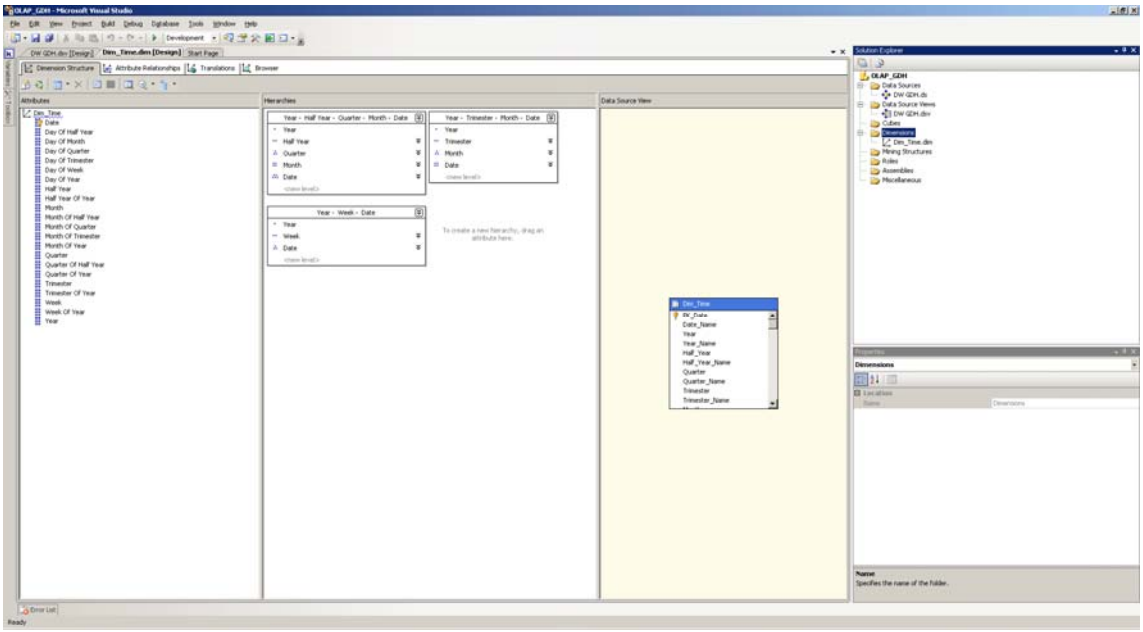


Figura 63 – Criação de um projecto de Analysis Services (parte 2).

O passo seguinte consistiu em adicionar ao projecto uma dimensão nova (Figura 64):

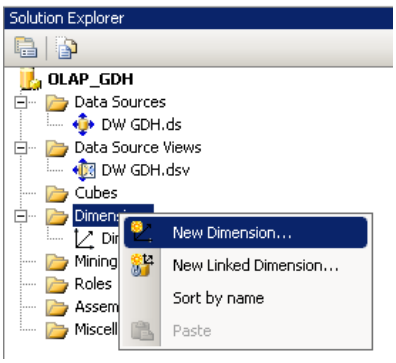


Figura 64 – Criação de uma nova dimensão no projecto de Analysis Services.

Seguidamente escolheu-se a opção “Generate a time table in the data source” (Figura 65) em se definiu como “data source” a base de dados DW_GDH, ou seja, é nesta BD que se pretendia criar com a dimensão time.

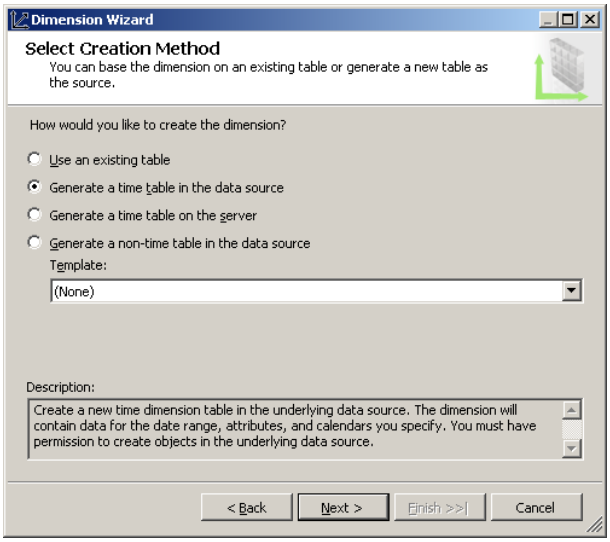


Figura 65 – Opção de gerar uma tabela de tempo na base de dados DW_GDH.

O passo seguinte consistiu em definir a data mínima e máxima, assim como a granularidade da dimensão time (Figura 66).

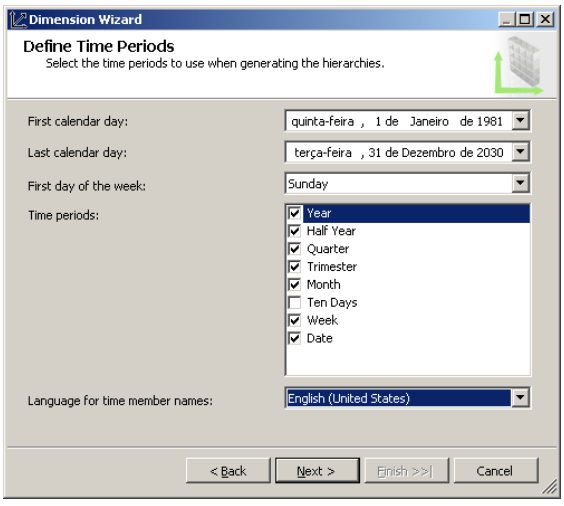


Figura 66 – Selecção da data mínima, máxima e granularidade da dimensão time.

Optou-se pela data de mínima de 1981, visto que foi o primeiro ano em Portugal em que os GDH´s foram aplicados. A data máxima de 2030, visto que nesta altura este sistema estará,

possivelmente, desactualizado. Quanto à granularidade, escolheu-se vários períodos de forma a permitir que o utilizador final analise os dados por dia, mês, trimestre, semestre e ano.

O próximo passo consistiu em gerar a dimensão time (Figura 67).

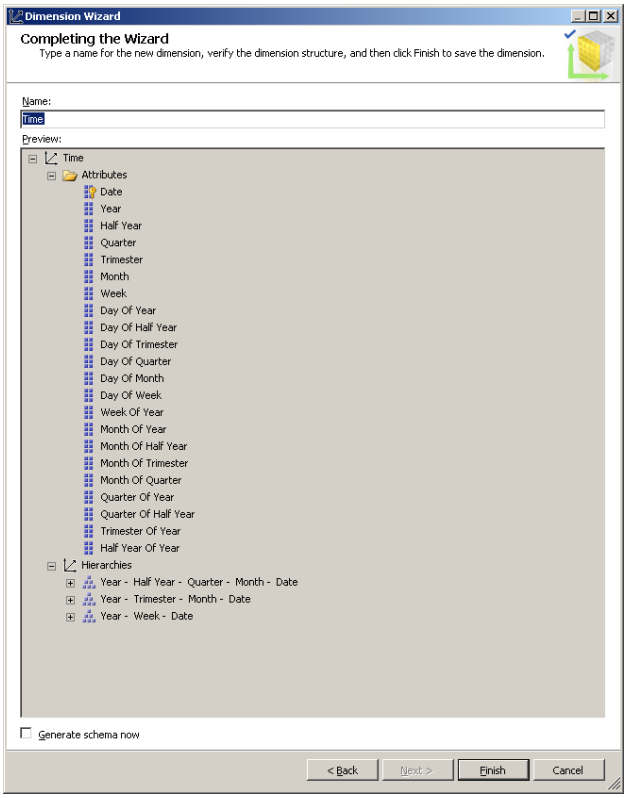


Figura 67 – Processo de geração da dimensão time.

Por último efectuou-se uma alteração no relacional (base de dados DW_GDH) que consistiu em transformar a chave primária criada por "default" (coluna PK_Date) com granularidade até ao dia e do tipo data (no formato AAAA-MM-DD 00:00:00.000) para outra coluna nova, denominada SK_DATE, do tipo inteiro e no formato (AAAAMMDD), Figura 68.

	PK_Date	SK_DATE
1	1981-01-01 00:00:00.000	19810101
2	1981-01-02 00:00:00.000	19810102
3	1981-01-03 00:00:00.000	19810103
4	1981-01-04 00:00:00.000	19810104
5	1981-01-05 00:00:00.000	19810105
6	1981-01-06 00:00:00.000	19810106
7	1981-01-07 00:00:00.000	19810107
8	1981-01-08 00:00:00.000	19810108
9	1981-01-09 00:00:00.000	19810109
10	1981-01-10 00:00:00.000	19810110

Figura 68 – Comparação entre a coluna default "PK_Date" e a coluna adicionada "SK_DATE".

Desta forma, o cruzamento de informação através de chaves do tipo inteiro é muito mais rápido do que qualquer outro tipo de dados. E quando falamos em grandes volumetrias de dados, como é o caso de um DW, esta técnica aumenta a performance de acesso aos dados.

11.4.5.7 Carregamento das Dimensões Distrito, Freguesia e Concelho

As dimensões distrito, freguesia e concelho encontram-se definidas como um ramo do modelo do DW, ou seja, estas dimensões estão representadas no modo de "snow-flake" em que a hierarquia mais elevada é o distrito, depois o concelho e o último nível (nível folha) é a freguesia. A estrutura das tabelas é a seguinte (Figura 69):

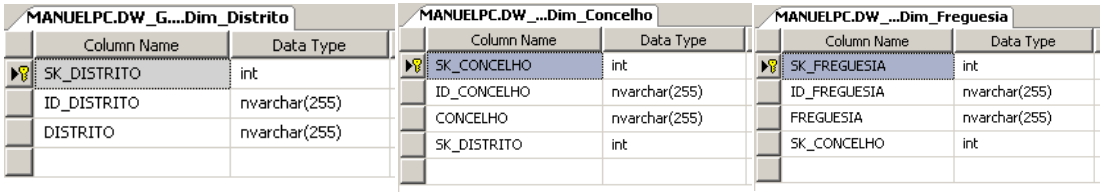


Figura 69 – Estrutura das dimensões distrito, concelho e freguesia.

Visto que as três dimensões estão dependentes da mesma fonte de dados (ficheiro Excel "SRC_FREGS_CONCELHO_DISTR.xlsx"), criou-se um "package" designado por "GDH_LOAD_DIM_RESIDE.dtsx" para o carregamento destas dimensões. Os dados no ficheiro fonte encontravam-se não normalizados o que obrigou a uma transformação específica para construir uma hierarquia. Optou-se por construir uma hierarquia de forma a que os dados estejam mais organizados/normalizados (minimizando a replicação de dados) e acima de tudo que facilite a navegação ("drill") ao utilizador final. Desta forma, os utilizadores podem navegar da freguesia do paciente até ao seu distrito e vice-versa, o que em termos de BI, traduz-se nas operações de "drill-down" e "drill-up". Esta flexibilidade permite aos utilizadores analisarem os dados sob diferentes perspectivas de granularidade.

Para compreender o processo de ETL que alimenta estas dimensões, é necessário observar um excerto de como é que os dados se encontram no ficheiro fonte (Figura 70).

DTCCFR	FREGUESIA	DTCC	CONCELHO	DT	DISTRITO
111301	A DOS CUNHADOS	1113	TORRES VEDRAS	11	LISBOA
100601	A DOS FRANCOS	1006	CALDAS DA RAINHA	10	LEIRIA
101201	A DOS NEGROS	1012	OBIDOS	10	LEIRIA
131301	A VER-O-MAR	1313	POVOA DE VARZIM	13	PORTO
030864	ABACAO	0308	GUIMARAES	03	BRAGA
171401	ABACAS	1714	VILA REAL	17	VILA REAL
030201	ABADE DE NEIVA	0302	BARCELOS	03	BRAGA
031201	ABADE DE VERMOIM	0312	VILA NOVA DE FAMALICAO	03	BRAGA
030401	ABADIM	0304	CABECEIRAS DE BASTO	03	BRAGA
040701	ABAMBRES	0407	MIRANDELA	04	BRAGANCA
160401	ABEDIM	1604	MONCAO	16	VIANA DO CASTELO
150901	ABELA	1509	SANTIAGO DE CACEM	15	SETUBAL
141601	ABITUREIRAS	1416	SANTAREM	14	SANTAREM
101501	ABIUL	1015	POMBAL	10	LEIRIA
130101	ABOADELA	1301	AMARANTE	13	PORTO
141301	ABOBOREIRA	1413	MACAO	14	SANTAREM
030701	ABOIM	0307	FAFE	03	BRAGA
130102	ABOIM	1301	AMARANTE	13	PORTO

Figura 70 – Estrutura das dimensões Distrito, Concelho e Freguesia.

Como se pode observar na figura acima, a chave operacional do distrito é unívoca, a do concelho é uma chave composta em que os dois primeiros dígitos pertencem ao distrito. A chave da freguesia, também ela composta, tem nos quatro primeiros dígitos os códigos do distrito (dois dígitos) e do concelho (dois dígitos).

Após a compreensão da estrutura do ficheiro fonte, definiu-se o fluxo de controlo do processo de ETL (Figura 71).

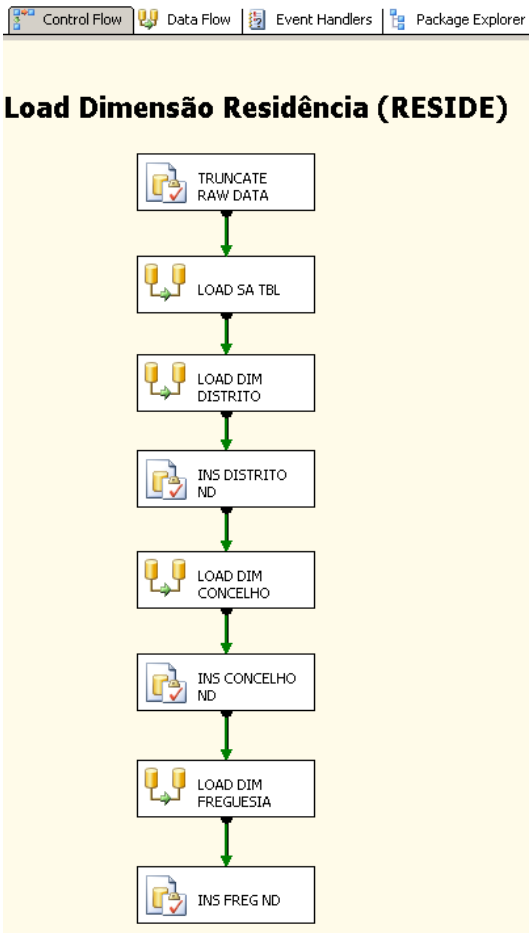


Figura 71 – Controlo de fluxo do processo de ETL para carregamento das três dimensões.

O primeiro passo ("TRUNCATE RAW DATA") consiste em apagar todos os registos de uma tabela intermédia ("TBL_RAW_DATA_RESIDE") que foi criada na Staging Area para importar os dados directamente da fonte para se ter independência do sistema fonte. O próximo passo ("LOAD SA TBL") consistiu no carregamento do ficheiro Excel para a tabela da Staging Area. Após este passo, a tarefa de fluxo de dados ("data flow task") renomeada para "LOAD DIM DISTRITO" consiste em carregar a dimensão distrito segundo o fluxo de dados (Figura 72):

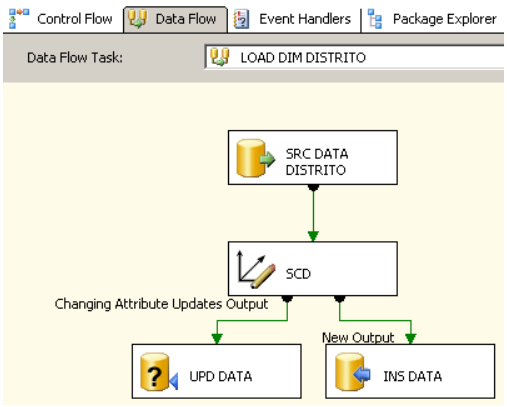


Figura 72 – Fluxo de dados para carregamento da dimensão distrito.

Em que a "query" fonte à tabela de Staging Area consiste apenas nos códigos e descritivos distintos do distrito, removendo-se o cabeçalho da coluna, visto que na fonte existem repetições (Figura 73).

```
select distinct DT,DISTRITO from  dbo.TBL_RAW_DATA_RESIDE
where DISTRITO not in ('DISTRITO/ILHA')
and DISTRITO is not null
order by DT
```

Figura 73 – "Query" de extracção para carregamento da dimensão distrito.

A tarefa de "slowly chaging dimension" permite identificar se já existe o distrito, o seu descritivo é actualizado (tarefa "UPD DATA") caso tenha sido alterado na fonte, se o distrito ainda não existe na dimensão é inserido (tarefa "INS DATA").

Regressando ao fluxo de controlo, o passo seguinte consiste em inserir um registo (caso ainda não exista) do distrito não definido em que a SK assume o valor "-1".

Para o carregamento dos concelhos teve-se em conta a dimensão distrito, visto que um distrito pode ter "n" concelhos. Partindo deste pressuposto, o carregamento da dimensão concelho é idêntico ao distrito ao nível de "slowly changing dimension", tirando apenas o cruzamento da tabela de Staging Area com a dimensão distrito de forma a que exista a chave estrangeira (SK_DISTRITO) na dimensão de concelho. Só desta forma é que será possível construir a hierarquia. A "query" de extracção para a dimensão concelho é a seguinte (Figura 74):

```
SELECT ID_CONCELHO=ISNULL(A.DTCC,'9999'),
CONCELHO=ISNULL(A.CONCELHO,'NÃO DEFINIDO'),
SK_DISTRITO=ISNULL(DIM.SK_DISTRITO,-1)
FROM (select distinct DTCC,CONCELHO from  dbo.TBL_RAW_DATA_RESIDE
where CONCELHO not in ('CONCELHO') and DTCC is not null) A
left join DW_GDH.dbo.Dim_Distrito DIM
on LEFT(A.DTCC,2)=DIM.ID_DISTRITO
```

Figura 74 – "Query" de extracção para carregamento da dimensão concelho.

No caso de o concelho ser nulo é atribuído o valor 9999, ou seja, não definido. Caso o descritivo do concelho seja igual à palavra "concelho" que no sistema fonte é o cabeçalho do ficheiro Excel que se repete algumas vezes pelo ficheiro e tem de ser eliminado visto que não representa nenhum concelho, mas sim uma repetição do cabeçalho. O cruzamento da

informação entre a tabela de SA que tem todos os dados e a dimensão distrito é efectuado através de um "left join" visto que se pretende carregar todos os concelhos tendo ou não distrito associado.

Como para a dimensão distrito, efectuou-se a mesma tarefa para a dimensão concelho que consistiu em inserir um registo (caso ainda não exista) do concelho não definido em que a SK assume o valor "-1".

Por último surge a dimensão freguesia que é idêntica à dimensão concelho. A única diferença consiste na "query" de extracção (Figura 75) que vai buscar ao SK do concelho (chave estrangeira) à dimensão concelho.

```
SELECT ID_FREGUESIA=ISNULL(A.DTCCFR,'999999'),
FREGUESIA=ISNULL(A.FREGUESIA,'NÃO DEFINIDO'),
SK_CONCELHO=ISNULL(DIM.SK_CONCELHO,-1)
FROM (select distinct DTCCFR,FREGUESIA from  dbo.TBL_RAW_DATA_RESIDE
where FREGUESIA not in ('FREGUESIA') and DTCCFR is not null) A
left join DW_GDH.dbo.Dim_Concelho DIM
on LEFT(A.DTCCFR,4)=DIM.ID_CONCELHO
```

Figura 75 – "Query" de extracção para carregamento da dimensão freguesia.

De notar que o cruzamento é efectuado pelos quatro primeiros dígitos da chave operacional da freguesia que corresponde ao código do concelho.

Através deste encadeamento entre as SK das dimensões foi possível construir a hierarquia das localidades onde os pacientes residem.

11.4.5.8 Carregamento da Dimensão Destino Após Alta

O processo de carregamento da dimensão destino após alta é efectuado de forma diferente em comparação com os processos anteriores.

Começou-se por criar uma tabela de dimensão destino após alta com a seguinte estrutura (Figura 76):

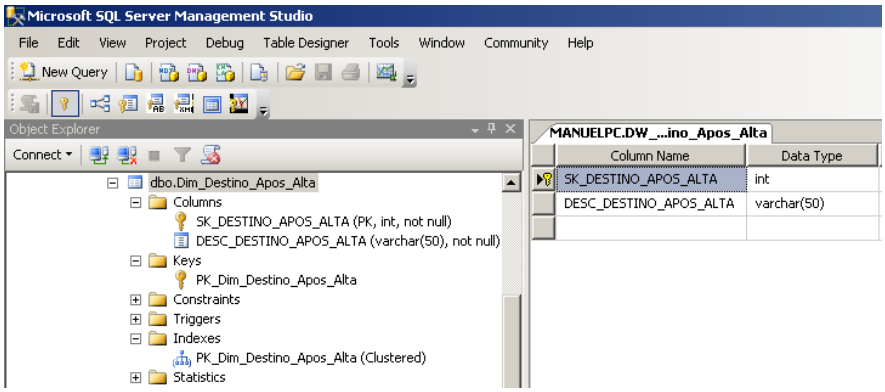


Figura 76 – Estrutura da tabela de dimensão destino após alta.

Seguidamente usou-se a chave operacional como SK, visto que a volumetria desta dimensão é bastante reduzida (apenas contém 7 registos). A fonte de dados é o ficheiro Excel "SRC_DSP_DESTINO_APOS_ALTA.xlsx" e o processo de ETL ("package") que efectua o carregamento da dimensão destino após alta, designa-se por "GDH_LOAD_DIM_DSP.dtsx". A sua estrutura ao nível do controlo de fluxo (Figura 77) consiste em apagar por completo a tabela de dimensão e efectuar um carregamento completo dos dados provenientes da fonte através da tarefa de fluxo de dados "LOAD DIM DATA" (Figura 78).

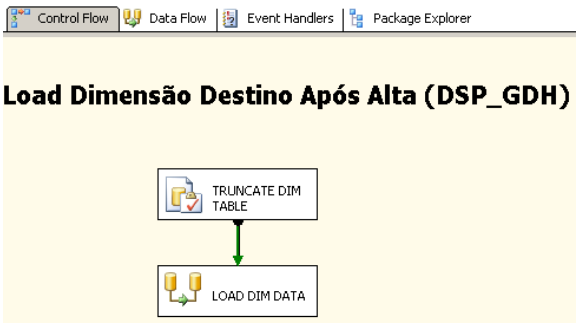


Figura 77 – Componente controlo de fluxo do package "GDH_LOAD_DIM_DSP.dtsx".

Neste caso, não existe uma tarefa de inserção do registo "não definido" visto que o ficheiro Excel já contém esse registo.

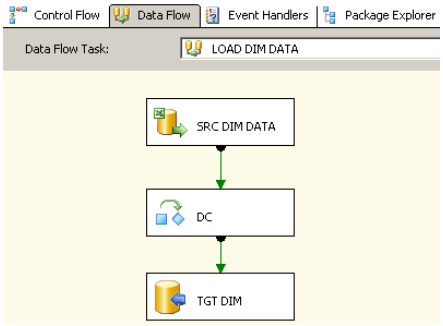


Figura 78 – Componente de fluxo de dados do package "GDH_LOAD_DIM_DSP.dtsx".

11.4.5.9 Carregamento da Dimensão Grandes Categorias de Diagnósticas (GCD)

O processo de carregamento da dimensão GCD, é efectuado de forma idêntica ao da dimensão destino após alta.

Começou-se por criar a tabela de dimensão GCD com a seguinte estrutura (Figura 79):

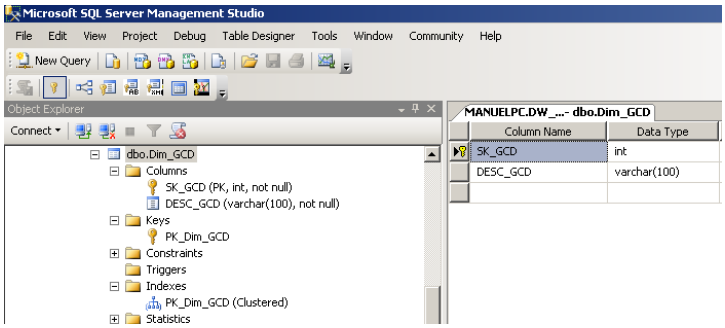


Figura 79 – Estrutura da tabela de dimensão GCD.

Usou-se a chave operacional como SK visto que a volumetria desta dimensão é bastante reduzida (apenas contem 29 registos). A fonte de dados é o ficheiro Excel "SRC_GCD_GRANDE_CATEGORIA_DIAGNOSTICO.xlsx" e o processo de ETL ("package") que efectua o carregamento da dimensão GCD designa-se por "GDH_LOAD_DIM_GCD.dtsx". A sua estrutura ao nível do controlo de fluxo (Figura 80) consiste em apagar por completo a tabela de dimensão e efectuar um carregamento completo dos dados provenientes da fonte através da tarefa de fluxo de dados "LOAD CGD DIM" (Figura 81).

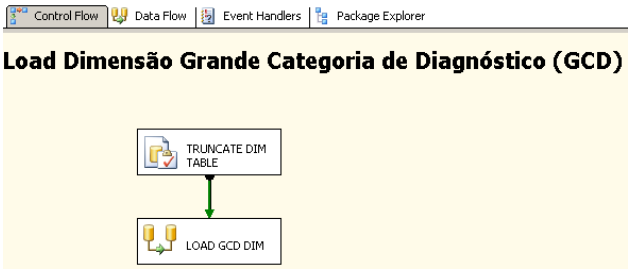


Figura 80 – Componente controlo de fluxo do package "GDH_LOAD_DIM_GCD.dtsx".

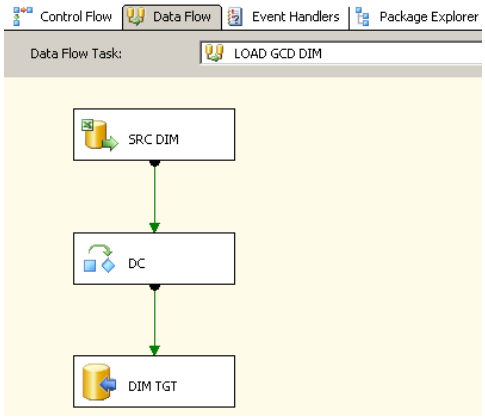


Figura 81 – Componente de fluxo de dados do package "GDH_LOAD_DIM_GCD.dtsx".

No capítulo seguinte iremos falar do carregamento da dimensão GDH que está relacionada com a dimensão GCD, ou seja, uma GCD tem N GDH. No modelo do DW apresentado anteriormente é possível verificar essa hierarquia destas dimensões.

Como é possível verificar no modelo do DW, para além da técnica aplicada do modelo em estrela, também se complementou com a técnica de "snow-flake" para a localidade dos pacientes e para o seu agrupamento/classificação em GDH e GCD.

11.4.5.10 Carregamento da Dimensão Grupo Diagnósticos Homogéneos (GDH)

O processo de ETL (GDH_LOAD_DIM_GDH.dtsx) responsável pelo carregamento da dimensão GDH tem como controlo de fluxo uma tarefa de fluxo de dados ("LOAD GDH DIM") e a inserção de um registo que representa o GDH não definido (Figura 82).

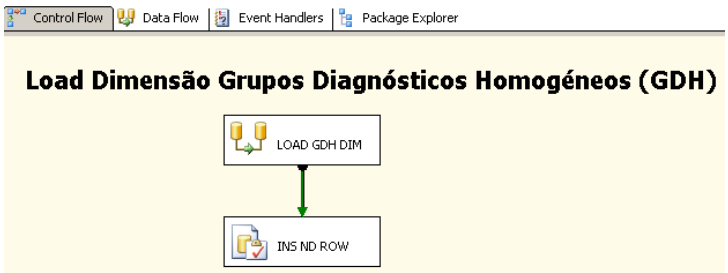


Figura 82 – Componente controlo de fluxo do package "GDH_LOAD_DIM_GDH.dtsx".

Ao nível de fluxo de dados (Figura 83), inicia-se o processo com a extracção dos dados do ficheiro Excel (SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx) que após convertidos para o tipo de dados correcto, atravessam a tarefa de "lookup", renomeada para "LK GCD DIM", que efectua o cruzamento dos registos fonte com a dimensão GCD de forma a determinar a que GCD pertence os dados de GDH's provenientes do ficheiro Excel.

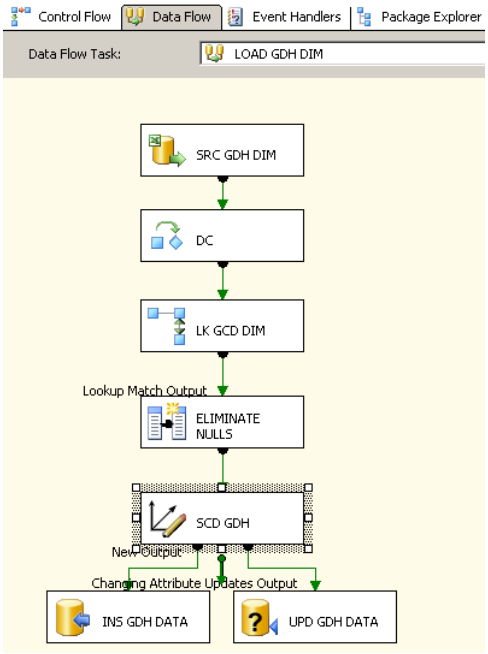


Figura 83 – Componente de fluxo de dados do package "GDH_LOAD_DIM_GDH.dtsx".

O passo seguinte consiste em substituir os registos "nulos" por um com valor (neste caso será zero). Por último, a tarefa de "slowly chaging dimension" actualiza todos os campos da dimensão ou insere novos registos.

A dimensão GDH é constituída por diversos campos (Figura 84) que são utilizados para a fórmula de cálculo do custo de internamento que não nos foi fornecida pela ACSS, no entanto, todos os campos/atributos foram importados para dimensão porque no futuro esta informação poderá ser necessária.

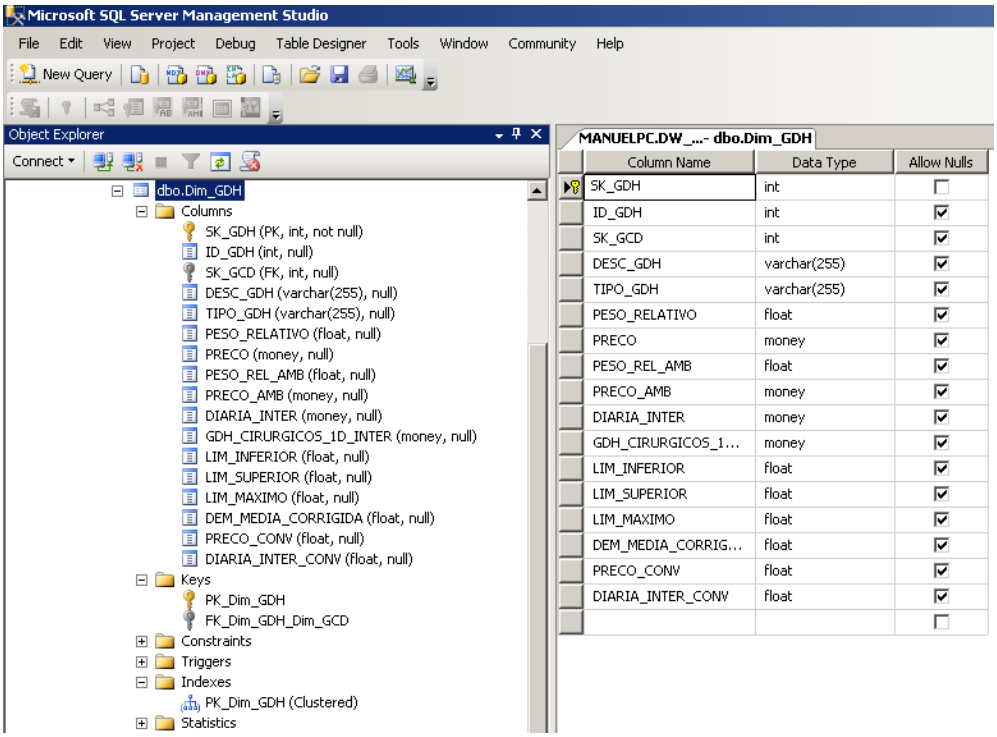


Figura 84 – Estrutura da tabela de dimensão GDH.

11.4.5.11 *Carregamento da Dimensão Tipo de Admissão*

A dimensão tipo de admissão apresenta também uma volumetria reduzida (7 registos), logo optou-se por desenvolver um processo de ETL ("GDH_LOAD_DIM_ADM_TIPO.dtsx") que efectuasse o carregamento total, ou seja, sempre que este processo de ETL é executado, os registos são apagados da dimensão e os dados que estiverem na fonte (ficheiro Excel "SRC_ADM_TIPO_TIPO_DE_ADMISSAO.xlsx") são todos carregados para a dimensão. O processo ao nível de controlo de fluxo (Figura 85) e fluxo de dados (Figura 86) é idêntico ao carregamento da dimensão destino após alta.

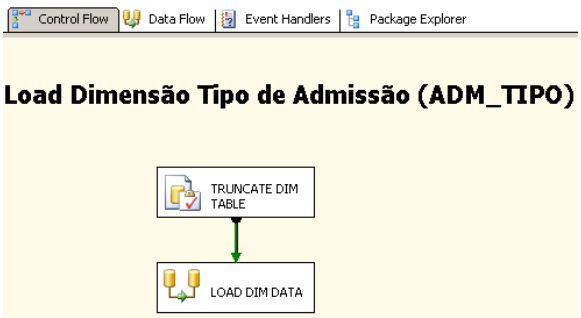


Figura 85 – Componente controlo de fluxo do package "GDH_LOAD_DIM_ADM_TIPO.dtsx".

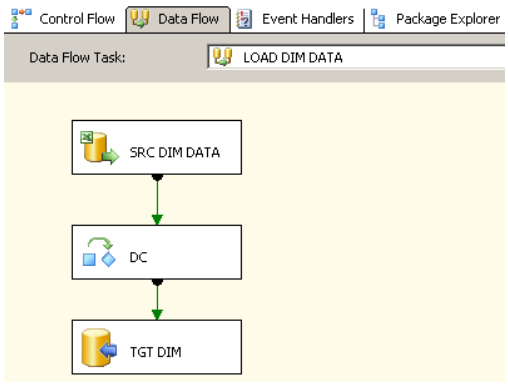


Figura 86 – Componente de fluxo de dados do package "GDH_LOAD_DIM_ADM_TIPO.dtsx".

Tal como o processo de carregamento de ETL do destino após alta, a chave operacional do tipo de admissão é usada como SK da dimensão.

A estrutura da dimensão tipo de admissão é constituída pela SK e o respectivo descritivo (Figura 87).

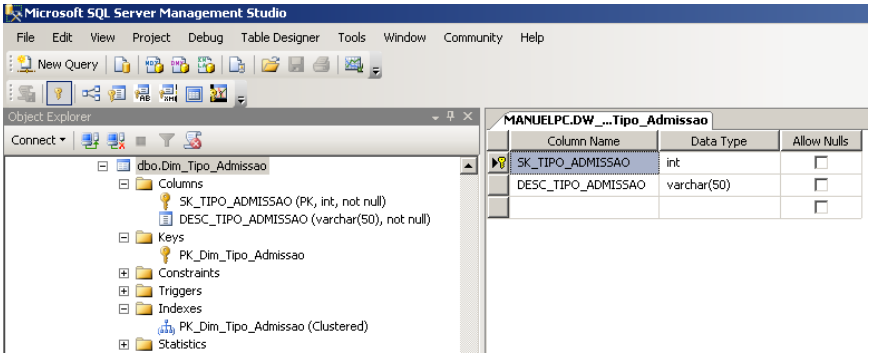


Figura 87 – Estrutura da tabela de dimensão tipo de admissão.

11.4.5.12 Carregamento da Dimensão Motivo de Transferência

O processo de ETL para o carregamento da dimensão motivo de transferência ("GDH_LOAD_DIM_MOT_TRANSF.dtsx") é bastante idêntico aos processos anteriores de carregamento da dimensão destino após alta e tipo de admissão.

No caso do motivo de transferência, a fonte de dados é o ficheiro Excel (SRC_MOT_TRANS_MOTIVO_TRANSFERENCIA.xlsx) e a estrutura da dimensão é a seguinte (Figura 88):

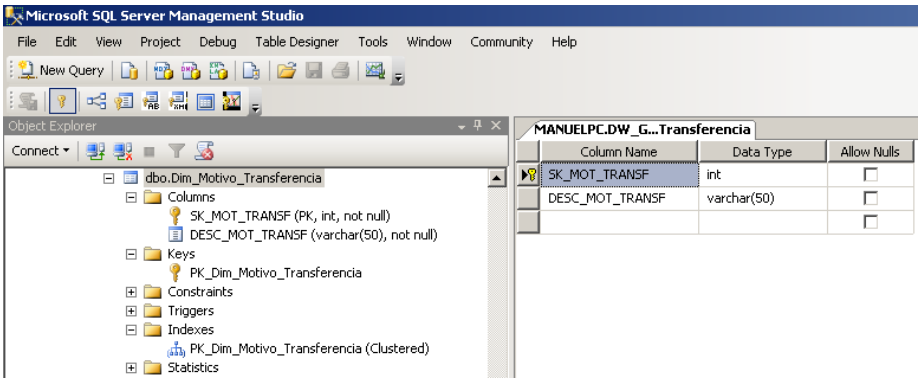


Figura 88 – Estrutura da tabela de dimensão motivo de transferência.

A estrutura de controlo de fluxo (Figura 89) e de fluxo de dados (Figura 90) segue a mesma linha que as dimensões anteriores de volumetria reduzida. Esta dimensão de motivo de transferência possui 6 registos.

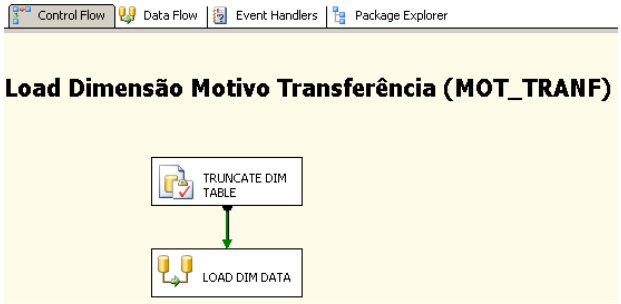


Figura 89 – Componente controlo de fluxo do package "GDH_LOAD_DIM_MOT_TRANSF.dtsx".

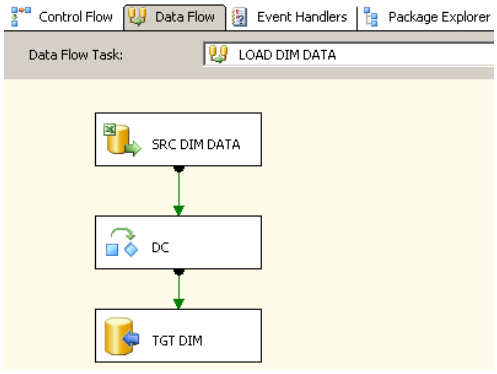


Figura 90 – Componente de fluxo de dados do package "GDH_LOAD_DIM_MOT_TRANSF.dtsx".

11.4.5.13 Carregamento da Dimensão Hospital

O carregamento da dimensão Hospital efectua-se através do processo ETL "GDH_LOAD_DIM_HOSPITAL.dtsx", em que o controlo de fluxo (Figura 91) consiste em inserir

um registo de Hospital não definido através da tarefa “execute SQL task”, renomeada para “INS ND ROW”, e a tarefa de fluxo de dados, renomeada para “LOAD DIM HOSPITAL”, trata de carregar os dados do ficheiro Excel fonte (“SRC_HOSPITAL_ID.xlsx”) para a dimensão.

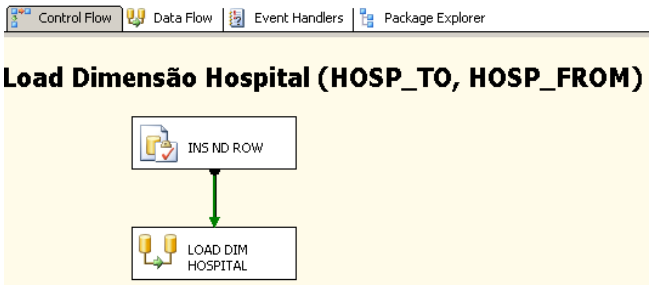


Figura 91 – Componente controlo de fluxo do package “GDH_LOAD_DIM_HOSPITAL.dtsx”.

Em relação ao fluxo de dados (Figura 92), são extraídos os dados do ficheiro fontes, depois convertidos para o tipo de dados de acordo com dimensão Hospital (tabela final) e após a conversão dos dados, estes passam pela tarefa de “slowly changing dimension”, renomeada para “SCD”, de forma a que os registos novos são inseridos directamente na tabela de dimensão, enquanto que os registos já existentes são actualizados caso o seu descritivo tenha mudado.

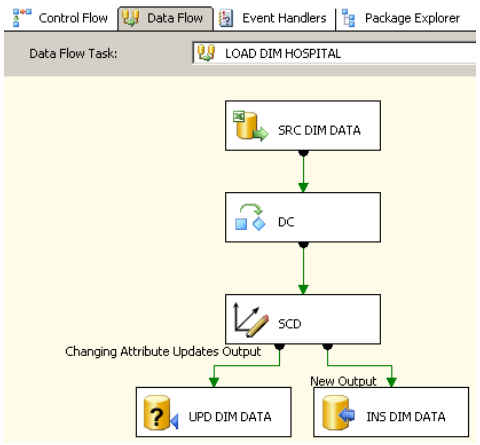


Figura 92 – Componente de fluxo de dados do package “GDH_LOAD_DIM_HOSPITAL.dtsx”.

A estrutura da tabela de dimensão Hospital (Figura 93) apresenta como SK (“SK_HOSPITAL”) uma chave incremental (“identity”), código (“ID_HOSPITAL”) e o descritivo (“DESC_HOSPITAL”) operacional que são provenientes da fonte de dados.

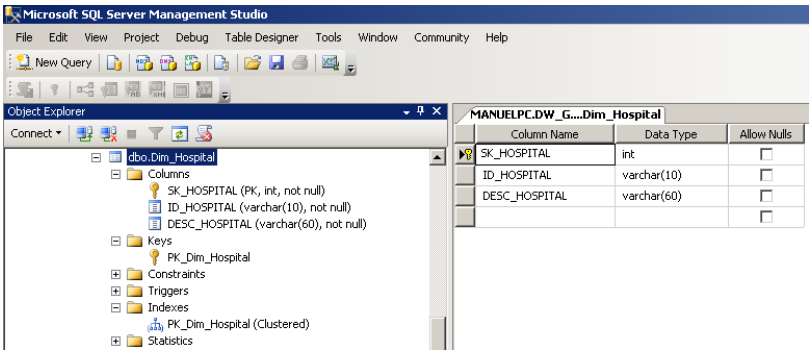


Figura 93 – Estrutura da tabela de dimensão Hospital.

11.4.5.14 Carregamento da Dimensão Sazonalidade

A dimensão sazonalidade assim como a dimensão tempo será estática, visto que as estações do ano são imutáveis, ou seja, nunca mudam (pelo menos num futuro próximo). Partindo deste pressuposto, criou-se na Staging Area uma tabela designada "TBL_SEASON" que possui 4 registos (um para cada estação do ano) com um identificador único ("ID_SEASON"), o descrito da estação do ano ("SEASON_DESC") que assume os valores Inverno, Primavera, Verão e Outono, o mês ("BEGIN_MONTH") e o dia ("BEGIN_DAY") de inicio da respectiva estação (Figura 94).

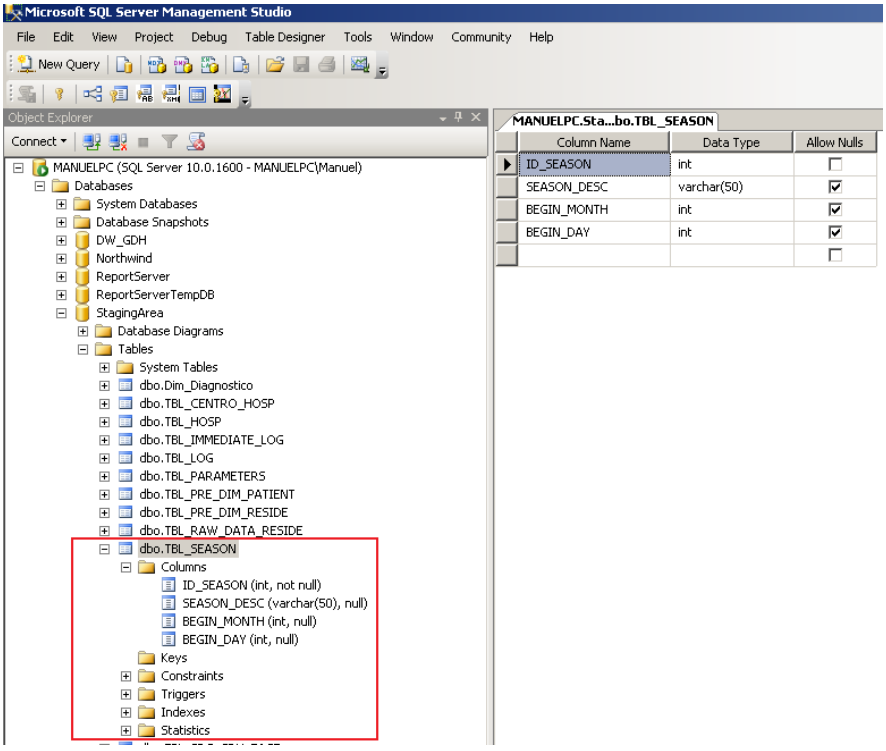


Figura 94 – Estrutura da tabela "TBL_SEASON".

As datas de inicio das estações do ano foram retiradas da fonte (Wikipédia, 2010). As datas inseridas na tabela "TBL_SEASON" foram as seguintes (Figura 95):

	ID_SEASON	SEASON_DESC	BEGIN_MONTH	BEGIN_DAY
1	1	INVERNO	12	21
2	2	PRIMAVERA	3	21
3	3	VERÃO	6	21
4	4	OUTONO	9	23

Figura 95 – Datas de inicio das estações do ano inseridas na tabela "TBL_SEASON".

Após definida a tabela fonte, criou-se um processo de ETL ("GDH_LOAD_DIM_SEASON.dtsx") que ao nível de controlo de fluxo (Figura 96) começa por apagar todos os registos da dimensão sazonalidade através da tarefa "execute SQL task", renomeada para "TRUNCATE DIM TABLE", e seguidamente carrega os dados da tabela fonte, através da tarefa de "fluxo de dados" (renomeada para "LOAD DIM DATA"), na dimensão sazonalidade.

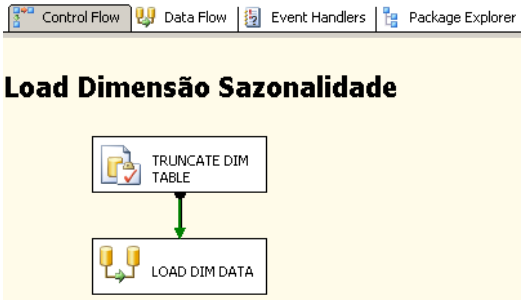


Figura 96 – Componente controlo de fluxo do package "GDH_LOAD_DIM_SEASON.dtsx".

A componente de fluxo de dados (Figura 97) do processo de ETL extrai os dados da tabela fonte ("TBL_SEASON") e coloca-os no destino, ou seja, na tabela de dimensão sazonalidade.

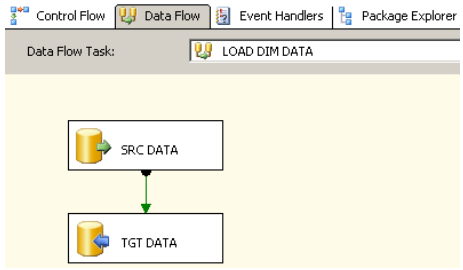


Figura 97 – Componente de fluxo de dados do package "GDH_LOAD_DIM_SEASON.dtsx".

11.4.5.15 Carregamento da Dimensão Serviço

Para a dimensão serviço optou-se por um carregamento total, embora a sua volumetria não seja assim tão insignificante em comparação com as restantes dimensões que são carregadas de forma total. O volume de dados desta dimensão situa-se nos 10.266 registos.

A opção de carregamento total foi tomada visto que para identificar um serviço de forma unívoca é necessário cruzar a informação fonte através de uma chave composta (ID_HOSPITAL e ID_SERVICO) visto que o mesmo serviço (código e descritivo) é igual para Hospitais diferentes, ou seja, apenas a chave do Hospital é que permite descobrir, sem duplicação de registos, os serviços de cada Hospital. Neste cenário e para manter a sincronização entre o identificador do Hospital e Serviço optou-se pelo carregamento total. No ficheiro de Excel criou-se uma coluna incremental que representa a SK do Serviço e que no DW identifica de forma unívoca os serviços.

Desta forma, o processo de ETL ("GDH_LOAD_DIM_SERV.dtsx") para o carregamento da dimensão serviço tem como fonte o ficheiro Excel "SRC_SERV_SERVIÇOS .xlsx" e ao nível de controlo de fluxo (Figura 98) começa por apagar na totalidade os dados da dimensão serviço e seguidamente executa a tarefa de carregamentos de dados. A última tarefa consiste em actualizar os registos inseridos na dimensão (a coluna ID_SERVICO) de forma a remover as pelicas a mais.

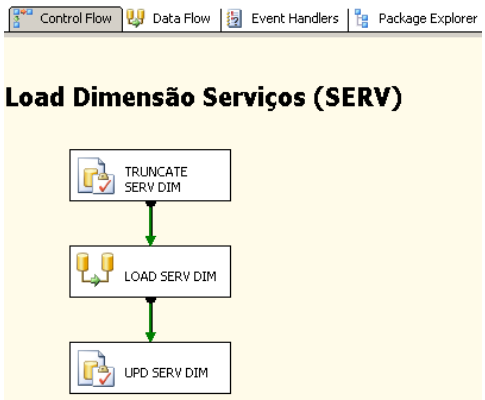


Figura 98 – Componente controlo de fluxo do package "GDH_LOAD_DIM_SERV.dtsx".

Quanto ao fluxo de dados (Figura 99), tarefa "LOAD SERV DIM", o fluxo de dados consiste em extrair os dados da fonte (ficheiro excel) e carregá-los directamente na dimensão serviço (tabela destino).

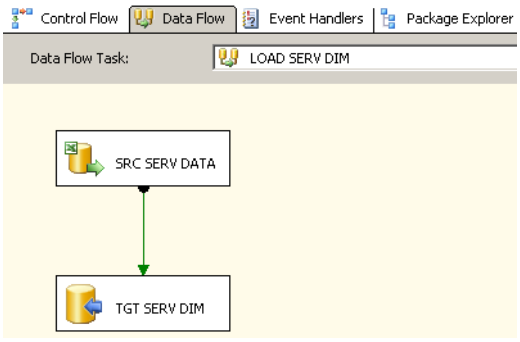


Figura 99 – Componente de fluxo de dados do package "GDH_LOAD_DIM_SERV.dtsx".

A dimensão serviço (tabela destino) tem a seguinte estrutura (Figura 100):

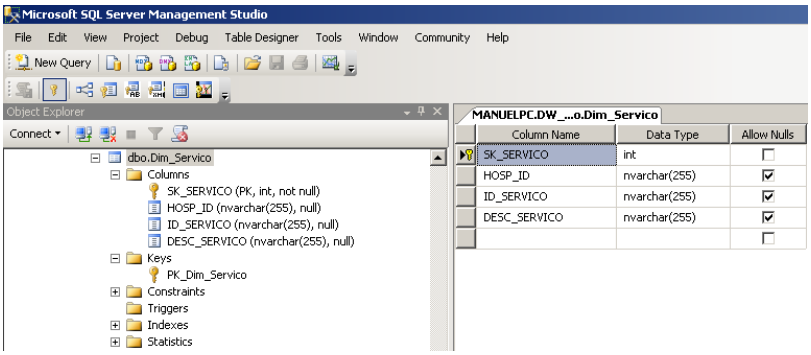


Figura 100 – Estrutura da tabela de dimensão serviço.

11.4.6Quinta Fase – Construção e carregamento dos Factos

O processo de ETL ("GDH_LOAD_SA.dtsx") que é responsável pelo carregamento dos factos apresenta uma complexidade maior visto que foi necessário efectuar várias transformações (passo a passo) aos dados fonte.

A fonte de dados como referido anteriormente, consistem em ficheiros *.dbf em que a sua estrutura pode ser revista no capítulo de "Primeira Fase – Análise dos Requisitos".

O primeiro passo consistiu em criar uma ligação dinâmica aos ficheiros (*.dbf), que desta vez será directa, ou seja, não será necessário passar por ficheiros Excel. O segredo consistiu em definir como fornecedor ("provider") da ligação o "Microsoft Jet 4.0 OLE DB Provider" e como "Extended Properties" a opção "dBase 5.0" (Figura 101). A partir deste momento estabeleceu-se a ligação à directoria onde residem os ficheiros que irão ser carregados na tabela de factos.

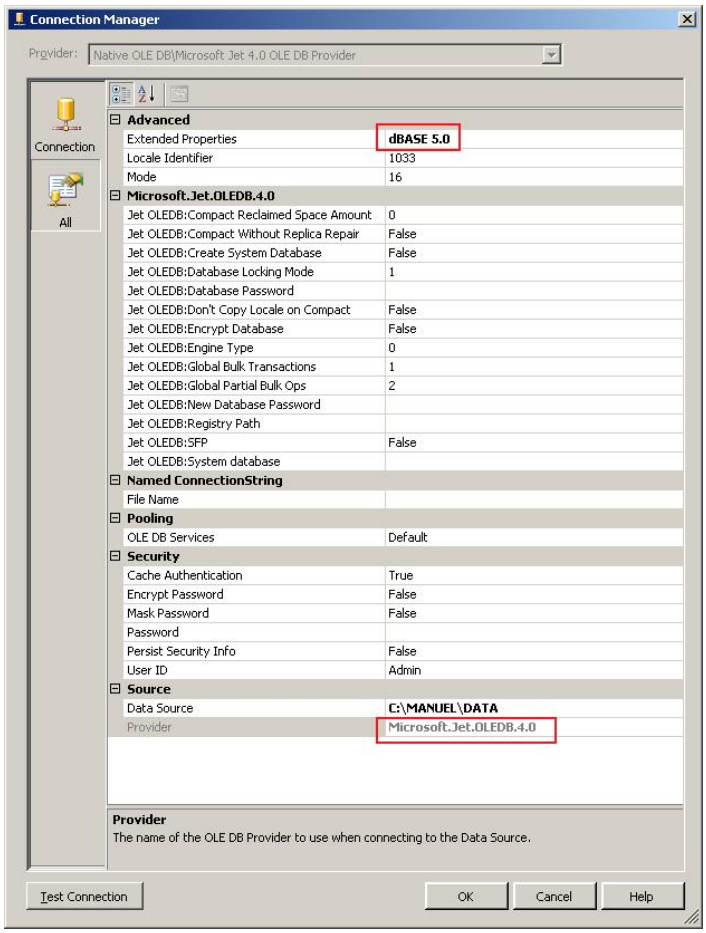


Figura 101 – Ligação aos ficheiros *.dbf.

Posto isto, construiu-se um fluxo de controlo (

Figura 102) que na primeira parte carrega os parâmetros necessários para a execução correcta do "package", ou seja, nesta primeira tarefa ("LOAD PARAMETERS") são lidos dados da tabela "TBL_PARAMETERS" da Staging Area dados como a directoria fonte onde estão os ficheiros *.dbf (que poderá ser alterada no futuro e desta forma só será necessário efectuar a alteração da directoria fonte apenas num sitio), a directoria destino dos ficheiros após o seu carregamento, a lista de emails a enviar, o servidor smtp e o remetente do email. Todos estes parâmetros são guardados em variáveis que são usadas ao longo do "package".

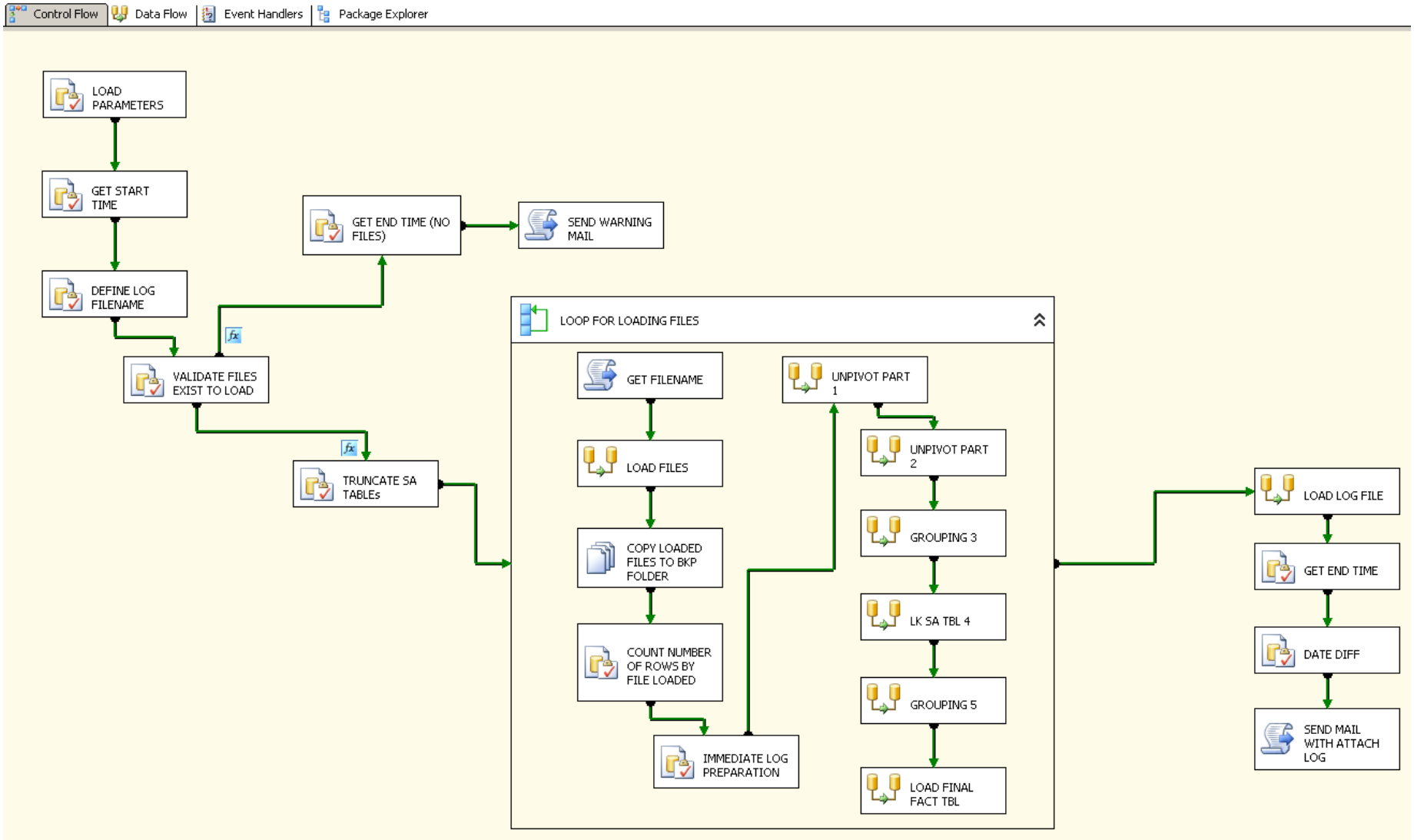


Figura 102 – Controlo de fluxo do "package" GDH_LOAD_SA.dtsx.

O passo seguinte consiste em guardar numa variável (“START_TIME”) a data de inicio do “package” através da tarefa “GET START TIME”. Seguidamente, a tarefa “DEFINE LOG FILENAME” permite definir dinamicamente o nome do ficheiro com base na data (até ao segundo) em que o “package” é executado, o que desta forma impossibilita que existam ficheiros com o mesmo nome como permite também manter um histórico das execuções, ou seja, cada execução deste “package” gera um ficheiro de log. Desta forma, os administradores do sistema de BI conseguem monitorizar de forma simples se o processo correu com sucesso.

Para criar nomes de ficheiros com data, escreveu-se o seguinte código SQL (Figura 103):

```
select 'C:\MANUEL\LOG_FILES\LOG_GDH_LOAD_SA_'+
cast(DATEPART("YY",GETDATE()) as varchar(4))+
cast(DATEPART("MM",GETDATE()) as varchar(2))+
cast(DATEPART("DD",GETDATE()) as varchar(2))+
cast(DATEPART("HH",GETDATE()) as varchar(2))+
cast(DATEPART("MI",GETDATE()) as varchar(2))+
cast(DATEPART("SS",GETDATE()) as varchar(2))+
'.txt' as 'FLATFILE_CONNSTR'
```

Figura 103 – Código SQL para gerar nomes de ficheiros com a data de execução do “package”.

O passo seguinte consiste em verificar, através da tarefa “VALIDATE FILES EXIST TO LOAD”, se existem ficheiros numa directoria específica para carregar (Figura 104).

```
declare @cmd varchar(1000)

create table #File_Exists(s varchar(1000))

select @cmd = 'dir /B C:\MANUEL\DATA'
insert #File_Exists exec master..xp_cmdshell @cmd

if (select COUNT(1) from #File_Exists)>1
        select 1 as 'FILE_EXISTS'
else
        select 0 as 'FILE_EXISTS'
drop table #File_Exists
```

Figura 104 – Código SQL para verificar a existência de ficheiros a carregar numa determinada directoria.

Através da procedimento (“stored procedure”) xp_cmdshell é possível no SQL Server invocar instruções de manipulação de ficheiros. Neste caso, o algoritmo implementado consistiu em listar todos os ficheiros de uma directoria e inseri-los numa tabela temporária. Após esse passo, é efectuada uma contagem ao número de registos da tabela temporária, caso existam devolve-se o registo “1”, caso contrário devolve “0”, com o nome de FILE_EXISTS (“alias”). Do lado do “package” é lido o resultado deste código SQL para uma variável com o mesmo nome (FILE_EXISTS) e caso não existam ficheiros para carregar (FILE_EXISTS==0), o processo de ETL segue o ramo superior em que o próximo passo consiste em obter a data de fim, através da tarefa “GET END TIME (NO FILES)” e por último é enviado um e-mail para os destinatários que se encontram definidos na tabela “TBL_PARAMETERS”. Para o envio de e-mail tentou-se utilizar uma tarefa (“send mail task”) específica do SQL Server para esta finalidade (Figura 105), no entanto, tem a limitação de apenas possibilitar “windows authentication” e o pretendido seria uma autenticação no servidor de smtp com outras credenciais, diferentes da autenticação “windows”.

Figura 105 – “Send mail task” do SQL Server.

Desta forma não foi possível utilizar a “send mail task” e para contornar o problema, partiu-se, uma vez mais, para o desenvolvimento à medida através da “script task”, renomeada para “SEND WARNING MAIL”. Esta tarefa permitiu construir na linguagem de programação “Visual Basic” código para o envio de mail com autenticação no servidor de smtp. O mail enviado para os destinatários é o seguinte caso não existam ficheiro para carregar (Figura 106):

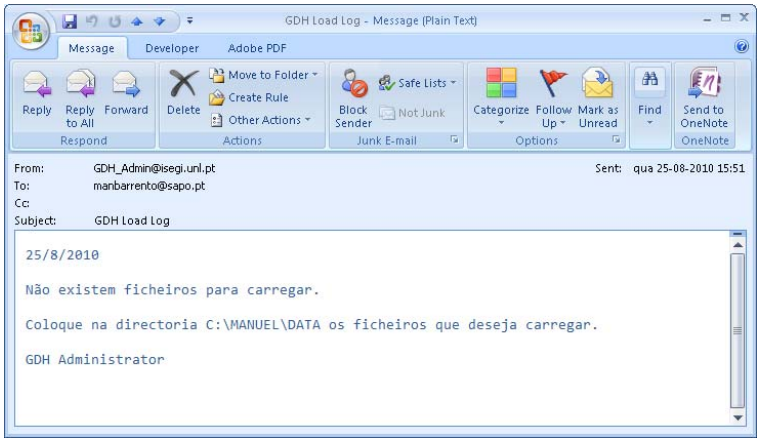


Figura 106 – Mail enviado aos destinatários.

No caso de existirem ficheiros para carregar, o fluxo de dados é outro. O passo seguinte consiste apagar, através da tarefa "TRUNCATE SA TABLEs", a tabela intermédia "TBL_SRC_GDH_FACT" que irá possuir o resultado da importação directa dos ficheiros e a tabela "TBL_IMMEDIATE_LOG" que irá conter o "log" do processo.

Após este passo, o processo entra em ciclo ("loop") que só termina quando não existirem ficheiros para carregar na directoria fonte.

A primeira tarefa pertencente ao ciclo consiste, através de código C#, em obter o nome do ficheiro a carregar. Seguidamente, o conteúdo do ficheiro é carregado na tabela intermédia "TBL_SRC_GDH_FACT" da SA atráves do fluxo de dados "LOAD FILES" (Figura 107).

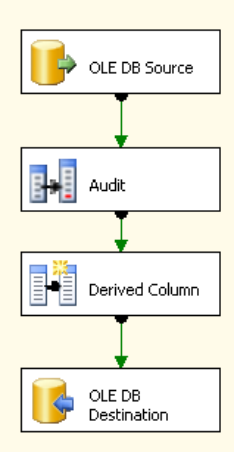


Figura 107 – Carregamento do conteúdo do ficheiro dbf para a tabela "TBL_SRC_GDH_FACT".

É de salientar que a tarefa de auditar ("Audit") consiste em acrescentar duas colunas, uma com o nome do utilizador que executou o processo e a outra com a data em que foram inseridos os registos na tabela "TBL_SRC_GDH_FACT". A tarefa de coluna derivada ("Derived Column") acrescenta mais uma coluna à tabela com o nome ficheiro a qual pertencem os registos.

Regressando ao controlo de fluxo do "package" de carregamento dos factos, o passo seguinte consiste numa tarefa de "file system" que consiste em copiar o ficheiro carregado para uma directoria de "backup". Seguidamente, são contados os registos inseridos na tabela "TBL_SRC_GDH_FACT" e o valor é guardado numa variável ("NUM_REG") do "package". O próximo passo consiste, através da tarefa "IMMEDIATE LOG PREPARATION", inserir na tabela de log, denominada "TBL_IMMEDIATE_LOG", o nome do ficheiro e o número de registos carregados.

Terminado o pré-processamento dos dados, ou seja, independência da fonte de dados, inicia-se uma outra fase que consiste na transformação dos dados na SA. A abordagem seguida consistiu em efectuar passo a passo as transformações em vez de tudo em uma só vez. Tomou-se esta opção devido às limitações das capacidades de processamento do computador em que o sistema foi desenvolvido. Em servidores de grande potência (com muita memória RAM e velocidade de processador) talvez fosse possível juntar alguns dos passos desenvolvidos neste processo ("package") de ETL. O primeiro passo da transformação dos dados (tarefa "UNPIVOT PART 1") consiste em transformar de colunas para linhas o diagnóstico principal (da coluna DDX1 à DDX20), os procedimentos (da coluna SRG1 à SRG20) e a causa externa (da coluna CAUSAD1 à CAUSAD20). A estas três colunas correspondem a cada 20 colunas originais. Esta transformação foi conseguida através de nova funcionalidade do SQL Server 2008 que consiste na instrução "UNPIVOT", Figura 108.

```
SELECT [ANO]
      ,[CODE]
      ,[NUMERO]
      ,[SEXO]
      ,[B_DATE]
      ,[FIN_RESP]
      ,[RESIDE]
      ,[DISTRITO]
      ,[CONCELHO]
      ,[FREGUESIA]
      ,[SERV1]
      ,[ENT1]
      ,[SAID1]
      ,[SERV2]
      ,[ENT2]
      ,[SAID2]
      ,[SERV3]
      ,[ENT3]
      ,[SAID3]
      ,[SERV4]
      ,[ENT4]
      ,[SAID4]
      ,[SERV5]
      ,[ENT5]
```

,[SAID5]

,[SERV6]

,[ENT6]

,[SAID6]

,[SERV7]

,[ENT7]

,[SAID7]

,[SERV8]

,[ENT8]

,[SAID8]

,[SERV9]

,[ENT9]

,[SAID9]

,[SERV10]

,[ENT10]

,[SAID10]

,[SERV11]

,[ENT11]

,[SAID11]

,[SERV12]

,[ENT12]

,[SAID12]

,[SERV13]

,[ENT13]

,[SAID13]

,[SERV14]

,[ENT14]

,[SAID14]

,[SERV15]

,[ENT15]

,[SAID15]

,[SERV16]

,[ENT16]

,[SAID16]

,[SERV17]

,[ENT17]

,[SAID17]

,[SERV18]

,[ENT18]
,[SAID18]
,[SERV19]
,[ENT19]
,[SAID19]
,[SERV20]
,[ENT20]
,[SAID20]
,[ADM_DIAG]
,[ID_DDX]
,[DDX]
,[ID_SRG]
,[SRG]
,[ID_CAUSA.ID_CAUSA]
,[ID_CAUSA.CAUSA_D]
,[MORF_TUM]
,[DSP]
,[BIRTH_WGT]
,[PRE_OP]
,[ICU]
,[ADM_TIP]
,[HOSP_TO]
,[HOSP_FROM]
,[TOTDIAS]
,[AGE]
,[GDH_HCFA16]
,[GDH_AP21]
,[GCD_HCFA16]
,[GCD_AP21]
,[NUM_EPISOD]
,[INTERV_CIR]
,[DSP_GDH]
,[MOT_TRANF]
,[SAIDLAST]
,[HORA_ENTRA]
,[HORA_SAIDA]
,[MODULO]
,[User name]

```
      ,[Execution start time]
      ,[FILE_NAME]
  FROM [StagingArea].[dbo].[TBL_SRC_GDH_FACT]
UNPIVOT
(ID_DDX FOR DDX IN ([DDX1]
      ,[DDX2]
      ,[DDX3]
      ,[DDX4]
      ,[DDX5]
      ,[DDX6]
      ,[DDX7]
      ,[DDX8]
      ,[DDX9]
      ,[DDX10]
      ,[DDX11]
      ,[DDX12]
      ,[DDX13]
      ,[DDX14]
      ,[DDX15]
      ,[DDX16]
      ,[DDX17]
      ,[DDX18]
      ,[DDX19]
      ,[DDX20])) as ID_DDX…
```

Figura 108 – Excerto de uma "query" de transformação dos dados fonte.

O passo seguinte (tarefa "UNPIVOT PART 2") consiste em transformar colunas em linhas, as datas de entrada (colunas ENT1 à ENT20) e saída (colunas SAID1 à SAID20) dos pacientes assim como os respectivos serviços (colunas SERV1 à SERV20). Neste mesmo passo é efectuado o cálculo (linha à linha) do total de dias de internamento, as colunas do tipo data são transformadas para inteiro no formato AAAAMMDD e as colunas Hospital origem (coluna HOSP_FROM) e destino (coluna HOSP_TO) quando vêm uma delas a nulo assumem o valor da sua homóloga, ou seja, caso o Hospital destino venha preenchido e o de origem venha a nulo, assumiu-se que o Hospital onde o paciente permaneceu não mudou. Os dados são posteriormente agrupados (tarefa "GROUPING 3") para eliminar qualquer possibilidade de existência de duplicados. A tarefa "LK SA TBL 4" efectua uma ligação ("left join") entre todos os registos da tabela de factos intermédia e as dimensão do DW. As ligações são efectuadas

através da chave operacional de forma a obter a SK (chave inteira) correspondente. Utilizou-se o "left join" na medidad em que não se pretende perder registos, ou seja, todos os registos que se encontram na tabela intermédia de SA serão carregados na tabela final de factos do DW mesmo que não existindo todos os códigos na dimensão. Para os códigos dos factos que não cruzem com as dimensões é atribuída a SK com o valor -1, que representa o código não definido. Como foi referido anteriormente, os cruzamentos de dados com chaves inteiras apresentam uma maior performance em comparação com chaves que, por exemplo, tenham caracteres na sua constituição. No entanto, a obtenção do valor da SK da estação do ano (coluna SK_SEASON) obedeceu ao desenvolvimento de um algoritmo de forma a classificar em que estação do ano o paciente deu entrada no Hospital. Colocando as estações do ano num formato de horizonte temporal (Figura 109), verifica-se que é possível classificar através de intervalos de datas a estação do ano.

Figura 109 – Horizonte temporal das estações do ano em que as datas consistem em dia e mês.

Visto que as SK's de entrada já são do tipo inteiro, converteu-se também para inteiro as datas das estações do ano (primeiro o mês e depois o dia) e compara-se se o dia e o mês de forma inteira, para se obter o intervalo de datas em que se situa. Caso a data de entrada do paciente não se situe no intervalo da Primavera, Verão ou Outono, então encontra-se no Inverno (Figura 110).

```
   SK_SEASON =
case
        when cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)>=321
                and cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)<621
        then 2
        when cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)>=621
                and cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)<923
        then 3
        when cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)>=923
```

```
                    and cast(SUBSTRING(convert(varchar(8),ENTRADA_YYYYMMDD,112),5,2) +
RIGHT(convert(varchar(8),ENTRADA_YYYYMMDD,112),2) as int)<1221
            then 4
            else 1
end
```

Figura 110 – Transformação de forma a obter a estação do ano em que o paciente deu entrada no Hospital.

O penúltimo passo antes efectuar o carregamento para tabela final de factos, consiste em efectuar um agrupamentos dos dados apenas pelas colunas que são relevantes para o modelo final do DW. Por exemplo, a coluna NUM_EPISOD é descartada visto que não traria uma mais-valia para as análises multidimensionais. A tarefa "LOAD FINAL FACT TBL" insere os registos de uma tabela intermédia de SA para a tabela de factos final do DW. É de referir que todos os passos intermédios dão origem a uma tabela física de SA (que vai desde a tabela TBL_SRC_GDH_FACT_STEP1 à TBL_SRC_GDH_FACT_STEP5).

Na tabela de factos Fact_GDH foram criados índices "unclustered" de forma a obter uma maior performance na consulta aos dados.

Por outro lado, uma grande limitação da estrutura dos dados fonte consistiu no facto da informação ser guardada em coluna, ou seja, um paciente só pode ter no máximo 20 diagnósticos, procedimentos, entre outras características, que são armazenadas no mesmo registo em colunas diferentes e não em registos diferentes. Esta limitação também se repercute no DW.

Após o carregamento dos dados no DW, é gerado, através da tarefa "LOAD LOG FILE", um ficheiro de texto com o "log" do processo de importação dos dados para a SA (Figura 111).

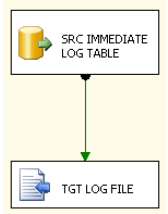


Figura 111 – Fluxo de dados da criação do ficheiro de log.

Os dados que se encontram na tabela de "log" são inseridos no ficheiro de texto que depois é enviado por e-mail como anexo. O nome do ficheiro segue o formato de exemplo

("LOG_GDH_LOAD_SA_20109805742.txt") em o que varia é a data. A estrutura do ficheiro encontra-se na Figura 112.

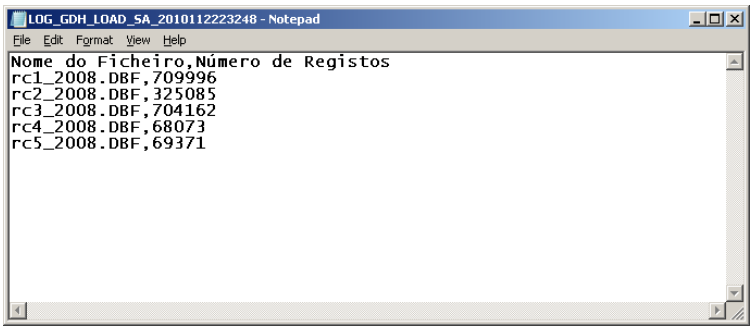


Figura 112 – Estrutura do ficheiro de log.

Após gerado o ficheiro de "log", obtém-se a data de fim do processo de ETL (tarefa "GET END TIME"), depois efectua-se a diferença entre a data de início e de fim para se saber a duração do processo (tarefa "DATE DIFF") e por último programou-se, através da linguagem de programação visual basic, o envio de mail (tarefa "SEND MAIL WITH ATTACH LOG"). A estrutura do mail enviado encontra-se na Figura 113.

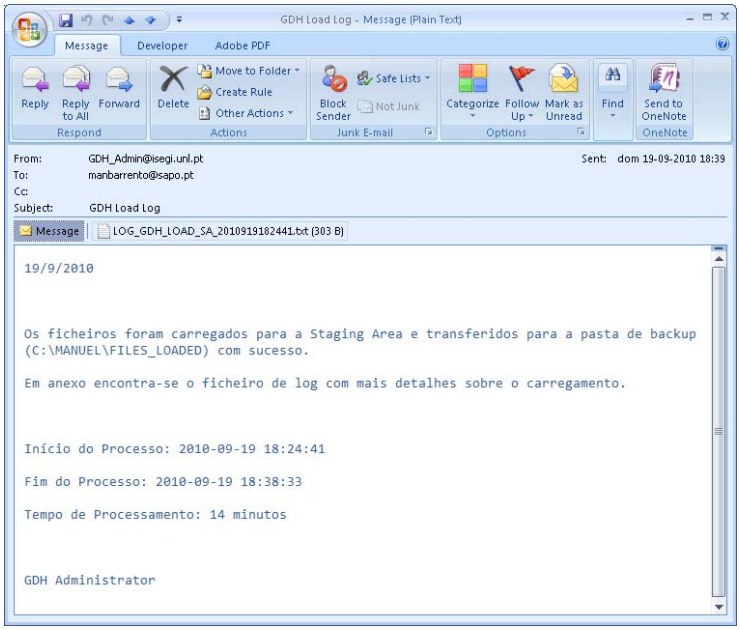


Figura 113 – Estrutura do mail enviado com o log do processo de carregamentos dos factos.

11.4.7Sexta Fase – Reporting

11.4.7.1 Estrutura de camada dupla

Após a componente de ETL, surge a camada de "reporting" que permite aos utilizadores interagirem com o sistema de BI para analisarem os dados. Como foi referido anteriormente, a

plataforma adoptada para o desenvolvimento do "reporting" foi o MicroStrategy, na medida em que reúne um conjunto de mais-valias em termos de utilização, escalabilidade, versatilidade, entre outras que as suas concorrentes ainda não conseguem oferecer. Para além destes pontos positivos, o que mais se destaca é a facilidade com que o utilizador final consegue analisar os dados.

O primeiro passo consistiu em construir um projecto directo (Figura 114) no MicroStrategy Desktop.

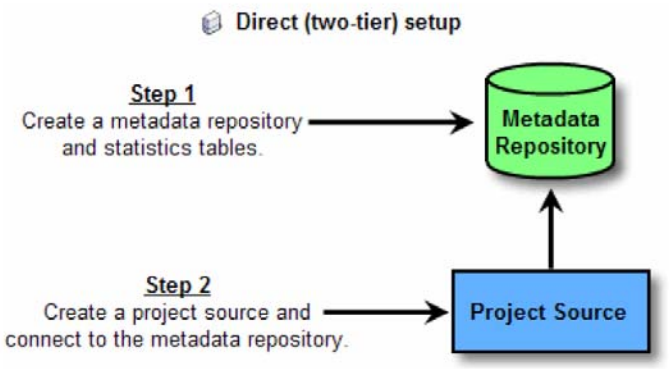


Figura 114 – Arquitectura de um projecto directo em MicroStrategy. Fonte: (MicroStrategy, 2010)**.**

O projecto directo apresenta uma dupla camada que é constituída pelo repositório de metadados e o projecto em si de "reporting". A metadata é obtida através de uma ligação ODBC ao DW dos GDH's.

O MicroStrategy Architect é uma componente do MicroStrategy Desktop onde se selecciona a metadata disponível (tabelas e vistas) do SQL Server, de forma a construir do lado do MicroStrategy o modelo lógico do DW (MicroStrategy, 2009). Começou-se por colocar no "Project Tables View" as tabelas necessárias para a construção do modelo de "reporting" e definiu-se quais os campos das tabelas que seriam atributos e factos (que resultam em métricas) para análise (Figura 115).

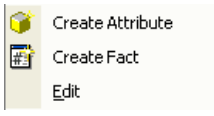


Figura 115 – Definição dos atributos e factos do modelo de "reporting" do MicroStrategy.

Depois definiu-se as ligações entre as tabelas de dimensão e a tabelas de factos. O resultado final consiste num modelo de "reporting" (Figura 116) onde sobre ele irão ser realizadas análises, relatórios e "dashboards".

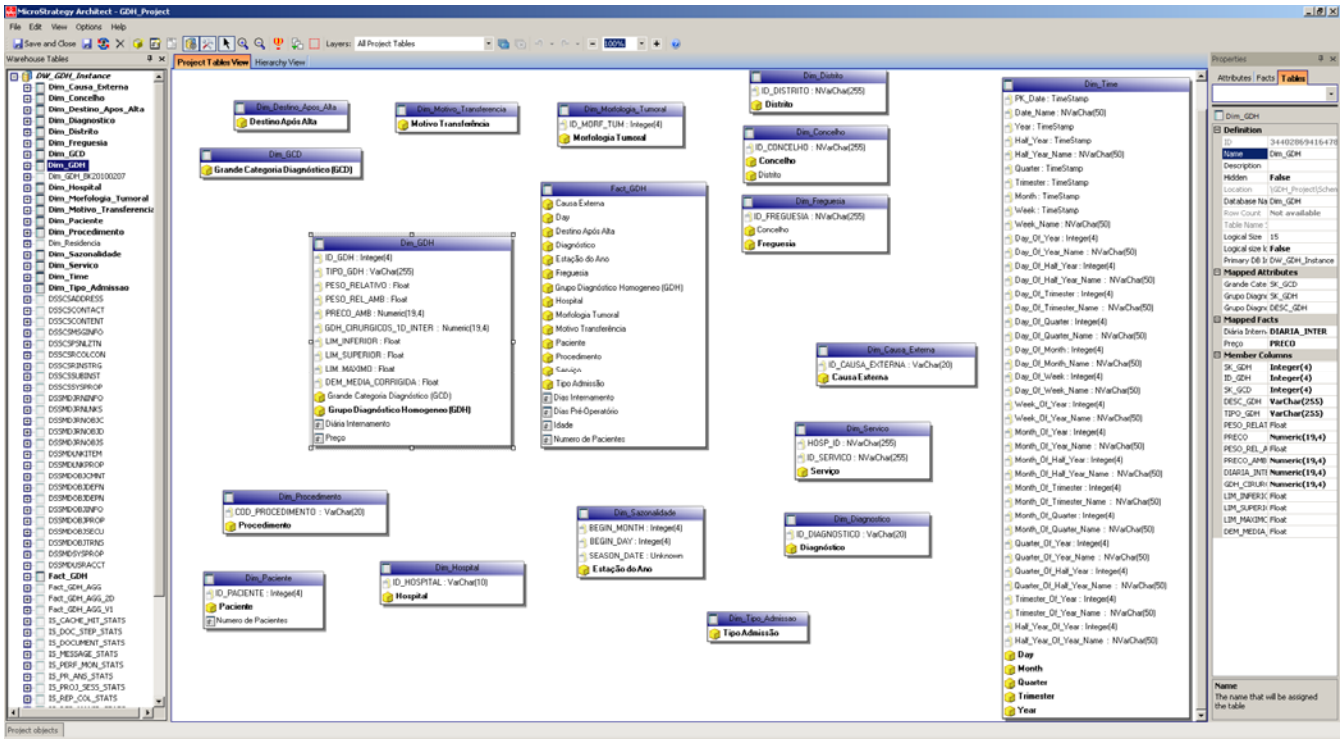


Figura 116 – Modelo de "reporting" do MicroStrategy.

Como foi referido anteriormente, existem algumas dimensões que obedecem ao conceito de "snow-flake", assim sendo, é necessário definir na "Hierarchy View", ou seja, a hierarquia destas dimensões (Figura 117).

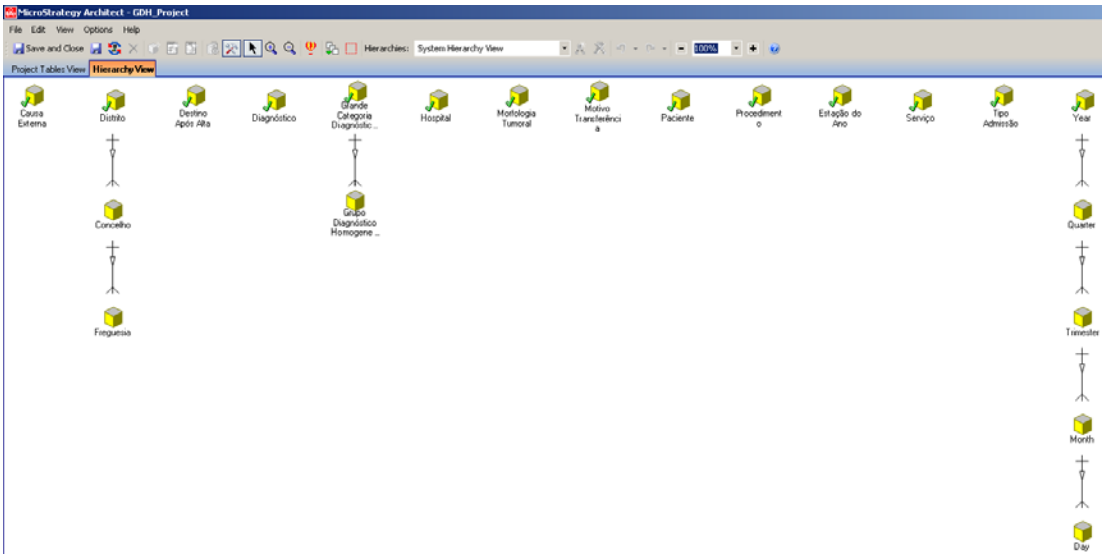


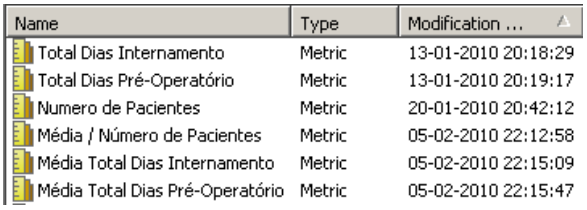
Figura 117 – Construção das hierarquias do modelo de "reporting".

Como se pode observar na figura anterior, foram definidas as hierarquias para a localização do paciente, para a GCD e GDH, em que cada nível da hierarquia corresponde a uma tabela física. Em relação à dimensão tempo apenas existe apenas uma tabela física que contém vários graus

de granularidade desde o ano até ao dia, o que se traduziu na hierarquia do ponto de vista lógico e não físico.

A definição destas hierarquias irá permitir ao motor de SQL do MicroStratey realizar consultas "queries" aos dados de forma mais eficiente. Do ponto de vista do utilizador, irá permitir-lhe navegar por diferentes níveis através da funcionalidade de "drill-up" ou "drill-down" que consiste em começar a analisar os dados por Concelho, depois pode ser necessário ver os dados a um nível mais macro, por exemplo por Distrito ("drill-up") ou ter mais detalhe indo até ao nível da Freguesia ("drill-down").

O passo seguinte consistiu em construir as métricas para análise com base nos campos que foram definidos como factos no MicroStrategy Architect. Definiram-se algumas métricas (Figura 118), no entanto, consoante as necessidades de análise podem ser acrescentadas mais de forma simples e rápida.



Name	Type	Modification ...
Total Dias Internamento	Metric	13-01-2010 20:18:29
Total Dias Pré-Operatório	Metric	13-01-2010 20:19:17
Numero de Pacientes	Metric	20-01-2010 20:42:12
Média / Número de Pacientes	Metric	05-02-2010 22:12:58
Média Total Dias Internamento	Metric	05-02-2010 22:15:09
Média Total Dias Pré-Operatório	Metric	05-02-2010 22:15:47

Figura 118 – Métricas definidas para o "reporting".

Em que a sua fórmula consiste em:

- Total Dias Internamento: Sum([Dias Internamento])
- Total Dias Pré-Operatório: Sum([Dias Pré-Operatório])
- Número de Pacientes: Count(distinct [Numero de Pacientes])
- Média / Número de Pacientes: Avg([Numero de Pacientes])
- Média Total Dias Internamento: Avg([Dias Internamento])
- Média Total Dias Pré-Operatório: Avg([Dias Pré-Operatório])

11.4.7.2 Estrutura de camada tripla

A primeira abordagem do projecto de "reporting" da MicroStrategy consistiu numa camada dupla de forma a que o acesso aos dados seja efectuado de forma directa e com a maior performance possível. Este tipo de projecto revelou-se bastante adequado para utilizadores que constroem os seus próprios relatórios e não pretendem "dashboards". O projecto directo também tem como limitação não permitir construir "dashboards". Esta limitação é

ultrapassada com a criação de um projecto de "reporting" de camada tripla em que a única diferença que existe em comparação com o projecto directo é a camada intermédia, ou seja, o "intelligence server" que é responsável por executar consultas ("queries"), cálculos, gerir os pedidos de consultas dos utilizadores e serve como ponto central da metadata (MicroStrategy, 2010), Figura 119.

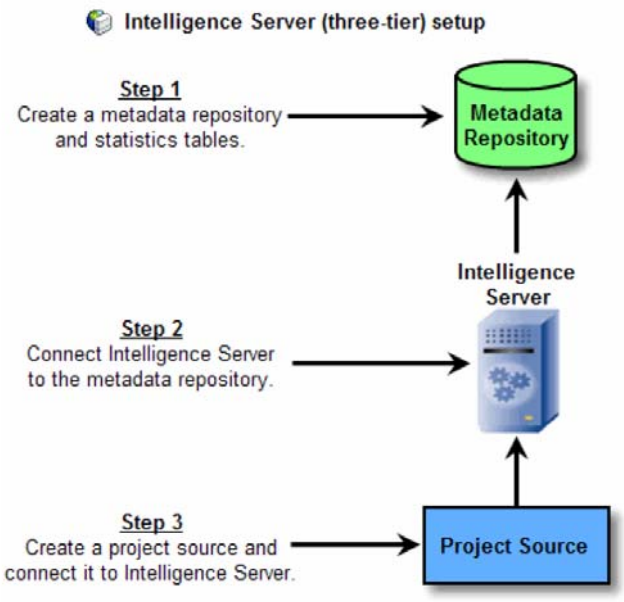


Figura 119 – Arquitectura de camada tripla. Fonte: (MicroStrategy, 2010).

A implementação desta arquitectura é idêntica à implementada para o projecto directo, a única alteração consiste no fluxo de informação que atravessa mais uma camada. A fonte do projecto é a mesma assim como o repositório da metadata. Esta arquitectura só foi implementada devido à necessidade de desenvolver "dashboards".

11.4.7.3 Relatórios

A arquitectura de "reporting" implementada permite três tipos de relatórios. O projecto directo permite a construção "ad-hoc" de relatórios no momento, como relatórios estáticos. Os "dashboards" são semi-dinâmicos, pois o utilizador pode alterar os filtros no momento, no entanto, apenas são disponibilizados na arquitectura de camada tripla.

Após seleccionar-se a opção de criar novo relatório, surge uma janela para o utilizador construir o seu "report" (por medida) com base nos atributos e métricas disponíveis (Figura 120). O utilizador tem apenas que arrastar ("drag and drop") os atributos e métricas que quer analisar para o interior do "Report View", e se pretender também pode trocar os atributos de linha por coluna ou até criar um gráfico. Como complemento ao "reporting", a área de "Report Filter" permite ao utilizador criar um filtro visto que a análise em questão poderá ser, por

exemplo, apenas para o distrito de Lisboa. Não existe qualquer limitação ao número de "reports" criados.

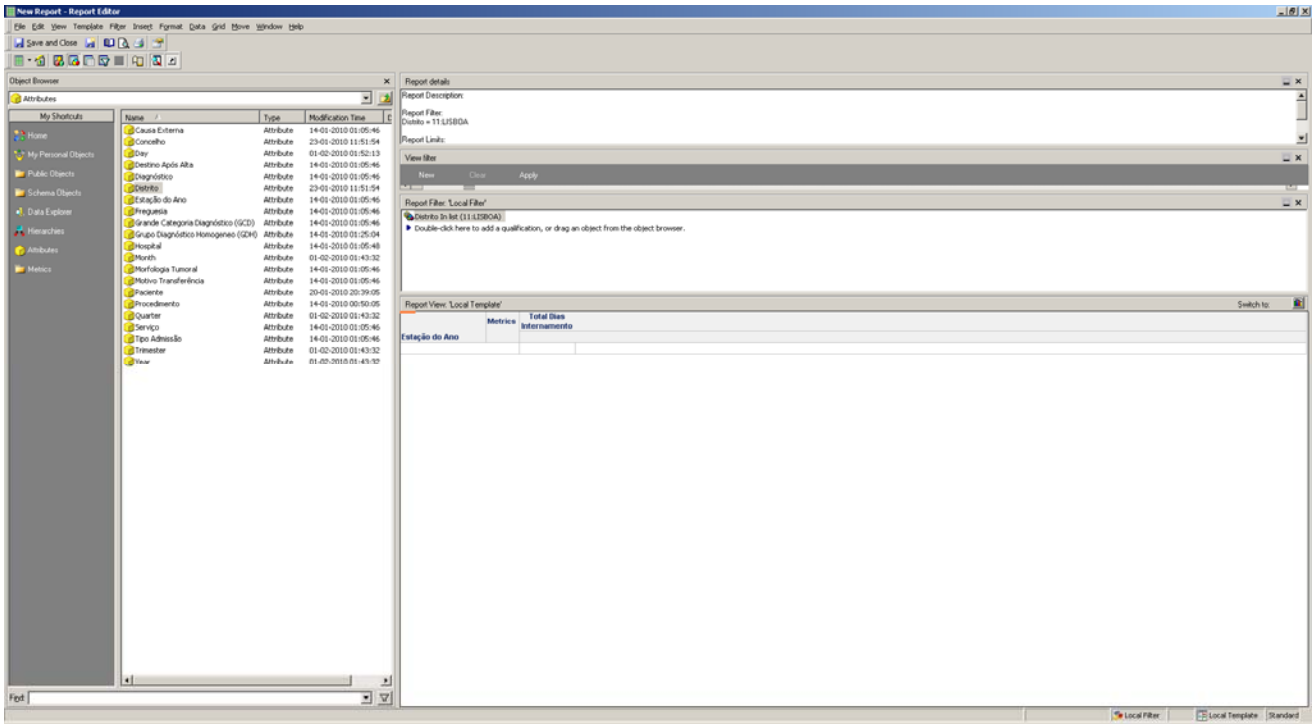


Figura 120 – Ambiente e funcionalidades para a criação de um novo report.

O segundo tipo de relatórios são construídos da mesma forma que o tipo anterior, no entanto, o utilizador final apenas tem a possibilidade de visualizar os resultados. Este relatório apenas podem ser actualizados (refrescados), visto que não podem ser alterados. No entanto, possuem a vantagem de poderem ser exportados para Excel, Word, PDF ou ficheiro HTML e enviados por e-mail a qualquer profissional de saúde que necessite de ter acesso a um respectivo conjunto de dados. É de salientar que a exportação dos relatórios para Excel permite uma grande flexibilidade, visto que podem ser realizados cálculos posteriores com base nos dados exportados. A Figura 121 ilustra um relatório que devolve o top 30 de GDH´s, por estação do ano, com mais número de pacientes.

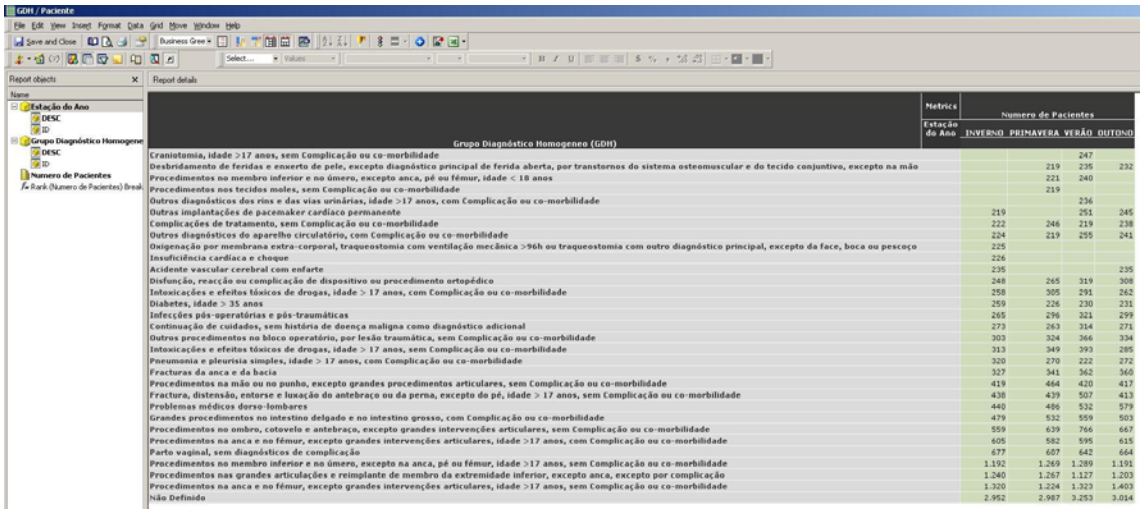


Figura 121 – Report do top 30 de GDH's com mais dias de internamento por estação do ano.

Um outro exemplo de relatório apresenta uma "prompt" antes de ser executado, ou seja, é perguntado ao utilizador final para que meses é que pretende analisar os dados (Figura 122). O utilizador selecciona os meses e o "report" é gerado para os meses seleccionados. Desta forma o horizonte temporal de análise é limitado por parte do utilizador, o que lhe permite alguma autonomia em comparação com o relatório anterior, que é completamente estático, ou seja, não permite aos utilizadores alterar os filtros do relatório.

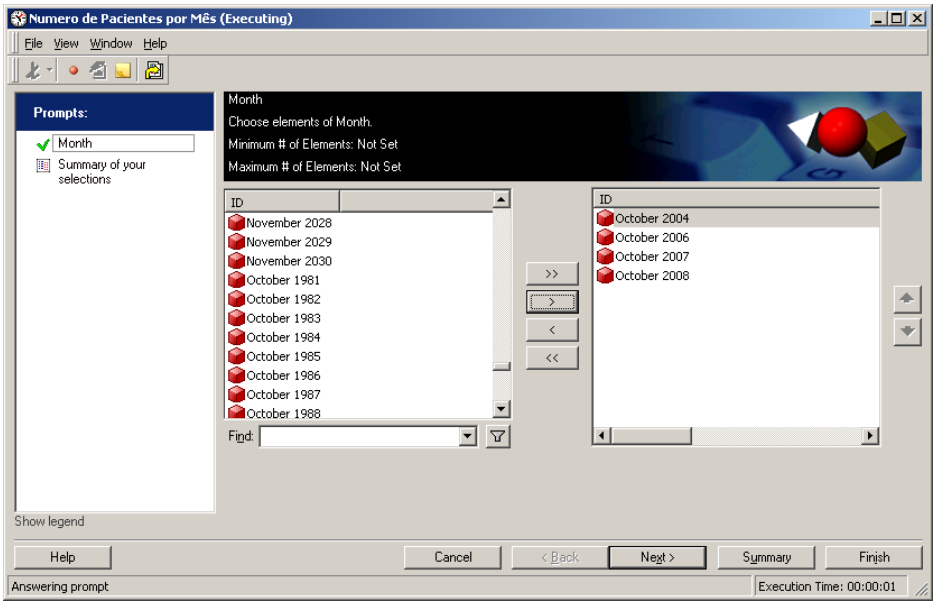


Figura 122 – Prompt para selecção de meses a analisar.

Após selecção da combinação meses/anos a analisar, é gerado um relatório do número de pacientes pelas combinações seleccionadas (Figura 123). Desta forma, é possível efectuar comparações entre períodos homólogos, como estar a par da evolução do número de pacientes (se aumenta ou reduz ao longo dos meses/anos).

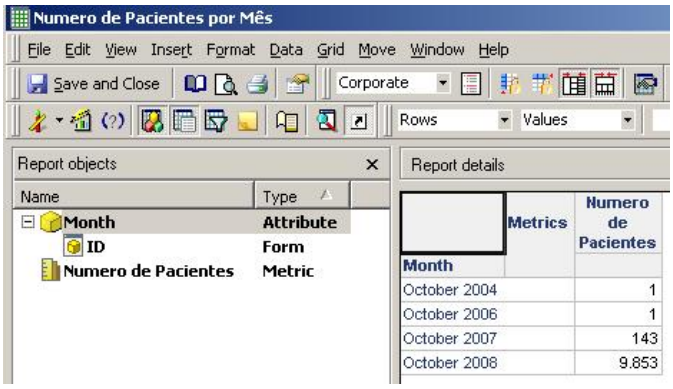


Figura 123 – Report do número de pacientes por Mês.

A componente gráfica do MicroStrategy ,que é bastante apelativa também foi explorada. Assim, criou-se um relatório que consiste num gráfico de barras que compara o total de dias de internamento por GCD entre 2007 e 2008 (Figura 124). Através de uma análise gráfica simples, é possível concluir o aumento de dias de internamento de 2007 para 2008 ao nível das perturbações do sistema nervoso e olhos, entre outras, aumentou.

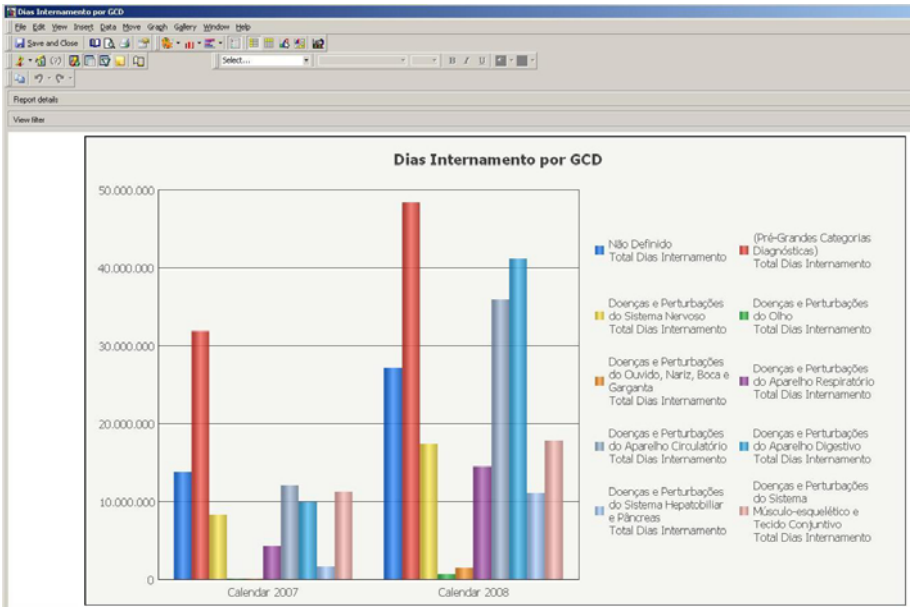
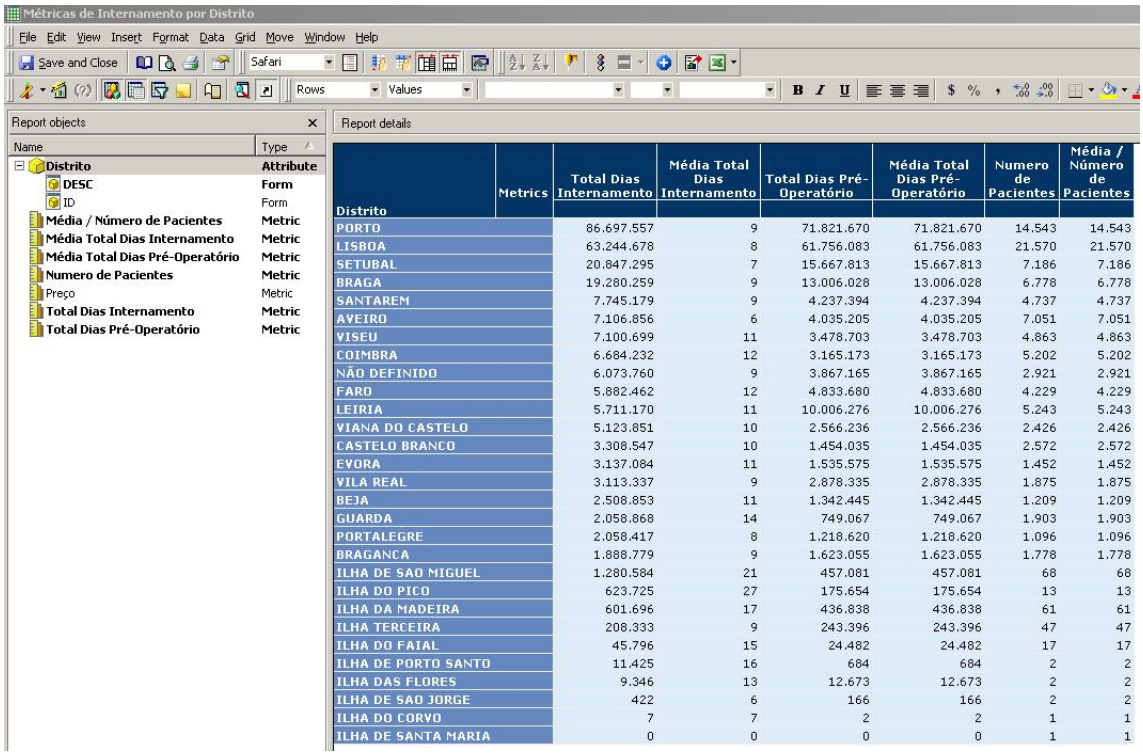


Figura 124 – Report de comparação entre os dias de internamento por GCD em 2007 e 2008.

Um outro exemplo de relatório (Figura 125) consistiu em mostrar de forma tabular, todas as métricas criadas no MicroStrategy que derivam da métrica base "Total Dias de Internamento". O relatório em si mostra, por distrito as métricas desenvolvidas.



Distrito	Metrics	Total Dias Internamento	Média Total Dias Internamento	Total Dias Pré-Operatório	Média Total Dias Pré-Operatório	Numero de Pacientes	Média / Número de Pacientes
PORTO		86.697.557	9	71.821.670	71.821.670	14.543	14.543
LISBOA		63.244.678	8	61.756.083	61.756.083	21.570	21.570
SETUBAL		20.847.295	7	15.667.813	15.667.813	7.186	7.186
BRAGA		19.280.259	9	13.006.028	13.006.028	6.778	6.778
SANTAREM		7.745.179	9	4.237.394	4.237.394	4.737	4.737
AVEIRO		7.106.856	6	4.035.205	4.035.205	7.051	7.051
VISEU		7.100.699	11	3.478.703	3.478.703	4.863	4.863
COIMBRA		6.684.232	12	3.165.173	3.165.173	5.202	5.202
NÃO DEFINIDO		6.073.760	9	3.867.165	3.867.165	2.921	2.921
FARO		5.882.462	12	4.833.680	4.833.680	4.229	4.229
LEIRIA		5.711.170	11	10.006.276	10.006.276	5.243	5.243
VIANA DO CASTELO		5.123.851	10	2.566.236	2.566.236	2.426	2.426
CASTELO BRANCO		3.308.547	10	1.454.035	1.454.035	2.572	2.572
EVORA		3.137.084	11	1.535.575	1.535.575	1.452	1.452
VILA REAL		3.113.337	9	2.878.335	2.878.335	1.875	1.875
BEJA		2.508.853	11	1.342.445	1.342.445	1.209	1.209
GUARDA		2.058.868	14	749.067	749.067	1.903	1.903
PORTALEGRE		2.058.417	8	1.218.620	1.218.620	1.096	1.096
BRAGANCA		1.888.779	9	1.623.055	1.623.055	1.778	1.778
ILHA DE SAO MIGUEL		1.280.584	21	457.081	457.081	68	68
ILHA DO PICO		623.725	27	175.654	175.654	13	13
ILHA DA MADEIRA		601.696	17	436.838	436.838	61	61
ILHA TERCEIRA		208.333	9	243.396	243.396	47	47
ILHA DO FAIAL		45.796	15	24.482	24.482	17	17
ILHA DE PORTO SANTO		11.425	16	684	684	2	2
ILHA DAS FLORES		9.346	13	12.673	12.673	2	2
ILHA DE SAO JORGE		422	6	166	166	2	2
ILHA DO CORVO		7	7	2	2	1	1
ILHA DE SANTA MARIA		0	0	0	0	1	1

Figura 125 – Report de métricas por Distrito no ano de 2008.

O número de relatórios que podem ser criados é ilimitado. Podem construir-se relatórios tabulares, apenas gráficos ou uma junção dos dois como iremos ver mais à frente. O foco nesta fase foi explorar as potencialidades do MicroStrategy. Por exemplo, um relatório interessante e bastante objectivo é o destino após alta dos pacientes, filtrado pelos anos de 2007 e 2008 (Figura 126). Observamos que uma grande parte dos pacientes, após terminarem o seu internamento, regressam ao seu domicílio.

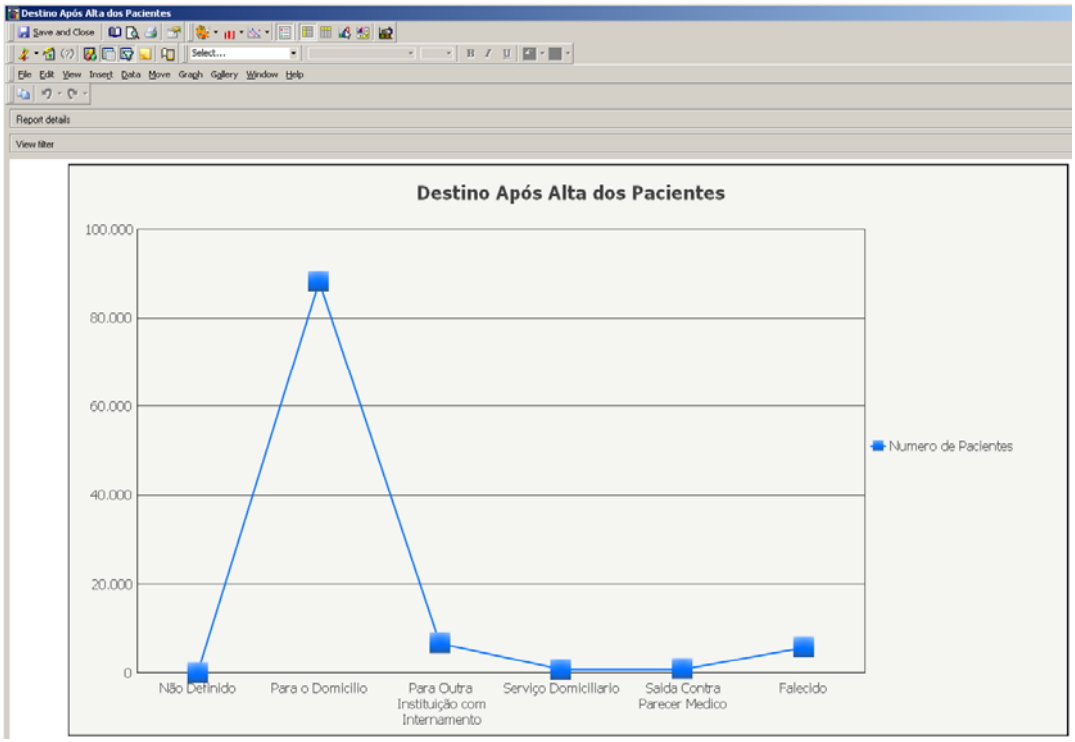


Figura 126 – Report destino após alta dos paciente em 2007 e 2008.

Como referido anteriormente, o relatório "Estação do Ano / Dias Internamento" apresenta uma junção de componente tabular e gráfica (Figura 127). Do lado esquerdo do relatório é possível ver os dados em tabela do total de dias de internamento por estação do ano em 2007 e 2008. Estes dados são representados do lado direito do relatório em forma de gráfico de barras. Como se pode concluir empiricamente, é mais fácil tirar conclusões a partir de um gráfico do que uma tabela. Observando novamente o relatório, conclui-se facilmente através do gráfico que houve mais internamentos no Inverno e na Primavera de 2008 do que em 2007.

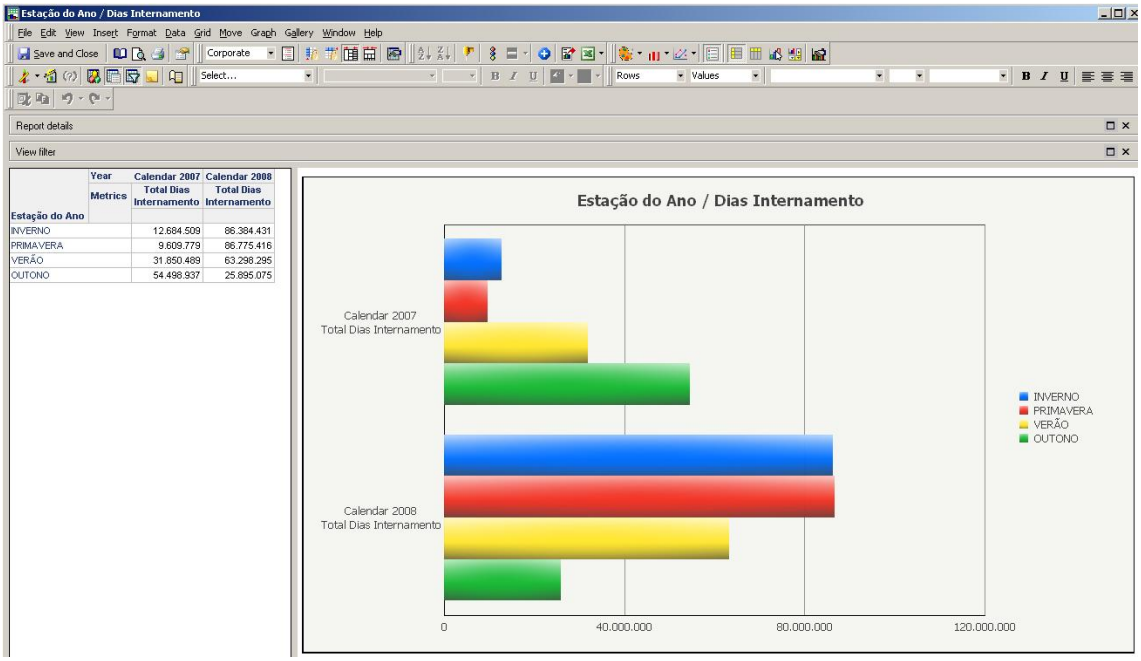


Figura 127 – Report dias de internamento por estação do ano em 2007 e 2008.

11.4.7.4 Dashboards

Os "dashboards" são elaborados tendo como base os relatórios que são acoplados num documento, em que pode ter ou não um filtro dinâmico.

Com esta base construiram-se dois "dashboards" semi-dinâmicos.

Num deles é possível que o utilizador escolha o distrito para analisar o número de dias de internamento por sexo (Figura 128).

Figura 128 – Dashboard que revela para o Distrito seleccionado o número de dias de internamento por sexo (para dados do primeiro semestre de 2008).

O outro dashboard representa as GCD por total de dias de internamento e pré-operatório em formato de tabela, onde o utilizador, ao escolher uma das GCD, filtra automaticamente os gráficos abaixo (Figura 129). O gráfico do lado esquerdo representa o total de dias de internamento ao longo do tempo para a GCD seleccionada. O gráfico do lado direito representa o total de dias de internamento por estação do ano e tempo para a GCD seleccionada. É de referir também que ao seleccionar-se na tabela a GCD para análise, os títulos dos gráficos também são alterados para a descrição da GCD seleccionada. Por último, encontra-se em baixo dos gráficos uma barra de deslocamento ("slider") que permite ao utilizador final alargar ou encurtar o horizonte temporal dos gráficos.

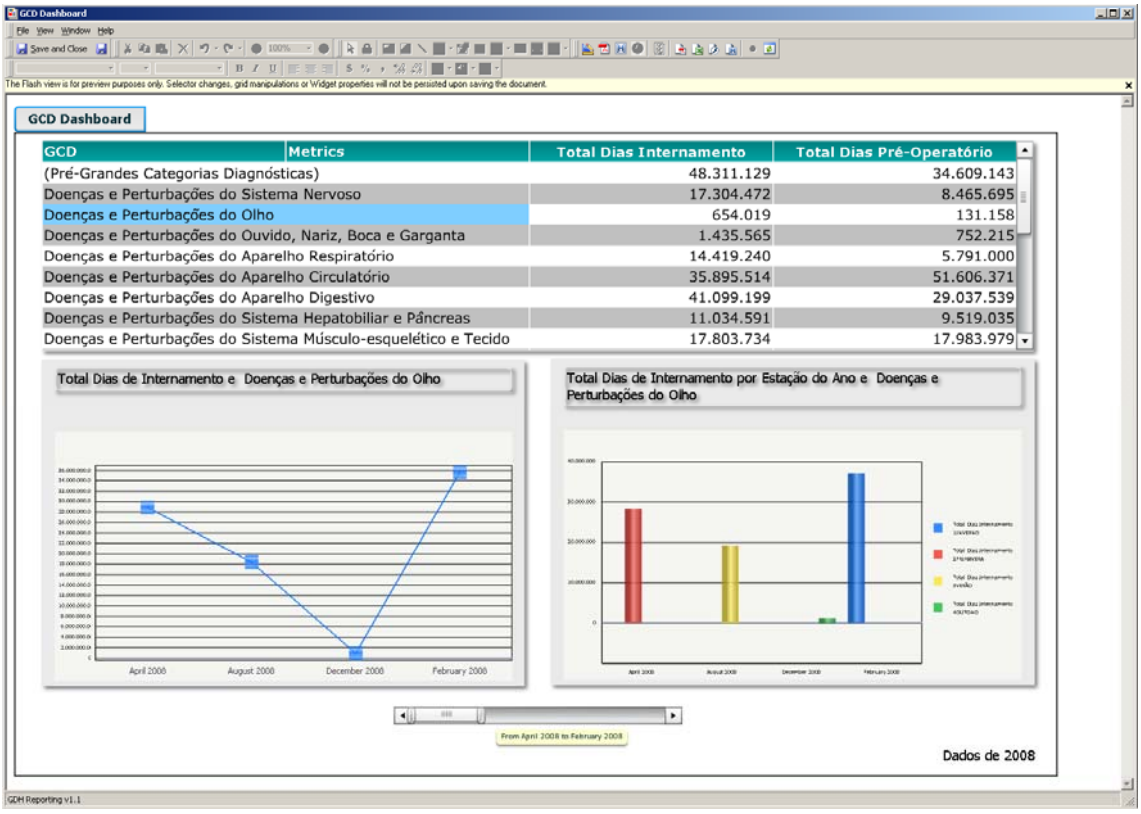


Figura 129 – Dashboard de análise das GCD sob diferentes perspectivas.

12 Conclusão

Não obstante as dificuldades que, invariavelmente, acompanham todos os processos de mudança, a implementação dos GDH's em Portugal, permitiu no universo do sistema de saúde, nomeadamente nas unidades de saúde hospitalares, dar um enorme passo na forma de encarar e fazer gestão, nomeadamente no apoio à tomada de decisão dos operacionais até á gestão de topo, visando melhorar o modo de funcionamento destas unidades, suportadas numa concepção inovadora do denominado produto hospitalar.

O modelo adoptado não constitui, com certeza, uma via definitiva, mas representa seguramente um caminho válido e coerente para se atingir um financiamento mais justo e racional para todos os hospitais englobados no Sistema do Serviço Nacional de Saúde (SNS).

Apesar de todos os condicionalismos impostos à partida, bem como aqueles que surgiram pelo caminho, contínuo seriamente convicto que é preferível ter um sistema não totalmente perfeito, mas dinâmico, do que esperar por um que esteja a um nível eventualmente inatingível de perfeição.

Em relação às diversas fases deste projecto de BI, análise, desenho e implementação, todas elas foram superadas com mais ou menos dificuldade. Foi um desafio constante, que começou nos dados fonte, que se encontravam num formato orientado para o sistema transaccional, ou seja, o modelo da base de dados que serviu como suporte na construção do sistema de BI foi desenvolvido segundo a denominada "orientação à coluna", condicionada por algumas limitações em termos do número de diagnósticos e noutras características dos pacientes. As limitações da base de dados fonte penalizaram em parte a componente analítica e multidimensional. No entanto, utilizando as melhores práticas, com base na literatura, de "data warehousing" foi possível superar todas as adversidades.

A fase de análise desta dissertação foi muito importante para o resultado final obtido, pois permitiu desenhar uma solução de como o sistema se deveria comportar e qual o aspecto que deveria ter após a implementação. Assim, foi possível construir um sistema que disponibiliza, de uma forma simples e rápida, dados através de uma ferramenta de "reporting" que permite aos profissionais de saúde construir as suas análises de forma autónoma, sem ser necessário a intervenção do departamento de informática.

Verifiquei que a fonte de dados não estava optimizada para análises multidimensionais, bem como para consultas mais complexas. Perante esta realidade, teve de ser empregue no

desenvolvimento do ETL as melhores técnicas de optimização, para atingir melhor a performance, devido ao volume final de dados (cerca de 30 milhões de registos). Neste sentido, foi dissecado um tema clássico para optimização de consultas a um DW que são os índices, estes são mecanismos de optimização, que assumem vários tipos, no entanto, a escolha de um tipo, a aplicar ao DW, não foi trivial porque não existe um tipo óptimo, mas sim diferentes tipos que consoante o padrão de consultas se adaptam melhor às necessidades. A escolha do tipo de índices a aplicado levou-me a uma análise cuidada do padrão de análises/consultas, ao qual o DW teve que responder. Após este estudo, optei por um tipo que reduziu o tempo de resposta do DW às consultas.

Por outro lado, a definição do tipo de dimensões ("slowly changing dimension") também me obrigou a um entendimento da evolução da codificação dos GDH's de forma a optar qual seria o melhor tipo que se adaptava ao cenário em questão, então, acabei por concluir que o tipo 1 (sobreposição do valor) seria o mais indicado visto que as alterações à codificação dos GDH's têm de ser reflectidas para todos os registos actuais e anteriores.

Ainda em relação às dimensões, foi criada a dimensão "Estação do Ano" que possibilita a quem analisa (utilizadores finais) mais uma fonte de informação para análise. Outra dimensão, deveras importante, que denominei como "Tempo" foi criada através dos Analysis Services (AS) que funcionou de forma inversa ao esperado, ou seja, geralmente os dados são transportados da camada relacional para a camada analítica (cubos), no entanto, para a criação desta dimensão e respectiva granularidade, utilizei os AS. Após a criação desta dimensão nos AS, a sua estrutura, incluindo os dados, foram exportados para a camada inferior denominada BD relacional. Desta forma, cheguei à conclusão que os AS são uma ferramenta que ao nível da dimensão"Tempo" podem enriquecer a componente relacional.

A ideia do envio de e-mail após a execução do processo de ETL, que carrega os factos no DW, com data de início e de fim do carregamento de dados, assim como o nome dos ficheiros carregados e volumetria de registos, obrigou-me a um desenvolvimento à medida muito específico. Esta funcionalidade permite facilitar a monitorização em ambiente remoto, assim como a manutenção do sistema de BI.

Quanto à componente de visualização de dados por parte dos utilizadores finais, utilizei a ferramenta da MicroStrategy para a implementação da camada de "reporting", esta ferramenta que desconhecia, teve de ser estudada, investigada de forma a poder utilizar as suas potencialidades para a construção de relatórios e "dashboards" com dinamismo e significado para quem os utiliza.

Ao nível das métricas, a questão do custo do número de dias de internamento por GDH's não foi considerado, visto que uma das tentativas de implementação através da fórmula, número de dias de internamento por GDH's a multiplicar pelo custo da diária, não estava correcta segundo a ACSS. Como não consegui obter uma fórmula exacta para o custo, optei por colocar as métricas provenientes do sistema fonte, sendo estas o total de dias de internamento, o número de dias de pré-operatório e o número de pacientes. Contudo, caso se venha a pretender construir fórmulas relativas a custos de internamentos ou outras, as métricas base encontram-se disponíveis no DW e na camada de "reporting", assim como as características dos GDH's (como por exemplo, a diária de internamento, o peso relativo, entre outras). Assim, o utilizador final apenas tem de seleccionar os atributos que pretende para construir a fórmula (nova métrica) que deseja implementar. Do exposto, pode-se concluir que o sistema de BI é versátil e dinâmico.

Cabe-me ainda destacar uma outra funcionalidade, que foi implementada na camada de "reporting". Esta consistiu na opção de "drill" que permite aos utilizadores navegar nas hierarquias das dimensões (do elemento mais elementar até ao mais agregador e vice-versa), como é o caso das dimensões tempo, GCD e GDH. Concluiu-se que esta funcionalidade irá facilitar a navegação pelos dados por parte dos utilizadores.

Realço a potencialidade da ferramenta de "reporting" da MicroStrategy que me ofereceu e poderá oferecer aos utilizadores finais um leque de opções de análise muito vasto, já que sob este ponto de vista é bastante "user-friendly".

Ao nível do "negócio" propriamente dito, a temática dos GDH's foi um tema novo para mim visto, que não tinha noção de que modo se processava o financiamento dos hospitais portugueses ao nível de internamentos. Este conceito de financiamento é susceptível de erros dos mais diversos, visto que está dependente do factor humano, desde o diagnóstico ao número de dias de internamento, no entanto, este sistema de BI irá permitir reforçar a auditoria aos dados dos GDH's de forma a facilitar a monitorização e identificação de erros. Para além disso, permite também efectuar análises comparativas entre hospitais no sentido de avaliar a coerência e evolução dos GDH's a nível nacional. Com este sistema de BI, a classificação dos pacientes em GDH's poderá ser analisada de múltiplas perspectivas com o intuito de melhorar e antecipar qualquer tipo de falha visto que os hospitais em 2010 foram financiados com os dados inseridos em sistema no ano de 2009.

Em suma, esta dissertação foi um fio condutor para uma aprendizagem constante à medida que a implementação ia avançando, o que me possibilitou um amadurecimento consistente

nas tecnologias usadas (Microsoft SQL Server e MicroStrategy), técnicas de "data warehousing" e do conceito de BI aplicado à área da Saúde, nomeadamente aos GDH's.

Não foi um desafio fácil, devido à sensibilidade do tema dos GDH's que envolve custos. Os hospitais Portugueses são financiados através deste sistema que funciona numa lógica de orçamento, ou seja, cada hospital efectua a sua classificação de internamentos em GDH's e com base na classificação, e respectivos custos de internamentos do ano anterior ao presente, é atribuído o valor do financiamento para o presente ano.

Por outro lado, existiram algumas limitações de software (todo ele gratuito) sobretudo na componente de "reporting" que apenas permite usar um CPU, do sistema operativo "Windows Vista Home Edition" que limitou bastante a performance no acesso aos dados, e do hardware (visto que seriam necessários 8 Gb de memória RAM), no entanto, com esforço e dedicação foi possível realizar este sonho.

Hoje, resultado da presente dissertação, temos ao dispor um sistema de BI aplicado aos GDH's que permite a qualquer utilizador realizar as análises que pretende, consultar "dashboards" e relatórios sem limites.

Finalmente, de referir que a partir desta dissertação de mestrado foi elaborado um artigo científico que se apresentou na CISTI'2010 (5ª Conferência Ibérica de Sistemas e Tecnologias de Informação).

13 Perspectivas Futuras

Os sistemas de BI estão em constante evolução, de forma a acompanhar as necessidades do negócio que vão sofrendo alterações ao longo do tempo, necessidades essas que cada vez mais se apresentam mais exigentes.

O sistema de classificação de pacientes com base em GDH's é utilizado em diversos países, e uma possível evolução e enriquecimento desta dissertação seria adicionar dados de outros países de forma a que fosse possível efectuar análises comparativas com outras realidades, como por exemplo, comparar por cada país quais os GDH's com mais dias de internamento e com que diagnóstico, entre outras análises.

Uma outra componente evolutiva deste trabalho seria a geo-referenciação, que iria permitir de forma gráfica compreender, por exemplo, o foco a nível regional de determinados diagnósticos, em que zonas do País se encontram os hospitais com mais dias de internamento, entre outros estudos.

Uma outra vertente poderá ser desenvolvida a partir do presente trabalho, que será disponibilizar este sistema de BI aos profissionais de saúde que lidam com os GDH's, sobretudo aqueles que estão directamente ligados à auditoria. Saliento, todos os relatórios que foram criados, com base em análises multidimensionais, tiveram como objectivo fazer transparecer o potencial da ferramenta de "reporting" para as áreas de saúde carentes de informação sobre os GDH's. Assim, num futuro próximo, há todo o interesse em envolver a maioria dos profissionais de saúde neste sistema de BI.

Por outro lado, a forma como o sistema de BI foi construído permite adaptar-se de forma rápida às evoluções dos GDH's, como por exemplo:

- Em caso de necessidade de introduzir uma nova métrica, esta será fácil de implementar, bastando para tal criar uma fórmula na componente de "reporting" (como por exemplo, a fórmula do custo de internamento por GDH's que foi referida anteriormente).

- Se surgirem mais atributos para caracterizar as dimensões, a sua implementação tem pouco impacto, visto que se trata apenas de metadados que não afectam as ligações entre dimensões e factos. Um exemplo deste tipo de questão, consiste em acrescentar o nome e o número do SNS do paciente na dimensão paciente, visto que para o desenvolvimento deste sistema de BI apenas nos foi fornecido dados mascarados com

um identificador sem significado em relação aos pacientes (o que foi o suficiente para esta tese visto que não havia qualquer interesse em saber mais informações sobre os pacientes) que permite identificar de forma unívoca cada paciente registado na base de dados fonte.

Por último, este sistema de BI permite apenas analisar dados históricos, ou seja, factos ocorridos e registados no passado que podem ser levados em conta pelos profissionais de saúde como suporte às suas decisões. No entanto, este sistema poderia ser complementado no futuro se fossem construídos modelos preditivos (utilizando técnicas de "data mining") com base nos dados do DW. Desta forma seria possível efectuar previsões com base nos dados dos GDH's, como por exemplo a realização de orçamentos, a previsão média do número de dias de internamento por diagnóstico, entre outras.

As dificuldades são como as montanhas. Elas só se aplainam quando avançamos sobre elas."
- Provérbio japonês

14 Bibliografia

ACSS - Administração Central do Sistema de Saúde. (2006). *SISTEMA DE CLASSIFICAÇÃO DE DOENTES EM GRUPOS DE DIAGNÓSTICO HOMOGÉNEOS (GDH).* Lisboa: ACSS.

Araújo, N. (2008). *Acções Autónomas de Enfermagem - Ganhos em Saúde.* Porto: Universidade Fernando Pessoa.

Averill, R. (1998). The evolution of case mix measurement : using diagnosis related groups. *3M HIS Research Report* , pp. 5-98.

Bentes, M. (1996). A utilização dos GDHs como instrumento de financiamento hospitalar. *Gestão Hospitalar* , 33-40.

Bentes, M. (1997). Formas de pagamento de serviços hospitalares: resumo da comunicação. *Jornadas Ibéricas de Gestão Hospitalar* .

Bentes, M. (1998). O financiamento dos hospitais. *IGIF* .

Bentes, M. (1991). Using DRGs to fund hospitals in Portugal: an evaluation of the experience. *Ministério da Saúde* .

Borysowich, C. (27 de Novembro de 2007). *Snowflake Schema Modelling (Data Warehouse)*. Obtido em 4 de Novembro de 2010, de http://it.toolbox.com/blogs/enterprise-solutions/snowflake-schema-modelling-data-warehouse-20809

Casas, M. (1991). Issues for comparability of DRG statistics in Europe : results from EURODRG. *Health Policy* , 121-132.

Charbonneau, C. (1988). Validity and reliability issues in alternative patient classification systems. *Medical Care* , 800-813.

Cincinnati Children's Hospital Medical Center. (03 de Janeiro de 2009). *Research Data Warehouse*. Obtido em 24 de Maio de 2009, de http://i2b2.cchmc.org/faq

Costa, C. (1990). Financiamento de serviços de saúde : a definição de preços. *Revista Portuguesa de Saúde Pública* , 65-72.

Costa, C. (1994). Os DRGs e a gestão do hospital. *Revista Portuguesa de Gestão* , 47-65.

Cunha, J. R. (10 de Maio de 2008). *Solução de Business Intelligence no Hospital Militar Principal.* Obtido em 8 de Abril de 2009, de http://www.sisaude.oninet.pt/apresenta08/14-BusinessIntelligenceHospitalMilitar-Entrega.pdf

Daniel, P. (10 de Março de 2007). *A Brief History of Decision Support Systems*. Obtido em 22 de Março de 2009, de DSSResources.COM: http://dssresources.com/history/dsshistory.html

Davenport, T., & Harris, J. (2007). The Architecture of Business Intelligence. Boston: Harvard Business School Press.

Dismuke, C. (1996). *A preliminary analysis of the DRG system in Portugal : hospital response as measured by length of stay.* Lisboa: APES.

Escoval, A. (1997). Sistemas de financiamento de saúde : análise e tendência. *ISCTE* .

Feiman, J., & MacDonald, N. (2010). Magic Quadrant for Business Intelligence Platforms. *Gartner RAS Core Research Note G00173700* , 2.

Fleming, M. (2010). *How b-tree database indexes work and how to tell if they are efficient (100' level)*. Obtido em 12 de Junho de 2010, de http://mattfleming.com/node/192

Hans, L. (1958). A Business Intelligence System. *IBM Journal* , 314-319.

Himmelsbach, V. (10 de Maio de 2005). *How business intelligence is making healthcare smarter*. Obtido em 23 de Março de 2009, de Signs of Intelligent Life: http://www.connectingforhealth.nhs.uk/newsroom/worldview/protti10/

Hughes, R. (1 de Novembro de 2004). *Optimal Data Architecture for Clinical Data Warehouses.* Obtido em 18 de Março de 2009, de Information Management: http://www.information-management.com/issues/20041101/1012400-1.html

Iezzoni, L. (1989). A description and clinical assessment of the computerized severity of illness. *Boston University Medical Center* .

Instituto de Gestão Informática e Financeira da Saúde. (2005). *Auditoria às Bases de Dados dos GDHs.* Lisboa: Departamento de Desenvolvimento de Sistemas de Financiamento e Gestão.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit* (2nd ed.). New York: John Wiley & Sons.

Lima, E. (2000). *The financing systems and the performance of Portuguese hospitals.* Lisboa: APES.

Maurer, J. (1 de Outubro de 2000). Obtido em 4 de Agosto de 2010, de http://www.auditmypc.com/acronym/dbf.asp

Meta Análise. (20 de Junho de 2008). *BI ajuda Hospital Samaritano a melhorar remuneração* . Obtido em 23 de Março de 2009, de Meta Análise - Inteligência de Mercado: http://www.metaanalise.com.br/inteligenciademercado/palavra-aberta/melhores-praticas/bi-ajuda-hospital-samaritano-a-melhora-remunerac-o.html

Microsoft. (2010). *Clustered Index Structures*. Obtido em 12 de Junho de 2010, de http://msdn.microsoft.com/en-us/library/ms177443.aspx

Microsoft. (2010). *How to: Connect to a dBASE or Other DBF File*. Obtido em 5 de Agosto de 2010, de http://technet.microsoft.com/en-us/library/aa337084.aspx

MicroStrategy. (2010). *Installation and Configuration Guide.* Virginia: MicroStrategy Incorporated.

MicroStrategy. (2010). *MicroStrategy*. Obtido em 1 de Junho de 2010, de http://www.microstrategy.com.br/Company/index.asp

MicroStrategy. (2009). *Project Design Guide.* Virginia: MicroStrategy Incorporated.

Owen, H. (6 de Abril de 2006). *Using Dashboard-Based Business Intelligence Systems*. Obtido em 22 de Março de 2009, de Graziadio Business Report: http://gbr.pepperdine.edu/034/bis.html

Pinto, J. (2002). *Performance organizacional : variabilidade no desempenho de dois serviços hospitalares com perfil de produção similar.* Lisboa: ENSP.

Primak, F. V. (8 de Novembro de 2009). Obtido em 5 de Abril de 2010, de Uma introdução simplista aos conceitos de Business Intelligence: http://www.oficinadanet.com.br/artigo/2131/uma_introducao_simplista_aos_conceitos_de_business_intelligence_-_parte_1

Spiner. (15 de Junho de 2006). *CRM: Customer Relationship Management*. Obtido em 2010 de Agosto de 31, de http://www.spiner.com.br/modules.php?name=Forums&file=viewtopic&t=409

Thorpe, K. (1987). The distributional implications of using relative prices in DRG payment systems. *Inquiry* , 85-95.

Vertrees, J. (1998). Incentivos globais e competição nos serviços. *Encontro sobre Financiamento dos Sistemas de Saúde.* Lisboa: Comunicação apresentada no Encontro.

Vertrees, J. (1998b). Using DRGs for contracting in Romania. *3M Health Information Systems* , 3-25.

Victor, J. (3 de Abril de 2009). *Tesco abre banco dentro de lojas*. Obtido em 3 de Junho de 2009, de http://www.hipersuper.pt/2009/04/03/tesco-abre-banco-dentro-de-lojas/

Wang, J. (15 de Setembro de 2006). *Develop SOA solutions for healthcare organizations using business-driven development*. Obtido em 23 de Maio de 2009, de IBM: http://www.ibm.com/developerworks/webservices/library/ws-soa-bddhealth/

Wikipedia. (Setembro de 2009). *Esquema em estrela*. Obtido em 2 de Junho de 2010, de http://pt.wikipedia.org/wiki/Esquema_estrela

Wikipédia. (21 de Julho de 2010). *Estação do ano*. Obtido em 2010 de Julho de 22, de http://pt.wikipedia.org/wiki/Esta%C3%A7%C3%A3o_do_ano

Wikipedia. (12 de Junho de 2010). *Index (database)*. Obtido em 2010 de Junho de 12, de http://en.wikipedia.org/wiki/Index_(database)

Willems, J. (1989). Use of diagnosis related groups for hospital management. *Health Policy* , 121-133.

15 Anexos

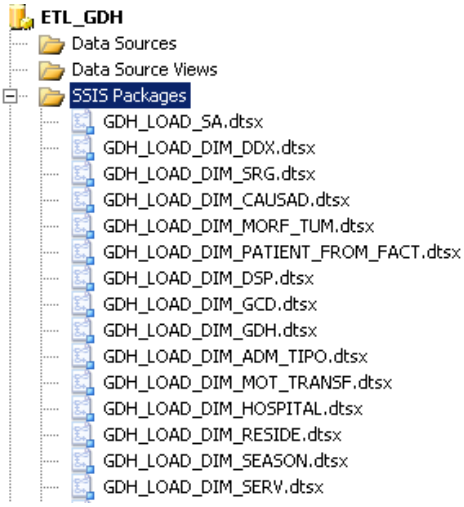
15.1 Mapeamento entre tabelas fonte e dimensões do DW

Tabela/Ficheiro Fonte	Campo/ Coluna Fonte	Data Warehouse	Tabela de Dimensão	Campo/ Coluna Destino	Tipo de dados
-	-	DW_GDH	Dim_Causa_Externa	SK_CAUSA_EXTERNA	int
SRC_CAUSAD_CAUSA_EXTERNA.xlsx	COD_CAUSA_EXT	DW_GDH	Dim_Causa_Externa	ID_CAUSA_EXTERNA	varchar
SRC_CAUSAD_CAUSA_EXTERNA.xlsx	DES_CAUSA_EXT	DW_GDH	Dim_Causa_Externa	DESC_CAUSA_EXTERNA	varchar
-	-	DW_GDH	Dim_Concelho	SK_CONCELHO	int
SRC_FREGS_CONCELHO_DISTR.xlsx	DTCC	DW_GDH	Dim_Concelho	ID_CONCELHO	nvarchar
SRC_FREGS_CONCELHO_DISTR.xlsx	CONCELHO	DW_GDH	Dim_Concelho	CONCELHO	nvarchar
Dim_Distrito	-	DW_GDH	Dim_Concelho	SK_DISTRITO	int
SRC_DSP_DESTINO_APOS_ALTA.xlsx	COD_DEST_ALTA	DW_GDH	Dim_Destino_Apos_Alta	SK_DESTINO_APOS_ALTA	int
SRC_DSP_DESTINO_APOS_ALTA.xlsx	DES_DEST_ALTA	DW_GDH	Dim_Destino_Apos_Alta	DESC_DESTINO_APOS_ALTA	varchar
-	-	DW_GDH	Dim_Diagnostico	SK_DIAGNOSTICO	int
SRC_DDX_DIAGNOSTICO.xlsx	COD_DIAGNOSTICO	DW_GDH	Dim_Diagnostico	ID_DIAGNOSTICO	varchar
SRC_DDX_DIAGNOSTICO.xlsx	DES_DIAGNOSTICO	DW_GDH	Dim_Diagnostico	DESC_DIAGNOSTICO	varchar
-	-	DW_GDH	Dim_Distrito	SK_DISTRITO	int
SRC_FREGS_CONCELHO_DISTR.xlsx	DT	DW_GDH	Dim_Distrito	ID_DISTRITO	nvarchar
SRC_FREGS_CONCELHO_DISTR.xlsx	DISTRITO	DW_GDH	Dim_Distrito	DISTRITO	nvarchar
-	-	DW_GDH	Dim_Freguesia	SK_FREGUESIA	int
SRC_FREGS_CONCELHO_DISTR.xlsx	DTCCFR	DW_GDH	Dim_Freguesia	ID_FREGUESIA	nvarchar
SRC_FREGS_CONCELHO_DISTR.xlsx		DW_GDH	Dim_Freguesia	FREGUESIA	nvarchar
Dim_Concelho	-	DW_GDH	Dim_Freguesia	SK_CONCELHO	int
SRC_GCD_GRANDE_CATEGORIA_DIAGNOSTICO.xlsx	GCD_COD	DW_GDH	Dim_GCD	SK_GCD	int
SRC_GCD_GRANDE_CATEGORIA_DIAGNOSTICO.xlsx	GCD_DESC	DW_GDH	Dim_GCD	DESC_GCD	varchar
-	-	DW_GDH	Dim_GDH	SK_GDH	int
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	COD_GDH	DW_GDH	Dim_GDH	ID_GDH	int
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	GCD_COD	DW_GDH	Dim_GDH	SK_GCD	int
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	GDH_DESC	DW_GDH	Dim_GDH	DESC_GDH	varchar
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Tipo GDH	DW_GDH	Dim_GDH	TIPO_GDH	varchar
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Peso Relativo	DW_GDH	Dim_GDH	PESO_RELATIVO	float
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Preço	DW_GDH	Dim_GDH	PRECO	money
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Peso Relativo em Ambulatório	DW_GDH	Dim_GDH	PESO_REL_AMB	float
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Preço em Ambulatório	DW_GDH	Dim_GDH	PRECO_AMB	money
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Diária de Internamento	DW_GDH	Dim_GDH	DIARIA_INTER	money
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	GDH Cirurgicos - Preço 1º dia de internamento	DW_GDH	Dim_GDH	GDH_CIRURGICOS_1D_INTER	money
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Limiar Inferior	DW_GDH	Dim_GDH	LIM_INFERIOR	float
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Limiar Superior	DW_GDH	Dim_GDH	LIM_SUPERIOR	float
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Limiar Máximo	DW_GDH	Dim_GDH	LIM_MAXIMO	float
SRC_GDH_GRUPO_DIAGNOSTICO_HOMOGENEO.xlsx	Demora Média Corrigida	DW_GDH	Dim_GDH	DEM_MEDIA_CORRIGIDA	float
-	-	DW_GDH	Dim_GDH	PRECO_CONV	float
-	-	DW_GDH	Dim_GDH	DIARIA_INTER_CONV	float
-	-	DW_GDH	Dim_Hospital	SK_HOSPITAL	int
SRC_HOSPITAL_ID.xlsx	Hosp_id	DW_GDH	Dim_Hospital	ID_HOSPITAL	varchar
SRC_HOSPITAL_ID.xlsx	Nome	DW_GDH	Dim_Hospital	DESC_HOSPITAL	varchar
-	-	DW_GDH	Dim_Morfologia_Tumoral	SK_MORF_TUM	int
SRC_MORF_TUM_MORFOLOGIA_TUMORAL.xlsx	COD_MORF_TUM	DW_GDH	Dim_Morfologia_Tumoral	ID_MORF_TUM	int
SRC_MORF_TUM_MORFOLOGIA_TUMORAL.xlsx	DES_MORF_TUM	DW_GDH	Dim_Morfologia_Tumoral	DESC_MORF_TUM	varchar
SRC_MOT_TRANS_MOTIVO_TRANSFERENCIA.xlsx	COD_MOTIVO_TRANSF	DW_GDH	Dim_Motivo_Transferencia	SK_MOT_TRANSF	int
SRC_MOT_TRANS_MOTIVO_TRANSFERENCIA.xlsx	DES_MOTIVO_TRANSF	DW_GDH	Dim_Motivo_Transferencia	DESC_MOT_TRANSF	varchar
-	-	DW_GDH	Dim_Paciente	SK_PACIENTE	int
Tabela de Factos	NUMERO	DW_GDH	Dim_Paciente	ID_PACIENTE	int
Tabela de Factos	SEXO	DW_GDH	Dim_Paciente	SEXO	varchar
Tabela de Factos	B_DATE	DW_GDH	Dim_Paciente	DATA_DE_NASCIMENTO	int
Tabela de Factos	BIRTH_WGT	DW_GDH	Dim_Paciente	PESO_NASCENCA	int
-	-	DW_GDH	Dim_Procedimento	SK_PROCEDIMENTO	int
SRC_SRG_PROCEDIMENTO.xlsx	COD_PROCEDIMENTO	DW_GDH	Dim_Procedimento	COD_PROCEDIMENTO	varchar
SRC_SRG_PROCEDIMENTO.xlsx	DES_PROCEDIMENTO	DW_GDH	Dim_Procedimento	DES_PROCEDIMENTO	varchar
TBL_SEASON	ID_SEASON	DW_GDH	Dim_Sazonalidade	SK_SEASON	int
TBL_SEASON	SEASON_DESC	DW_GDH	Dim_Sazonalidade	SEASON_DESC	varchar
TBL_SEASON	BEGIN_MONTH	DW_GDH	Dim_Sazonalidade	BEGIN_MONTH	int
TBL_SEASON	BEGIN_DAY	DW_GDH	Dim_Sazonalidade	BEGIN_DAY	int
TBL_SEASON	SEASON_DATE	DW_GDH	Dim_Sazonalidade	SEASON_DATE	date
SRC_SERV_SERVIÇOS.xlsx	SK_SERV	DW_GDH	Dim_Servico	SK_SERVICO	int
SRC_SERV_SERVIÇOS.xlsx	HOSP_ID	DW_GDH	Dim_Servico	HOSP_ID	nvarchar
SRC_SERV_SERVIÇOS.xlsx	COD_SERV	DW_GDH	Dim_Servico	ID_SERVICO	nvarchar
SRC_SERV_SERVIÇOS.xlsx	DES_ESPECIALIDADE	DW_GDH	Dim_Servico	DESC_SERVICO	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	PK_Date	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Date_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Year	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Half_Year	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Half_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Trimester	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Trimester_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Week	datetime
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Week_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Half_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Half_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Trimester	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Trimester_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Quarter	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Quarter_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Month	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Month_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Week	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Day_Of_Week_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Week_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Week_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Half_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Half_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Trimester	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Trimester_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Quarter	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Month_Of_Quarter_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter_Of_Half_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Quarter_Of_Half_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Trimester_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Trimester_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Half_Year_Of_Year	int
Microsoft Analysis Services	-	DW_GDH	Dim_Time	Half_Year_Of_Year_Name	nvarchar
Microsoft Analysis Services	-	DW_GDH	Dim_Time	SK_DATE	int
SRC_ADM_TIPO_TIPO_DE_ADMISSAO.xlsx	ADM_TIPO	DW_GDH	Dim_Tipo_Admissao	SK_TIPO_ADMISSAO	int
SRC_ADM_TIPO_TIPO_DE_ADMISSAO.xlsx	DESCRIÇÃO	DW_GDH	Dim_Tipo_Admissao	DESC_TIPO_ADMISSAO	varchar

15.2 Tabela de volumetria de registos

Nome da Tabela	Número de Registos
Dim_Causa_Externa	1265
Dim_Concelho	309
Dim_Destino_Apos_Alta	7
Dim_Diagnostico	15208
Dim_Distrito	30
Dim_Freguesia	4256
Dim_GCD	29
Dim_GDH	670
Dim_Hospital	108
Dim_Morfologia_Tumoral	1078
Dim_Motivo_Transferencia	6
Dim_Paciente	973954
Dim_Procedimento	4548
Dim_Residencia	4620
Dim_Sazonalidade	4
Dim_Servico	10266
Dim_Time	18262
Dim_Tipo_Admissao	7
Fact_GDH	31200387

15.3 Lista de packages de ETL desenvolvidos



15.4 Criação de uma base de dados de teste e de um processo de importação directa de um ficheiro

Para criar efectivamente uma base de dados de teste, é necessário no MS carregar com o botão esquerdo do rato na pasta "Databases" e escolher a opção "New Database…" (Figura 130).

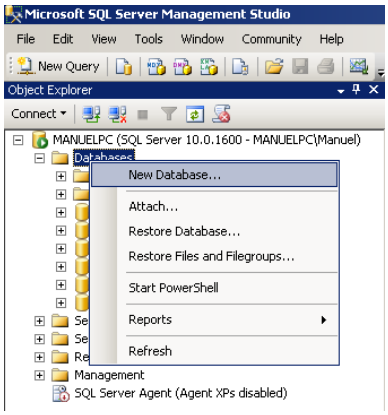


Figura 130 – Opção de criação de uma nova BD.

Seguidamente, surge o seguinte menu para dar um nome à nova BD a criar (Figura 131).

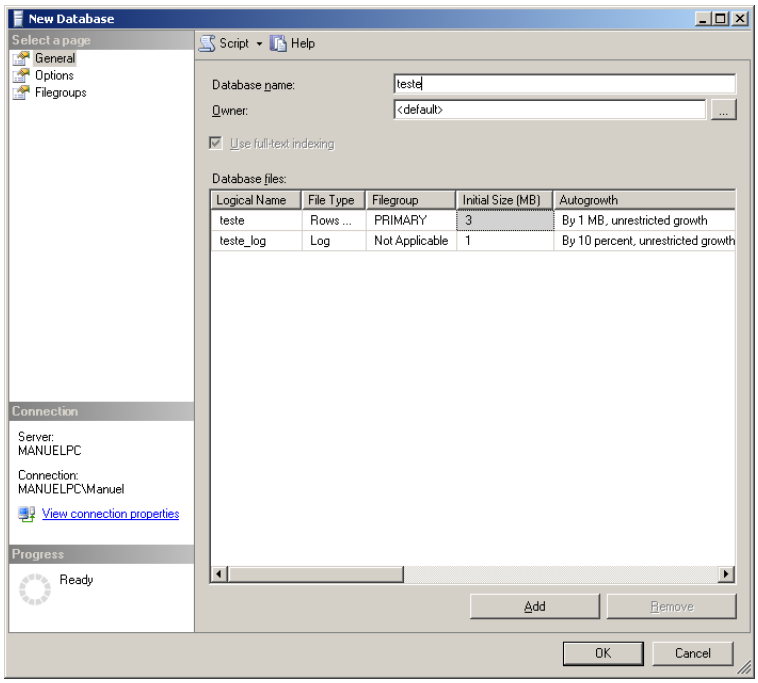


Figura 131 – Criação da BD de "teste" no SQL Server.

Após a criação da BD de "teste" experimentou-se utilizar a funcionalidade "Import Data…" (Figura 132) do Management Studio (MS) para importar directamente os dados do ficheiro fonte (*.dbf) para a BD de "teste".

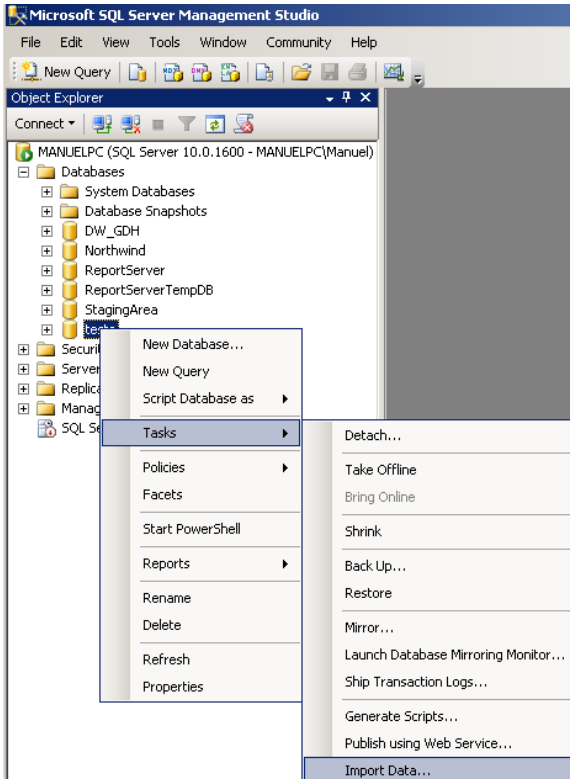


Figura 132 – Opção de importação directa de dados para BD's do SQL Server.

No entanto, não foi possível realizar esta operação visto que o wizard de importação/exportação do MS não permite esta importação directa (Microsoft, 2010), o que se revelou uma limitação da ferramenta que foi contornada. A sugestão da Microsoft consiste em abrir o ficheiro dbf no Excel ou no Access e depois importá-lo de forma directa para o SQL Server. Abriu-se o ficheiro no Excel e gravou-se como ficheiro do tipo *.xlsx e após esse passo regressou-se ao SQL Server para efectuar a importação dos dados. Ao seleccionar novamente a opção "Import Data…" como ilustra a figura acima e que consiste num wizard, efectuaram-se os seguintes passos:

1. Definição da fonte de dados que consiste no Excel 2007 e a primeira linha do ficheiro contem o nome das colunas (Figura 133).

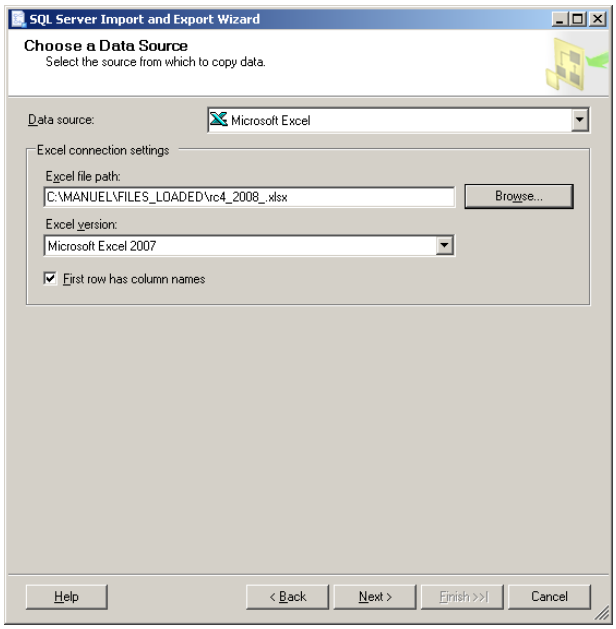


Figura 133 – Definição da fonte de dados.

2. O próximo passo consistiu em definir o destino dos dados, ou seja, a BD de "teste" que foi criada no SQL Server (Figura 134).

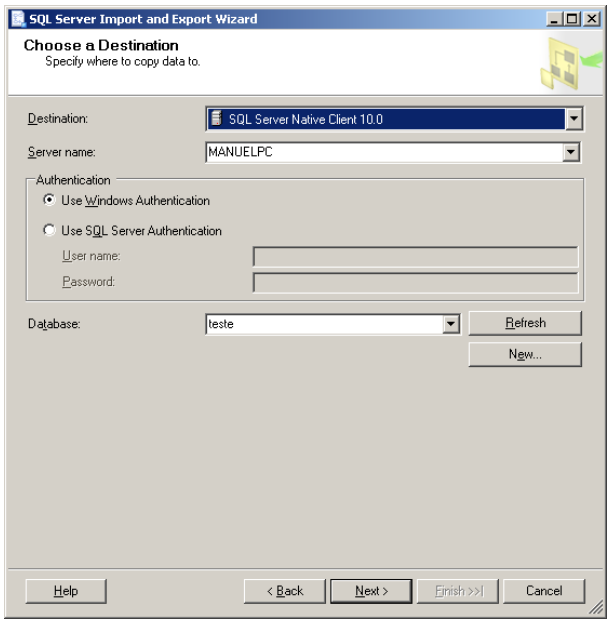


Figura 134 – Definição do destino dos dados.

3. Após o passo anterior, especificou-se que se pretendia copiar dados de uma tabela que neste caso corresponderá a uma sheet do Excel (Figura 135).

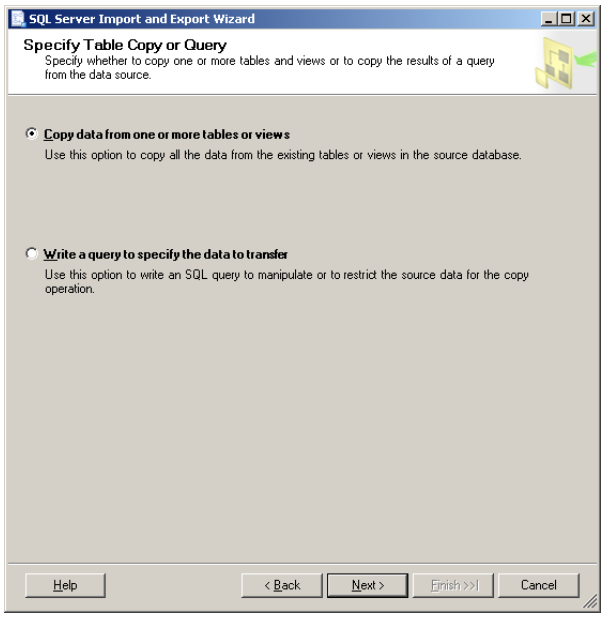


Figura 135 – Especificação de dados a tranferir.

4. Seguidamente, especificou-se que os fonte da sheet de Excel iriam ser copiados para uma tabela designada “tbl_teste” (Figura 136).

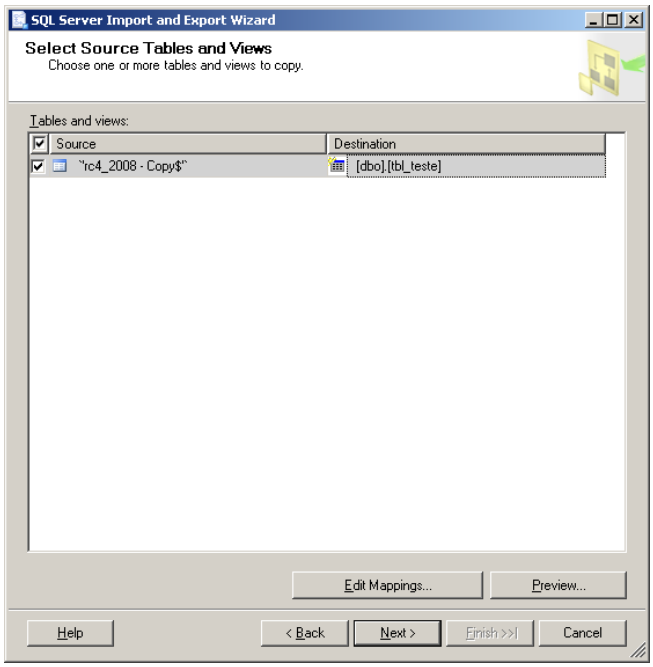


Figura 136 – Especificação da fonte de dados (sheet de Excel) e o destino que consiste numa tabela designada "tbl_teste".

5. No passo seguinte, executou-se o processo (Figura 137).

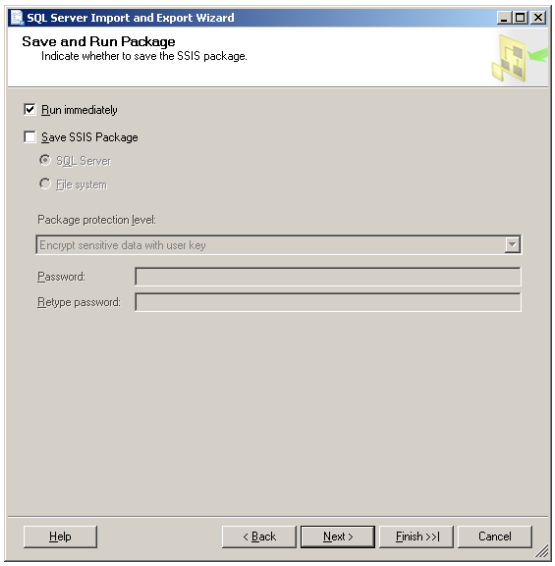


Figura 137 – Execução do processo.

6. Por último, verificou-se que o conteúdo do ficheiro Excel (68073 linhas) foi copiado com sucesso para a tabela "tbl_teste" da BD de "teste" do SQL Server (Figura 138).

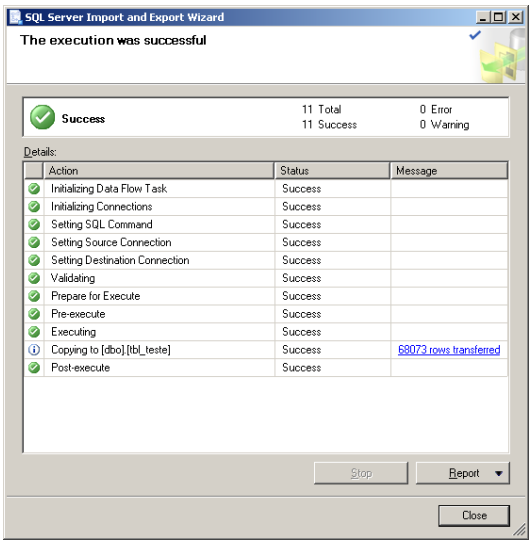


Figura 138 – Status da importação de dados.

15.5 Criação de um projecto directo no MicroStrategy

Para a criar um projecto directo de "reporting" no MicroStrategy iniciou-se o programa MicroStrategy Desktop (Figura 139).

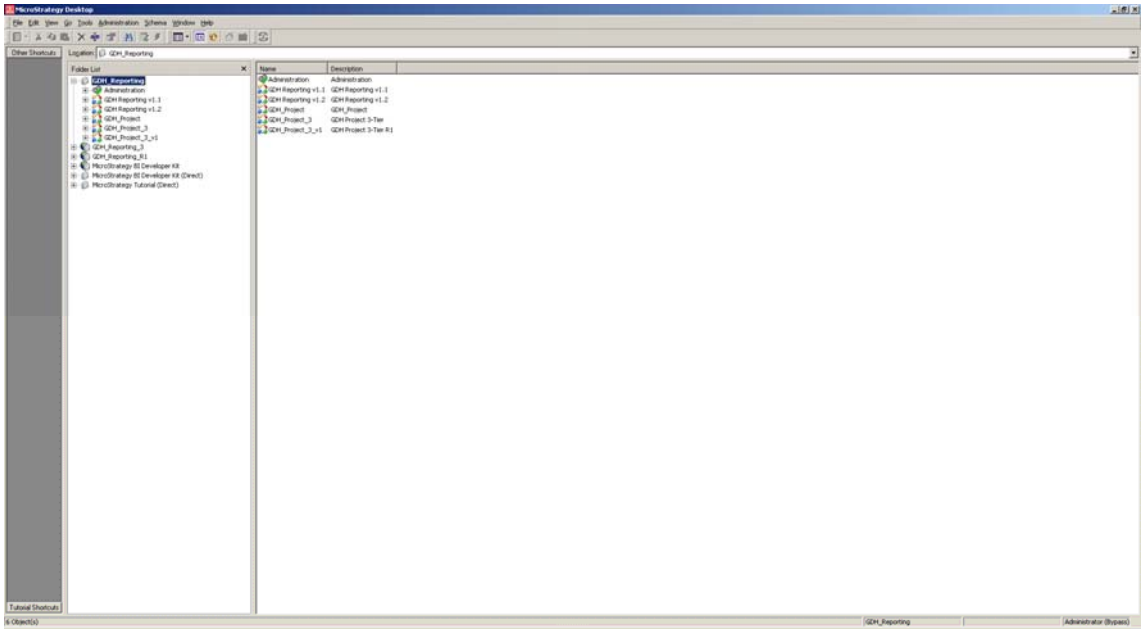


Figura 139 – Ecrã principal do MicroStrategy Destopk onde se criam os projectos de "reporting".

Seguidamente, no menu "Schema", escolhe-se a opção "Create New Project…" (Figura 140).

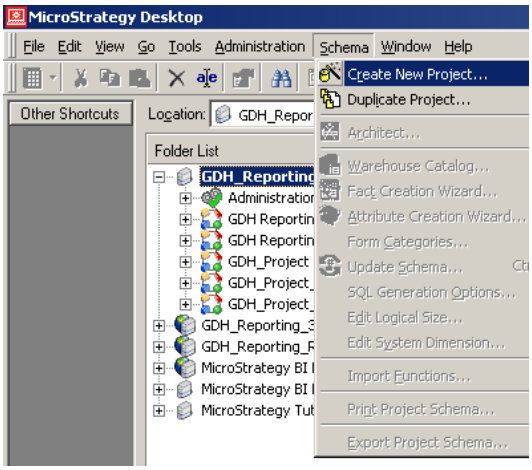


Figura 140 – Criação de um projecto directo no MicroStrategy.

Surge o assistente de criação de projectos (Figura 141) e escolhe-se a opção “Create Project”.

Figura 141 – Assistente de criação do projecto.

O passo seguinte consiste em definir o nome (GDH_Reporting), a descrição (GDH_Project) e a directoria do projecto (Figura 142).

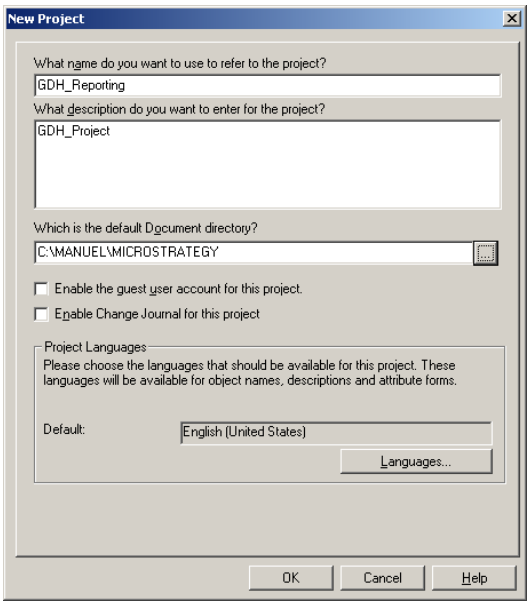


Figura 142 – Criação de um novo projecto.

Após o projecto criado com sucesso, o passo seguinte consiste em construir o catálogo de tabelas do DW (Figura 143).

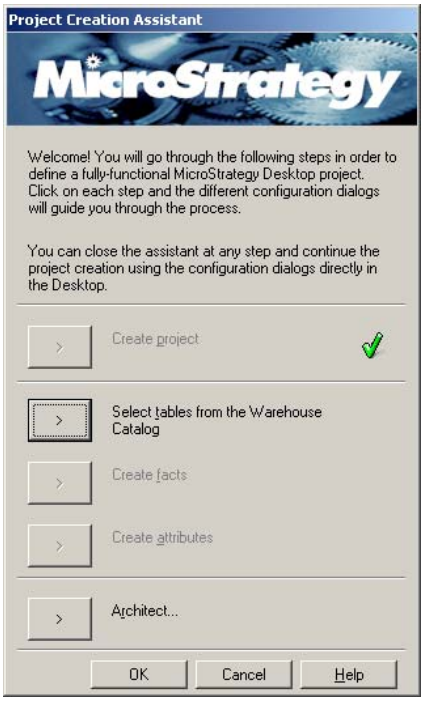


Figura 143 – Criação do catálogo de tabelas do DW.

Para aceder ao catálogo de tabelas do DW é necessário criar uma instância de forma a que o MicroStrategy se consiga ligar ao SQL Server (Figura 144).

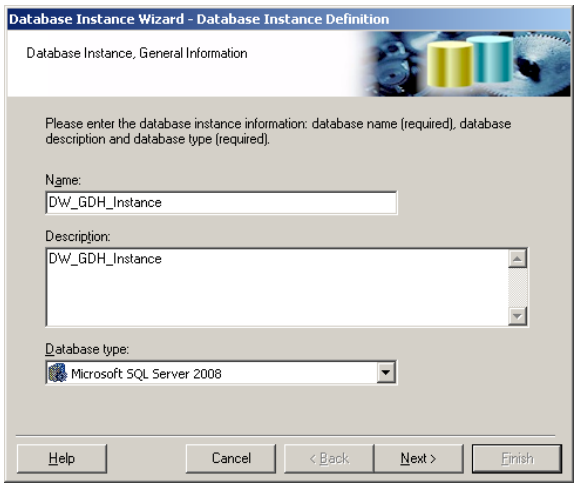


Figura 144 – Criação da instância para ligação ao SQL Server.

Após definir-se a instância da BD e o seu tipo (SQL Server) é necessário escolher-se a opção de "Configure ODBC" de forma a configurar o ODBC para efectuar à BD (Figura 145).

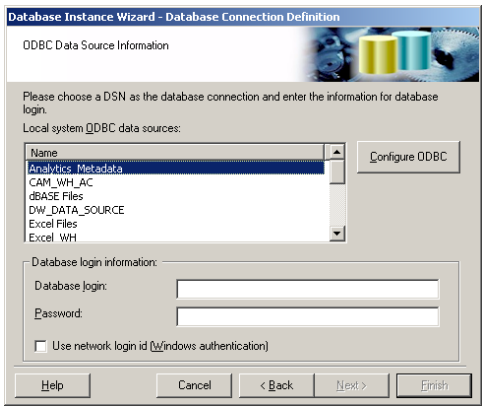


Figura 145 – Configuração da ligação ao DW.

A configuração do ODBC baseou-se nas credenciais do SQL Server onde reside o DW (Figura 146).

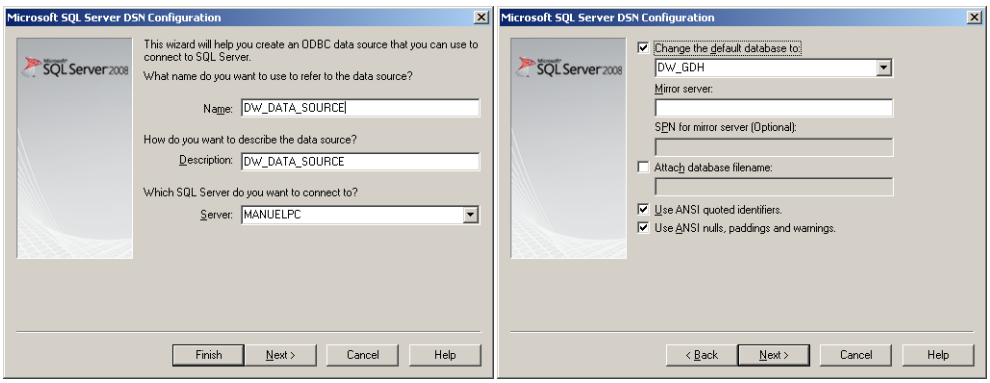


Figura 146 – Configuração do ODBC (ligação do DW).

Em seguida, regressa-se ao assistente de criação do projecto de forma a escolher a opção "Architect…" para desenhar a arquitectura de "reporting". Assim sendo, escolhe-se a instância anteriormente criada (DW_GDH_Instance) para ligação ao DW (Figura 147).

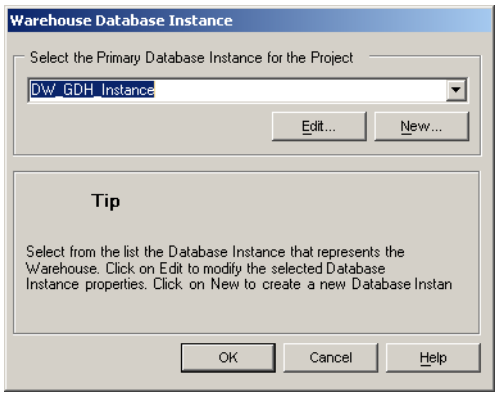


Figura 147 – Escolha da instância para ligação ao DW.

Após definida a instância de SQL Server para a ligação ao DW surge o passo seguinte que consiste em definir o modelo lógico de "reporting" do MicroStrategy. Este é o último passo do assistente de criação de projecto que direcciona para o "Architect…" onde será realizada essa construção (Figura 148).

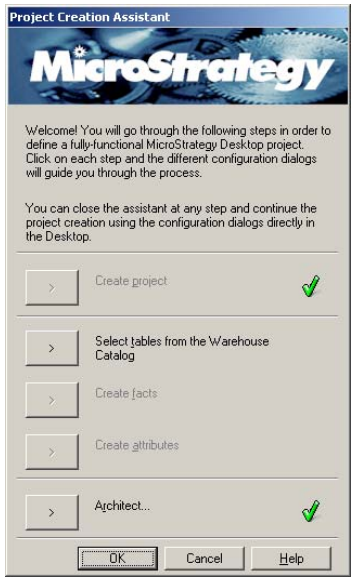


Figura 148 – Criação da arquitectura de "reporting".