UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

ESTUDO DE IMPLEMENTAÇÃO DE UM DATA WAREHOUSE PARA UMA INSTITUIÇÃO DO ENSINO SUPERIOR

NORBERTO LUÍS DE JESUS ALBINO

(LICENCIADO)

DISSERTAÇÃO PARA A OBTENÇÃO DO GRAU DE MESTRE EM

ENGENHARIA INFORMÁTICA

ORIENTADOR: PROFESSOR DOUTOR CARLOS PAMPULIM CALDEIRA

ÉVORA, DEZEMBRO DE 2009

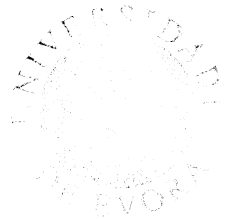
UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

ESTUDO DE IMPLEMENTAÇÃO DE UM DATA WAREHOUSE PARA UMA INSTITUIÇÃO DO ENSINO SUPERIOR

NORBERTO LUÍS DE JESUS ALBINO

(LICENCIADO)



DISSERTAÇÃO PARA A OBTENÇÃO DO GRAU DE MESTRE EM

ENGENHARIA INFORMÁTICA

185 664

ORIENTADOR: PROFESSOR DOUTOR CARLOS PAMPULIM CALDEIRA

ÉVORA, DEZEMBRO DE 2009

ESTUDO DE IMPLEMENTAÇÃO DE UM DATA WAREHOUSE PARA UMA INSTITUIÇÃO DO ENSINO SUPERIOR

RESUMO

As organizações actuais, em geral, e as Instituições de Ensino Superior, em particular, produzem e manipulam grandes volumes de informação.

Neste trabalho é estudada a implementação de um Data Warehouse para uma Instituição de Ensino Superior, com vista a permitir dotá-la de uma ferramenta analítica para auxiliar os seus decisores no processo de tomada de decisão.

Para demonstrar que esta abordagem tem vantagens para os elementos da organização que necessitam de tratar a informação, aumentando a qualidade e facilidade de acesso à mesma e maximizando o tipo e a quantidade de análises que podem efectuar aos dados, foi desenvolvido um protótipo com base na arquitectura BUS e na modelação dimensional.

A solução desenvolvida consiste numa ferramenta que torna possível o acesso à informação por utilizadores não técnicos, garantindo assim uma maior autonomia face aos recursos humanos informáticos da organização.

No final, são apresentados os resultados obtidos pela utilização do protótipo.

Palavras-chave

Data Warehouse

Sistema de Informação

Modelo Dimensional

A DATA WAREHOUSE IMPLEMENTATION STUDY FOR A HIGHER EDUCATION INSTITUTION

ABSTRACT

Nowadays organizations in general and higher education institutions in particular produce and manage high volumes of data.

In this work, we study a Data Warehouse implementation for a higher education institution, in order to endow it with an analytical tool that assists its managers in the process of decision making.

To demonstrate that this approach has advantages for the people in the organization that need to treat the information, increasing the quality and easiness of access to that information and maximizing the type and the amount of analyses that can be made to the data, a prototype was developed based in the BUS architecture and in Dimensional Modeling.

The solution developed consists in a tool which enables the access to information by non-technical users, therefore allowing an increased independence from the organization's technical IT staff.

Keywords

Data Warehouse

Information System

Dimensional Model

A presente proposta de Dissertação de Mestrado preenche o requisito parcial para a obtenção do grau de Mestre em Engenharia Informática, na Universidade de Évora.

Agradecimentos

Aos meus pais

À Filipa Pires

Ao Professor Doutor João Vinagre Santos

ÍNDICE

1. Introdução 1

1.1 Apresentação 1

1.2 Objectivos 2

1.3 Estrutura 3

2. Fundamentação (considerações teóricas) 5

2.1 Sistemas de Informação 5

2.2 Sistemas Transaccionais 8

ERP – Enterprise Resources Planning 9

2.3 Sistemas Analíticos 10

O Data Warehouse 12

Arquitectura 14

Componentes 17

2.4 O Modelo Dimensional 19

Processo de negócio 21

Tabelas de Factos 23

Tipos de Factos 23

Chaves na tabela de Factos 25

Tipos de tabelas de Factos 25

Tabelas de factos sem Factos 27

Tabelas Dimensão 28

Hierarquias 29

Dimensões de alteração lenta 32

Minidimensões 34

A Dimensão Data 34

Dimensões "role-playing" 36

Dimensões degeneradas 37

Dimensão floco de neve (snowflaking, em inglês) 37

As Datas no Modelo Dimensional 38

3. O Problema 39

3.1. Motivação 39

3.2. Caracterização do sistema existente 41

Componentes do Sistema de Informação 44
4. O Modelo Desenvolvido 49
4.1. Construção do Modelo Dimensional 50
Planeamento e Gestão do projecto 51
Definição dos requisitos do negócio 54
A matriz BUS inicial 60
Eixo de Dados 62
Modelação Dimensional 62
4.2. Construção do Protótipo 77
Eixo de Dados 78
Configuração e Desenho Físico 78
Área de Estágio e Desenvolvimento 89
Eixo Tecnológico 95
Eixo Analítico 97
Implementação 98
Manutenção e crescimento 98
5. Apresentação de Resultados 99
5.1. Caso de estudo 99
5.2. Análises efectuadas 105
6. Conclusão e estudos futuros 107
6.1. Estudos Futuros 109
7.Referências 111
Anexos 115
I. Scripts de criação do Data Warehouse 115
II. ETL – Gráficos de transformação 121
III. Bases de dados relacionais 129

ÍNDICE DE FIGURAS

Fig. 1 – Relações entre dados, informação e conhecimento 6

Fig. 2 – Sistema de infomação e seus utilizadores 8

Fig. 3 – Exemplo de uma transacção 9

Fig. 4 – Relação entre OLTP e OLAP 11

Fig. 5 – Distribuição por região geográfica de vendas de BI em 2006 13

Fig. 6 – Exemplo de um esquema BUS 15

Fig. 7 – Componentes do Data Warehouse 18

Fig. 8 – Exemplo de um esquema em estrela 20

Fig. 9 – Estrutura hierárquica 21

Fig. 10 – Processo de requisição de material 22

Fig. 11 – Exemplo de uma Tabela de Factos 23

Fig. 12 – Tabela de Factos de sumarização acumulada 26

Fig. 13 – Tabela de factos sem factos (para registo de ocorrências) 27

Fig. 14 – Exemplo de uma tabela Dimensão 29

Fig. 15 – Estrutura hierárquica com profundidade fixa 29

Fig. 16 – Exemplo de hierarquias para a dimensão Data 30

Fig. 17 – Hierarquia no modelo relacional (normalizada) 31

Fig. 18 – Estrutura hierárquica com profundidade variável 31

Fig. 19 – Exemplo de uma Minidimensão 34

Fig. 20 – Exemplo de dimensão Role Playing 36

Fig. 21 – Exemplo de dimensão em floco de neve 37

Fig. 22 – Dimensão outtrigger "região NUTS II" para a Dimensão Aluno 38

Fig. 23 – Diagrama conceptual da arquitectura do SI 43

Fig. 24 – Diagrama de Use Cases do módulo Gestão de Recursos Humanos 46

Fig. 25 – Diagrama de Use Cases do módulo Gestão Académica 47

Fig. 26 – Diagrama de Use Cases do módulo Produtos e Serviços 48

Fig. 27 – Diagrama de Use Cases do módulo Serviço Docente 48

Fig. 28 – Ciclo de vida do desenvolvimento dimensional da organização 50

Fig. 29 – Fluxograma da definição de requisitos de negócio 54

Fig. 30 – Organograma da ESTBarreiro/IPS 55

Fig. 31 – Diagrama UML com Use Cases para o Data Warehouse 58

Fig. 32 – Matriz de prioritização 61

Fig. 33 – Estrutura organizativa por Áreas Científicas 66

Fig. 34 – Modelo dimensional inicial do processo "Percurso Académico" 70

Fig. 35 – Tabela Ponte para a dimensão Corpo Docente 76

Fig. 36 – Modelo Dimensional para o processo de negócio "Percurso Académico" 77

Fig. 37 – Agregação Semestral – por número de inscrição 87

Fig. 38 – Agregação Semestral (detalhe por tipo de Inscrição/avaliação) 87

Fig. 39 – Processo ETL – Nível 1 90

Fig. 40 – Processo ETL – Nível 2 91

Fig. 41 – Processo ETL – Nível 3 (Dimensões) 92

Fig. 42 – Processo ETL - Nível 3 (Área de estágio) 93

Fig. 43 – Processo ETL - Nível 3 (Tabela de factos) 94

Fig. 44 – Arquitectura técnica do modelo desenvolvido 96

Fig. 45 – Componentes de BI do Microsoft SQL Server 2008 97

Fig. 46 – Integração do Sharepoint no BI da Microsoft 99

Fig. 47 – Ligação do Excel ao Analysis Services 100

Fig. 48 – Portal Sharepoint da ESTBarreiro/IPS 100

Fig. 49 – Barra de navegação do Portal 101

Fig. 50 – Ligações automáticas ao Data Warehouse 101

Fig. 51 – Livro de Excel com conexão ao Data Warehouse 102

Fig. 52 – Exemplo de análise de dados em Excel 102

Fig. 53 – Gráfico exemplo de análise de dados em Excel 103

Fig. 54 – Gravação da folha de Excel nos serviços de Excel 103

Fig. 55 – Publicação do objecto de Excel no portal 104

Fig. 56 – Apresentação dos resultados no portal 104

Fig. 57 – Modelo 46: Ficha estatística da Unidade Curricular 105

Fig. 58 – Evolução dos resultados por Unidade Curricular 106

Fig. 59 – Distribuição de inscrições por ano lectivo 106

ÍNDICE DE TABELAS

Tabela 1 – Vendas mundiais de sistemas ERP a empresas médias 10

Tabela 2 – Receitas mundiais de venda de ferramentas de BI 12

Tabela 3 – Aspectos principais dos modelos de Bill Inmon e Ralph Kimball..... 16

Tabela 4 – Exemplos de aditividade em factos..... 24

Tabela 5 – Exemplo da Dimensão Funcionário com uma alteração Tipo 1 32

Tabela 6 – Exemplo da Dimensão Funcionário com uma alteração Tipo 2 33

Tabela 7 – Exemplo da Dimensão Unidade Curricular com uma alteração Tipo 3 . 33

Tabela 8 – Exemplo de registos para a Dimensão Data..... 35

Tabela 9 – Vendas mundiais de plataformas de Data Warehouse por fabricante... 41

Tabela 10 – Identificação de Use Cases por Actor..... 45

Tabela 11 – Plano de execução do projecto..... 53

Tabela 12 – Lista de principais entrevistas para desenvolver o modelo 56

Tabela 13 – Resumo dos temas de análise identificados 57

Tabela 14 – Identificação dos Processos de negócio 59

Tabela 15 – Matriz BUS Inicial 60

Tabela 16 – Atributos da dimensão "Tipo de Inscrição e Estado de Avaliação"..... 68

Tabela 17 – Lista de Dimensões do Processo de negócio "Percurso Académico" . 69

Tabela 18 – Factos do processo de negócio “Percurso Académico” 69

Tabela 19 – Atributos da Dimensão Data 71

Tabela 20 – Atributos da Dimensão Semestre 71

Tabela 21 – Atributos da Dimensão Professor 72

Tabela 22 – Atributos da Dimensão Aluno 73

Tabela 23 – Atributos da Dimensão Curso 74

Tabela 24 – Atributos da Dimensão Tipo de Inscrição e Estado de Avaliação..... 74

Tabela 25 – Atributos da Dimensão Área Científica 75

Tabela 26 – Atributos da Dimensão Unidade Curricular 75

Tabela 27 – Dimensão Data: tabela física 80

Tabela 28 – Dimensão Área Científica: tabela física 81

Tabela 29 – Dimensão Semestre: tabela física 81

Tabela 30 – Dimensão Unidade Curricular: tabela física 82

Tabela 31 – Dimensão Tipo Inscrição e Estado de Avaliação: tabela física 82

Tabela 32 – MiniDimensão Curso: tabela física..... 83

Tabela 33 – Dimensão Aluno: tabela física 83

Tabela 34 – Dimensão Professor: tabela física 84

Tabela 35 – Dimensão Corpo Docente: tabela física..... 85

Tabela 36 – Tabela de Factos do modelo desenvolvido..... 85

Tabela 37 – Tabela de Auditoria..... 88

1. INTRODUÇÃO

1.1 APRESENTAÇÃO

As tecnologias de informação são uma ferramenta cada vez mais importante no seio de uma organização, na medida em que suportam ou apoiam o funcionamento das suas mais variadas áreas de negócio, influenciando o modo como as tarefas são desempenhadas e, consequentemente, os resultados daí obtidos. A importância que colocamos na gestão deste recurso deverá rivalizar com a gestão de outros recursos fundamentais como os recursos financeiros, os equipamentos ou mesmo os recursos humanos da organização. Por outro lado, a velocidade a que o mundo "corre" actualmente faz com que os gestores não tenham um minuto a perder nas decisões que necessitam tomar.

Desta forma, o modo como gerimos este "precioso" recurso pode determinar o sucesso ou fracasso do negócio. As Instituições de Ensino, e particularmente as de Ensino Superior, não são excepção. É necessário garantir que a informação é tratada de uma forma coerente de acordo com as necessidades de cada elemento da organização, garantindo o fácil acesso não só à informação mais actual mas também à armazenada ao longo do funcionamento do negócio. Isto significa que cada elemento deve ser munido da informação necessária para desempenhar o seu papel no negócio da forma mais eficaz e eficiente possível, pois assim ele próprio se torna um ponto forte da engrenagem e não mais um grão de areia que lhe cause entropia.

Esta necessidade aumenta à medida que se sobe na hierarquia da organização, pois com o aumento de importância dos elementos para a mesma, aumentam também as consequências que uma decisão errada pode acarretar.

Esta dissertação aborda o desenvolvimento de um *Data Warehouse* aplicado a uma Instituição de Ensino Superior. É objectivo demonstrar que esta abordagem tem vantagens para os elementos da organização que necessitam de tratar a informação, aumentando a qualidade do acesso à mesma e maximizando o tipo e a quantidade de análises que podem efectuar aos dados. A solução desenvolvida consiste numa ferramenta para auxiliar os gestores da organização no processo de tomada de decisões, através da análise cuidada da informação disponível.

As Instituições de Ensino Superior produzem informação de um conjunto de áreas muito alargado, onde podemos destacar a área financeira, a área relacionada com a gestão do pessoal, a área relacionada com os produtos e serviços oferecidos pela instituição e a gestão da informação académica dos seus "clientes". Dado o leque tão diverso de temas e áreas abrangidas, é sobre esta última àrea de actividade que este trabalho se irá focalizar em termos de desenvolvimento do protótipo: a informação académica dos alunos, com especial enfoque sobre a informação que permita analisar o seu sucesso/insucesso escolar.

Para construir este Data Warehouse foi utilizado o modelo dimensional desenvolvido por Ralph Kimball (Kimball & Ross, 2002), que consiste numa arquitectura de organização da informação que permite construir os vários *data marts* que possam compor o Data Warehouse. Estes *Data Marts* são constituídos por tabelas de factos e tabelas de dimensão, responsáveis por conter, respectivamente, os atributos numéricos mensuráveis e os dados textuais descritivos do negócio.

O sistema que fornece os dados ao data warehouse é um sistema de informação baseado num conjunto de bases de dados relacionais, desenvolvidas à medida para a Instituição de Ensino Superior, envolvendo também uma componente de integração de dados dispersos pelas várias àreas de negócio da organização e cujo formato (ficheiros de folha de cálculo, documentos de texto, etc.) não se encontra estruturado ao nível das bases de dados.

As bases de dados relacionais utilizadas pelo Sistema de Informação foram desenvolvidas utilizando o motor de bases de dados Microsoft SQL Server. O protótipo desenvolvido para o Data Warehouse utilizou o pacote de ferramentas de *Business Intelligence* da Microsoft (também designado por BI). Este pacote é constituído por três ferramentas que iremos abordar no capítulo 4, sobre o modelo desenvolvido, e que permitem:

- O desenvolvimento da solução analítica de Data Warehouse;
- Elaboração de relatórios;
- O desenvolvimento de aplicações de integração de sistemas heterogéneos.

Quanto a *front-end* do protótipo, foi utilizado o portal Microsoft Office SharePoint Server 2007 (MOSS 2007). Por último, as análises de dados foram efectuadas utilizando o Microsoft Office Excel, na versão 2007. Estas ferramentas serão abordadas em detalhe no Capítulo 5, destinado à apresentação de resultados.

1.2 Objectivos

Como anteriormente referido, o objectivo deste trabalho é desenvolver um *Data Warehouse* para uma Instituição de Ensino Superior. A instituição utilizada como modelo para o desenvolvimento do Data Warehouse é a Escola Superior de Tecnologia do Barreiro, do Instituto Politécnico de Setúbal – adiante designada por ESTBarreiro/IPS.

São objectivos específicos a atingir com este trabalho:

- Estudar o modelo de dados a implementar para a construção do *Data Warehouse*;
- Criar uma ferramenta que permita fornecer os dados necessários ao processo de decisão, disponibilizando a informação de uma forma que se pretende simples e organizada, coerente com este tipo de sistemas;

- Validar os dois pontos anteriores através do desenvolvimento de um protótipo.

Com vista ao desenvolvimento deste projecto, foi efectuado um estudo que envolveu desde o levantamento das necessidades da Instituição ao nível do armazenamento de dados - analisando o conjunto de bases de dados relacionais existentes que compõem o sistema de informação transaccional existente - e das decisões que frequentemente são tomadas, até à especificação do modo como a informação circula entre os vários intervenientes nos processos de decisão. Para o efeito foram conduzidas várias entrevistas aos responsáveis dos vários departamentos que, em conjunto com as técnicas de construção de *Data Warehouses*, permitiram identificar os processos de negócio mais importantes para a organização e que importam modelar.

1.3 Estrutura

Em termos de estrutura, este trabalho está dividido em 5 partes. Em primeiro lugar é feito um enquadramento teórico onde se caracterizam os sistemas de informação actuais nas instituições de Ensino Superior. São apresentados os sistemas transaccionais (OLTP – *On-Line Transaction Processing*) e analiticos (OLAP – *On-Line Analytical Processing*), discutindo-se a sua aplicação. É também apresentado o modelo dimensional, abordando os fundamentos teóricos da sua implementação.

No capítulo seguinte apresenta-se o problema que este trabalho pretende abordar e contribuir para a respectiva solução, a disponibilização da informação a quem dela necessita para a tomada de decisões cada vez mais informadas. É feita a caracterização do sistema existente, baseado no modelo transaccional, explicando de que forma este vai ser utilizado para a construção do protótipo de Data Warehouse.

Em seguida, apresenta-se o modelo desenvolvido, caracterizando detalhadamente todas as suas componentes e explicando as fases do seu desenvolvimento. É apresentado o protótipo desenvolvido.

Por último, são apresentados os resultados obtidos utilizando um caso de estudo baseado no protótipo desenvolvido e as várias análises efectuadas aos dados obtidos. É apresentada a conclusão do trabalho e delineado o trabalho futuro.

2. FUNDAMENTAÇÃO (CONSIDERAÇÕES TEÓRICAS)

2.1 SISTEMAS DE INFORMAÇÃO

A informação é um recurso fundamental para qualquer organização. A capacidade de gerir este precioso recurso permite atestar a capacidade para enfrentar os desafios com que as organizações se deparam.

As instituições de Ensino Superior não são excepção. Constantemente recolhem e armazenam elementos relacionados com os mais variados aspectos do seu funcionamento, de forma a conseguir acrescentar valor às decisões que tomam e melhorar o serviço prestado aos seus utentes. E é na forma como posteriormente são tratados os elementos recolhidos que pode estar a diferença entre tomar uma decisão acertada ou escolher um caminho menos certo.

É aqui que actuam os sistemas de informação, na recolha dos dados da forma mais completa e robusta, no tratamento e organização desses dados de uma forma eficaz, e na disponibilização do resultado desse anterior tratamento a quem deles necessita, no momento em que necessita. Segundo Turban e Rainer (2009), os sistemas de informação têm como objectivo apresentar a informação certa às pessoas certas, no tempo certo, em quantidade certa e no formato correcto.

No entanto, ao falarmos de sistemas de informação, rapidamente sentimos necessidade de definir alguns conceitos fundamentais. É comum, quando se discutem sistemas de infomação, ouvir falar em termos como dados, informação ou conhecimento. Estes termos, embora normalmente interligados, são de natureza diferente e podem ser definidos de várias formas. Uma definição genérica é proposta por (Turban & Aronson, 1997):

- Dados – são os factos sobre as coisas, eventos, actividades e transacções que são registados, classificados e armazenados mas que não estão organizados de forma a possuir um significado específico. Os dados podem ser numéricos, alfanuméricos, figuras, sons ou imagens;
- Informação – a informação surge quando os dados são organizados de forma a ter significado para o seu recipiente. Pode dar-nos a conhecer algo de novo ou confirmar suspeitas já existentes sobre determinado aspecto do negócio;
- Conhecimento – consiste em dados organizados e processados para fornecer compreensão, experiência, aprendizagem acumulada e qualificação técnica à medida que os mesmos se aplicam a um problema ou actividade corrente. O conhecimento pode traduzir-se como a aplicação dos dois tópicos anteriores em conjunto para a tomada de uma decisão.

A informação é também definida por Peter Drucker (1988) como "dados dotados com relevância e funcionalidade", as organizações transformam os dados em

informação dotando-os de um contexto que permita a sua análise com significado. Por exemplo, o registo de um conjunto de notas armazenado num determinado suporte físico sem qualquer relacionamento não passa de um conjunto de dados. Mas se existir a relação com outros dados como por exemplo os alunos que obtiveram essas notas, as unidades curriculares associadas, o tipo de exame realizado, etc. permite tornar este conjunto de dados em informação. A análise efectuada à evolução das notas a uma unidade curricular pode resultar num processo de aquisição de conhecimento sobre determinado aspecto ou actividade em estudo. A Fig. 1 apresenta a evolução do processo que permite transformar dados em conhecimento.

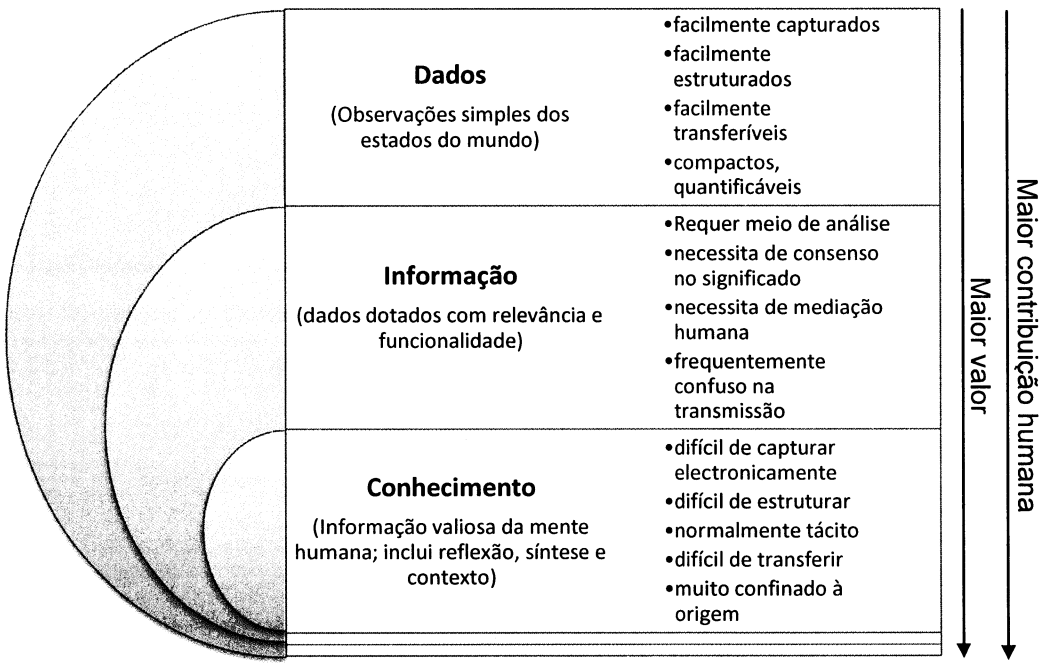


Fig. 1 – Relações entre dados, informação e conhecimento

(Pearlson & Saunders, 2009)

Os Sistemas de Informação (adiante designados por SI) surgiram na década de 60 e começaram por ser utilizados para satisfazer as necessidades internas das organizações. Saunders e Pearlson dividem a evolução dos SI em 5 eras (Pearlson & Saunders, 2009). Em primeiro lugar (Era I, década de 60) pretendeu-se automatizar os processos manuais realizados normalmente em papel e reduzir os custos relacionados com os produtos e serviços transaccionados – os modelos de informação eram centrados nas aplicações e os sistemas utilizados eram os *mainframes*. Depois, na década de 70, Era II, surgiu a necessidade de aumentar a produtividade e melhorar a qualidade das decisões aumentando a informação disponível aos decisores – os modelos de informação passaram a ser orientados aos dados. Na Era III, década de 80, a preocupação passou a ser a vantagem competitiva face à concorrência, visto a construção de SI estar já disseminada nas organizações, começou a olhar-se para o

SI como ferramenta estratégica da organização e o enfoque passou a ser o aumento da contribuição do indivíduo e do grupo para a organização. Na década de 90, Era IV, o SI reforçou a sua importância estratégica, passando a ser utilizado como ferramenta para transformar a forma de fazer negócio da organização, o enfoque passou a ser o processo de negócio e o sistema passou a possuir uma inteligência distribuída, baseada numa arquitectura cliente-servidor. Finalmente, na Era V, que começa na década de 2000, o papel principal dos SI passa a ser a criação de valor, o modelo de informação passa a ter uma orientação para o conhecimento e a preocupação da organização passa a ser a relação com o cliente. A tecnologia dominante passa a estar centrada na internet e a arquitectura do sistema começa a utilizar inteligência ubiqua[1].

Esta evolução permite observar que os SI desempenham um papel cada vez mais vital no funcionamento da organização, o que conduz a sua correcta gestão a uma vantagem competitiva sobre a concorrência.

Mas se a maior parte das organizações, de uma forma ou de outra, hoje possuem e utilizam sistemas de informação, torna-se relevante perceber de que forma é que essa vantagem competitiva se traduz. E a resposta está na forma como os sistemas de informação são utilizados: podemos ter o equipamento mais sofisticado do mercado que, se não o utilizarmos da forma correcta e explorando todas as suas capacidades, não iremos tirar partido da vantagem que à partida parecíamos ter.

Quer isto dizer que só a eficiente e eficaz utilização dos recursos de informação de uma organização pode tornar o seu sistema de informação uma capacidade competitiva sobre os seus concorrentes. A Fig. 2 apresenta as componentes de um SI de uma organização, apresentando a relação dos seus colaboradores com o mesmo. O SI é composto por várias componentes ou módulos separados por área funcional na parte superior da pirâmide e na base temos os sistemas transaccionais e eventual ERP (sigla para *Enterprise Resurces Planning* – estes sistemas serão abordados no próximo subcapítulo, Sistemas Transaccionais) que permite a integração de todas as componentes do SI. Nas laterais da pirâmide podemos analisar à direita os vários tipos de colaboradores dentro da organização e, à esquerda, as respectivas ferramentas do SI com que cada um desempenha as suas funções. À medida que nos aproximamos do topo, as ferramentas utilizadas passam de operacionais a um cariz mais analítico, como poderemos analisar na Fig. 2.

[1] A computação ubíqua é uma área integradora de várias competências – programação, engenharia de software, sistemas de informação, inteligência artificial, redes, sistemas distribuídos, etc.

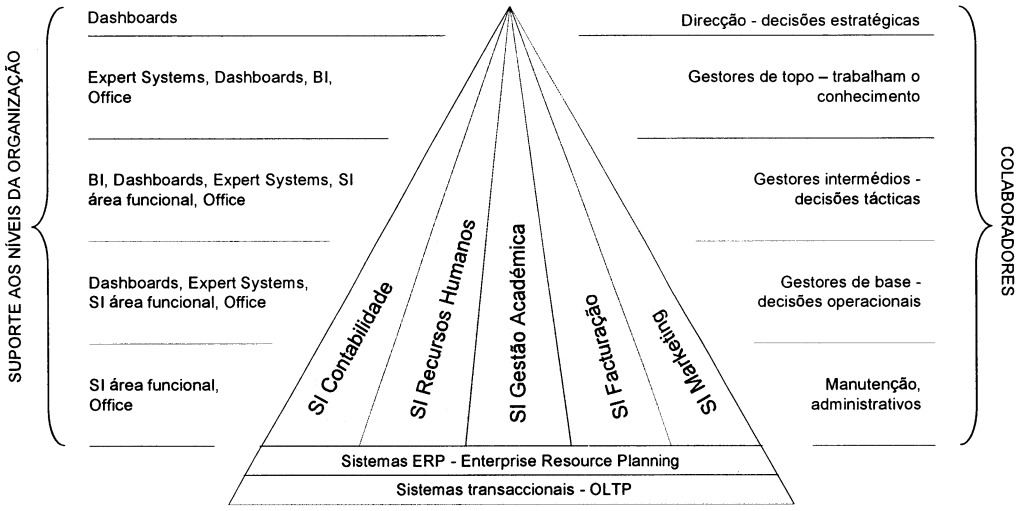


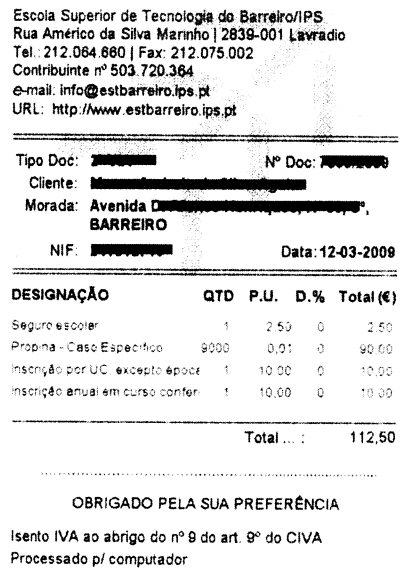
Fig. 2 – Sistema de infomação e seus utilizadores

(adaptado de Turban & Rainer, 2009)

2.2 SISTEMAS TRANSACCIONAIS

Nas organizações em geral e nas instituições de Ensino Superior em particular ocorrem centenas (ou mesmo milhares, dependendo da dimensão) de transacções por dia. Entenda-se por transacção todo o evento que produza qualquer dado com relevância para o negócio e interesse de recolha e armazenamento numa base de dados, ou noutros tipos de ficheiros transaccionais.

Os sistemas transaccionais ou OLTP – *On-Line Transaction Processing* são, como o próprio nome indica, sistemas orientados à transacção, isto é, são desenvolvidos com o objectivo de registar as operações isoladas, efectuadas no dia-a-dia pelos vários sectores da organização, como por exemplo o registo de inscrição de um aluno a uma unidade curricular ou o pagamento de um serviço. São sistemas que garantem a operacionalidade do negócio, normalmente também designados por sistemas operacionais. Como a Fig. 2 representa, os sistemas de cariz operacional são essencialmente utilizados pelos colaboradores da base da pirâmide. Não quer isto dizer que a importância destes colaboradores para a organização seja menor, é precisamente na qualidade da informação introduzida no SI que está a sua riqueza: se nesta fase não reduzirmos ao máximo os erros e omissões de introdução de dados, o trabalho das fases posteriores será normalmente mais dificultado e as análises de informação poderão não ter a qualidade desejada ou mesmo ficar comprometidas. Outro aspecto que permite relevar a importância dos colaboradores na base da pirâmide é o seu contacto directo com os utilizadores e a percepção que podem ter de novos dados que possamos recolher e que permitam enriquecer a qualidade da análise efectuada.



Escola Superior de Tecnologia do Barreiro/IPS
Rua Américo da Silva Marinho | 2839-001 Lavradio
Tel.: 212.064.660 | Fax: 212.075.002
Contribuinte nº 503.720.364
e-mail: info@estbarreiro.ips.pt
URL: http://www.estbarreiro.ips.pt

Tipo Doc: [redacted] Nº Doc: [redacted]
Cliente: [redacted]
Morada: Avenida [redacted], BARREIRO
NIF: [redacted] Data: 12-03-2009

DESIGNAÇÃO	QTD	P.U.	D.%	Total (€)
Seguro escolar	1	2,50	0	2,50
Propina - Caso Específico	9000	0,01	0	90,00
Inscrição por UC, excepto época	1	10,00	0	10,00
Inscrição anual em curso confer	1	10,00	0	10,00

Total ... : 112,50

OBRIGADO PELA SUA PREFERÊNCIA

Isento IVA ao abrigo do nº 9 do art. 9º do CIVA
Processado p/ computador

Fig. 3 – Exemplo de uma transacção

Como referido no ínicio, os sistemas transaccionais são habitualmente de utilização diária. Quer isto dizer que a recolha da informação é efectuada praticamente em tempo real, fazendo depender o funcionamento do serviço da disponibilidade destes sistemas. Este aspecto torna o sistema transaccional um factor crítico para a organização, não só em termos do funcionamento do SI como uma ferramenta global dentro da mesma mas também na relação da organização com os seus clientes, pois o sistema transaccional também tem a função de servir de interface entre a organização e estes últimos (pode nem envolver a presença de colaboradores da organização, como por exemplo no caso dos quiosques de aquisição de serviços).

As bases de dados mais recentes utilizam normalmente o modelo relacional no seu desenvolvimento, e devem ser realizadas com vista a optimizar a introdução de informação, minimizando a redundância (através da normalização) e a garantir a sua disponibilidade, através por exemplo da utilização de índices e vistas que permitam o acesso pontual aos registos da base de dados.

ERP – Enterprise Resources Planning

Os ERP's são uma ferramenta utilizada para permitir a integração dos SI das várias áreas funcionais da organização. Este problema ocorre normalmente quando o sistema de informação da organização é composto por vários SI de área funcional que não comunicam entre si, quer seja por questões tecnológicas ou mesmo como resultado de um mau planeamento. Este problema tem menor tendência para ocorrer quando o SI da organização é desenvolvido à medida das necessidades, em que os vários módulos do SI que compõe cada área funcional são desenvolvidos com comunicação entre eles, estando o próprio desenvolvimento de cada um dependente da forma como os restantes funcionam. Neste caso, o SI da organização funciona desde o início com as funcionalidades que um sistema ERP permitiria usufruir.

Normalmente os ERP são também baseados em sistemas transaccionais compostos por um repositório de dados central e vários módulos de software comunicantes entre si. O quadro abaixo apresenta os 5 principais fornecedores mundiais de ferramentas ERP nos anos 2005 e 2006, em termos de volume de vendas, a pequenas e médias empresas (entre 1 e 499 trabalhadores).

Tabela 1 – Vendas mundiais de sistemas ERP a empresas médias

IDC (2007a)

Fornecedor	Vendas em 2005 (milhões dólares)	Vendas em 2006 (milhões dólares)	Share de mercado em 2006 (%)
SAP	1.926	2.013	22,6
Oracle	813	1.228	13,8
Infor	580	706	7,9
Lawson Software	250	289	3,2
Microsoft	235	288	3,2
TOTAL	3.804	4.524	50,7

2.3 Sistemas Analíticos

Enquanto os sistemas transaccionais se destinam a assegurar a operacionalidade do negócio, os sistemas analíticos permitem colocar os recursos informacionais produzidos por esses sistemas ao serviço dos decisores da organização. Comummente designados por sistemas OLAP – *On-Line Analytical Processing*, os sistemas analíticos surgiram da necessidade de utilizar a informação gerada pela organização para auxiliar os gestores nas suas tarefas de análise e decisão estratégica. São ferramentas de análise e não ferramentas de registo operacional.

Estes sistemas têm como característica o facto de estarem orientados para o tratamento de grandes volumes de dados e de necessitarem, na maior parte das vezes, de utilizar simultaneamente diferentes e diversas fontes de dados para proceder às análises desejadas. A utilização dos dados acontece a um nível de detalhe completamente diferente dos sistemas OLTP – em vez de se ocupar da manipulação de operações atómicas, como por exemplo os items de uma factura, as bases de dados OLAP analisam informação organizada por período ou situação específica da empresa (por exemplo, a evolução das notas obtidas a uma unidade curricular ao longo de um determinado período de tempo). Importa também destacar que as operações realizadas sobre OLAP são maioritariamente consultas em deterimento de introduções ou alterações de dados.

A Fig. 4 apresenta a forma como os dois sistemas se relacionam.

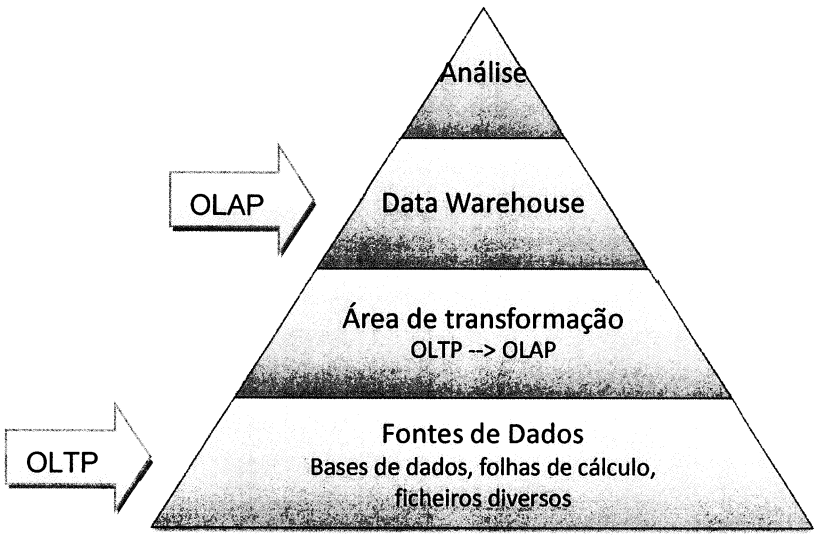


Fig. 4 – Relação entre OLTP e OLAP

(Caldeira, 2008)

Dada a sua diferença face aos sistemas operacionais de transaccção, as bases de dados OLAP necessitam de possuir uma organização da informação também diferente, que permita armazenar grandes volumes de informação mas ao mesmo tempo possibilite a sua consulta de uma forma que se pretende eficaz, robusta e o mais simples possível.

De acordo com Cortes (2005), existem algumas situações que devemos evitar e que podem ocorrer se optarmos por construir um sistema analítico baseado numa base de dados de cariz operacional:

- Incapacidade de integração – o nível de detalhe pretendido é diferente nos dois sistemas;
- Comprometimento do desempenho do sistema operacional – ao passar a ter de lidar com interrogações de cariz analítico a performance baixa, pois o sistema está optimizado para interrogações delimitadas;
- Incapacidade de utilizar outras fontes de informação para proceder à análise – um sistema analítico agrega normalmente informação transversal a toda a organização, sendo possível essa informação existir em diferentes sistemas transaccionais ou mesmo em fontes de dados não estruturadas como ficheiros de texto ou tabelas de folha de cálculo.

Podemos ainda acrescentar aos pontos anteriores as questões de segurança: quem utiliza os sistemas analíticos são, como vimos na Fig. 2, os directores e decisores de topo da organização, enquanto os sistemas transaccionais podem ser utilizados por todos os outros colaboradores. Se o sistema transaccional se "misturar" com o analítico, será mais difícil conseguir um controlo eficaz sobre quem acede a que informação.

De entre as ferramentas que podemos utilizar para construir um sistema OLAP podemos encontrar o Data Warehouse, objecto de estudo desta dissertação. A informação das várias fontes de dados utilizadas pela organização é recolhida, organizada, transformada de modo a ter um sentido lógico para a(s) actividade(s) que se deseja(m) analisar, limpa de erros que possam existir e carregada no Data Warehouse, que por sua vez irá funcionar como *input* para as ferramentas de análise que a organização utilizar, como por exemplo os sistemas de apoio à decisão ou as ferramentas de *Business Intelligence* (também designadas por BI).

O DATA WAREHOUSE

O Data Warehouse surge assim como a ferramenta utilizada para permitir a disponibilização da informação aos analistas da organização. O interesse ou preocupação do analista não é o registo de uma qualquer operação isolada ou a forma como a organização realiza uma qualquer tarefa mas sim o estudo da evolução de determinada componente ou processo da organização. Para tal é necessário que, em primeiro lugar, possamos combinar informação das várias fontes disponíveis para o negócio, quer sejam fontes internas ou externas[2], e em seguida devemos conseguir tratar essa informação para que a mesma tenha um significado lógico e coerente. Por último, devemos disponibilizar aos decisores esse conhecimento de uma forma simples de perceber e de fácil acesso.

A Tabela 2 apresenta as vendas mundiais de ferramentas de análise de dados, divididas por dois grandes grupos: as ferramentas de consulta, relatórios e análise e as ferramentas de análise avançada.

Tabela 2 – Receitas mundiais de venda de ferramentas de BI

IDC (2007b)

	Receitas (milhões de dólares)			Quota (%)			Crescimento (%)	
	2004	2005	2006	2004	2005	2006	2004-2005	2005-2006
Ferramentas de consulta, relatório e análise	4.004,9	4.487,6	5.008,5	79,5	80,0	80,1	12,1	11,6
Análise avançada	1.031,9	1.118,6	1.244,6	20,5	20,0	19,9	8,4	11,3
TOTAL	5.036,7	5.606,2	6.253,0	100,0	100,0	100,0	11,3	11,5

No primeiro grupo estão incluídas as consultas ad-hoc, os dashboards, servidores OLAP e Data Warehouses, Data Marts, etc. O segundo grupo inclui as ferramentas de data mining e software estatístico, tecnologias como as redes neuronais, indução de

[2] Segundo Turban e Aronson (1997), as fontes de informação podem ser a) internas – sistemas transaccionais da organização ou b) externas – informação de fontes externas à organização (ex. INE, DGES,etc.).

regras e *clustering*. Como podemos verificar, estas tecnologias crescem de uma forma sustentada, apresentando um aumento percentual de cerca de 12% ao ano.

No gráfico da Fig. 5, podemos observar a distribuição mundial em termos geográficos das vendas da Tabela 2, o que permite interpretar os valores de uma forma contextualizada.

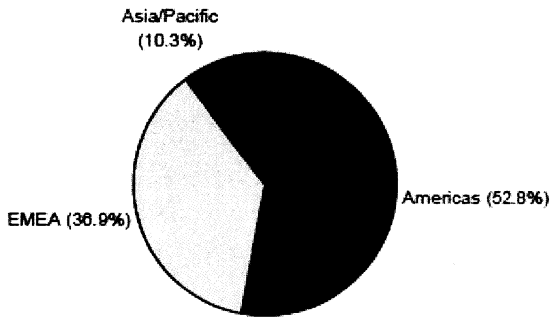


Fig. 5 – Distribuição por região geográfica de vendas de BI em 2006

IDC (2007b)

O Data Warehouse possui um conjunto de características importantes (Misner, 2009; Turban & Rainer, 2009):

- Histórico – a informação no Data Warehouse resulta do registo de funcionamento da organização ao longo do tempo nas diferentes áreas de negócio ou processos em análise, a informação é normalmente armazenada por vários anos. Torna-se assim possível construir uma base histórica que permite estudar comportamentos, evoluções ou até efectuar previsões. A utilidade da informação e importância do Data Warehouse aumenta com o decorrer do tempo pois as análises a efectuar vão dispor de maior base de sustentação;
- Consistência – a informação do Data Warehouse provém de sistemas heterogéneos, possivelmente desenvolvidos e mantidos por diferentes equipas ou empresas. Isto leva a que a mesma informação possa ser tratada de forma diferente, por exemplo uma base de dados de pessoal pode tratar o estado civil utilizando os valores "Casado" e "Casada" enquanto outra utiliza o termo genérico "Casado(a)". No Data Warehouse a informação é normalizada para obter consistência nos dados armazenados. É nestes casos que surge a importância da utilização de normas para classificar os dados sempre que possível, como iremos ver mais à frente;
- Agregação dos dados – num sistema analítico pretende-se analisar a informação com diferentes níveis de detalhe, como por exemplo as notas de um semestre ou aproveitamento anual por curso, devendo estas agregações estar pré construídas no Data Warehouse de forma a facilitar a utilização e potenciar o tipo de análises;

- Ligação a SGBDR – a ligação a Sistemas de Gestão de Bases de Dados Relacionais e a possibilidade de construção do próprio Data Warehouse utilizando um SGBD permite facilitar o processo de implementação do mesmo, pois não só o processo de carregamento de informação pode ser mais facilitado, também os custos serão mais reduzidos face à aquisição de sistemas proprietários que exigem formação e equipamento adicional face ao que já existe na organização;
- Forma de organização – a organização é realizada por assuntos ou características do negócio (informação disponível por produto ou por aluno, por exemplo) que permite apresentar a informação relevante para a análise de dados e o suporte à tomada de decisões.

ARQUITECTURA

Podemos implementar o Data Warehouse de acordo com duas arquitecturas distintas (Caldeira, 2008; Ramamurthy, Sen, & Sinha, 2008): o Data Warehouse Empresarial ou o Data Warehouse composto por Data Marts conformes – também designada por arquitectura BUS.

O primeiro é proposto por Bill Inmon[3] (Inmon, 2005) e assenta no conceito de visão global da organização: o Data Warehouse é um repositório de dados central único, construído como uma base de dados relacional, normalizada até à 3ª forma normal. Esta arquitectura é também designada por arquitectura *top-down*, ou seja, a construção do Data Warehouse é elaborada do topo para a base da organização.

Tem a particularidade de ser uma abordagem em que o tempo de desenvolvimento é elevado: embora, segundo Bill Inmon, a visão consista em não construir o Data Warehouse baseado na teoria do "Big Bang" (Inmon, 2005), ou seja, construir tudo de uma vez, a verdade é que a elaboração do modelo para toda a organização e a construção do Data Warehouse segundo esta arquitectura é bastante mais complicado, demorado e consequentemente dispendioso até se obterem resultados. Por outro lado, a estrutura relacional é demasiado complexa para ser utilizada por pessoal não técnico ou sem formação específica na área das bases de dados relacionais. O nível de normalização utilizado é outro factor desfavorável: o elevado número de tabelas que este modelo origina leva a que o número de ligações que é necessário efectuar para conseguir consultar a informação (designadas por *join's*), para além de dificultarem a construção das consultas também têm um impacto negativo na performance das mesmas, principalmente quando estamos a consultar grandes volumes de informação.

---

[3] Considerado por muitos como o pai do Data Warehouse por ter sido a primeira pessoa a utilizar este conceito em 1991 no seu livro *"Building the Data Warehouse"* (Breslin, 2004; Caldeira, 2008)

A outra arquitectura, designada por arquitectura em BUS foi desenvolvida por Ralph Kimball (Kimball & Ross, 2002) e consiste em elaborar o Data Warehouse como um conjunto de Data Marts consolidados construídos de forma faseada, representando cada um um processo de negócio da organização. Esta abordagem é também designada por abordagem *bottom-up*, ou da base para o topo (Breslin, 2004), visto começar por casos mais ou menos particulares no seio da organização (as vendas de produtos, o percurso académico, etc.), generalizando à medida que são adicionados mais Data Marts à solução até se atingirem todas as áreas da organização a incluir no Data Warehouse.

A arquitectura em BUS[4] assenta num esquema em estrela em que cada Data Mart é constituido por uma tabela central de indicadores ou medidas do negócio – as tabelas de factos, e um conjunto de tabelas periféricas (formando assim uma espécie de estrela), com as características textuais descritivas do mesmo – as tabelas dimensão (Moody & Kortink, 2000).

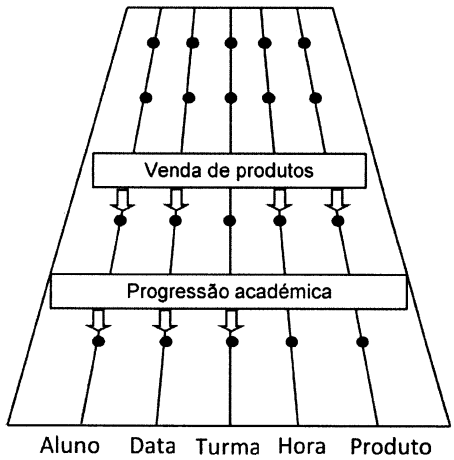


Fig. 6 – Exemplo de um esquema BUS

A interligação entre os vários Data Marts é efectuada pelas entidades comuns que existem entre eles e que permite assim construir um BUS de ligação.

Sempre que se efectuem análises envolvendo dados de mais do que um processo de negócio estamos a tirar partido desta interligação. A esta forma de consulta transversal da informação dá-se a designação de *drill-across* (ou "escavação transversal", numa tentativa de tradução para português). Por exemplo, se quisermos saber qual o valor médio que o aluno paga de propinas até concluir o curso, temos de utilizar a dimensão Aluno que participa no processo de vendas e no processo Percurso Académico. Em vez de utilizar o modelo relacional como suporte ao desenvolvimento do sistema, na arquitectura BUS o desenvolvimento dos Data Marts assenta no

[4] BUS é um termo que deriva da electrónica e que pretende definir uma rede ou caminho onde vários equipamentos estão ligados e que, consequentemente, comunicam entre si

modelo dimensional, que iremos ver em seguida. Este modelo é mais simples de utilizar, mais perceptível pelo utilizador e está direccionado para a consulta de informação.

Esta arquitectura tem a vantagem de possuir um investimento inicial reduzido, pois ao permitir a construção de forma faseada também permite o investimento faseado e a apresentação de resultados é mais célere (Abraham et al. 2004). Este factor, para além de facilitar a integração dos utilizadores, permite também que desde logo se comece a testar o funcionamento da soução. O processo de construção de um Data Warehouse é sempre um processo que resulta de uma abordagem iterativa, em que o produto final é o resultado de uma série de afinações que devem ser feitas com os utilizadores do sistema. Na Tabela 3 podemos consultar uma comparação entre as duas arquitecturas:

Tabela 3 – Aspectos principais dos modelos de Bill Inmon e Ralph Kimball.

Adaptado de Breslin (2004)

	Empresarial	Dimensional
Metodologia e arquitectura		
Abordagem geral	Top-down	Bottom-up
Estrutura arquitectónica	Data Warehouse empresarial	Conjunto de Data Marts conformes ligados por um BUS
Complexidade	Elevada	Reduzida
Comparação com metodologias existentes	Desenvolvimento em espiral[5]	Processo em 4 etapas; deriva da modelação relacional
Discussão do desenho físico	Bastante exaustivo	Relativamente ligeiro
Modelação de dados		
Orientação dos dados	Orientado ao sujeito ou aos dados	Orientado ao processo
Ferramentas	Diagrama entidade-relação	Modelação dimensional
Acessibilidade do utilizador final	Baixa	Elevada
Filosofia		
Principais destinatários	Profissionais de IT	Utilizadores finais
Espaço na organização	Parte integral do Sistema de Informação empresarial	Área de transformação e repositório de dados operacionais
Objectivo	Oferecer uma solução técnica sólida baseada em modelos de bases de dados e tecnologias comprovadas	Oferecer uma solução que permita ao utilizador final consultar directamente os dados e obter tempos de resposta razoáveis

Neste trabalho vamos utilizar o modelo Dimensional, visto ser o que permite o desenvolvimento por fases, que permite disponibilizar uma ferramenta de trabalho

[5] O desenvolvimento em espiral é um processo de desenvolvimento de software proposto por Bohem et al. (1998), que consiste num processo iterativo de resolução dos problemas.

mais rapidamente (Abraham et al. 2004), e que facilita o acesso à informação por utilizadores não especialistas.

Componentes

Em termos da implementação de um projecto de Data Warehouse, podemos identificar 4 componentes (Caldeira, 2008; Kimball & Ross, 2002):

- Os **sistemas Fonte** – são constituidos pelas fontes de dados utilizados para o carregamento do Data Warehouse. Normalmente são as bases de dados que compõem os sistemas transaccionais da organização, complementadas por informação de outras fontes, como ficheiros de folha de cálculo, tabelas estatísticas, ficheiros de texto, etc.. Esta informação existe e é recolhida no seio da organização, mas também pode ser informação externa, obtida em fontes externas, como por exemplo no caso de informação estatística (dados demográficos, tendências de consumo, etc.);
- A **área de estágio ou transformação** – nesta componente, que é considerada a mais exigente e importante da construção do Data Warehouse – se mal executada pode colocar em causa o sucesso do projecto – extrai-se a informação dos sistemas fonte, aplicam-se as operações de transformação necessárias à obtenção de conformidade e qualidade dos dados (filtros, "limpezas", reformatações, validações, eliminação de duplicados, etc.) e por fim carregam-se as bases de dados finais que constituem o Data Warehouse. A este processo dá-se o nome de ETL – *Extract, Transform and Load.* A execução desta fase, devido à elevada exigência de recursos que possui, deverá ser executada numa altura em que o Data Warehouse não esteja a ser utilizado, normalmente num período nocturno;
- A **área de apresentação** – consiste na estrutura de armazenamento dos dados para utilização pelos analistas. O Modelo Dimensional permite analisar a informação de uma forma temporal, ou seja, podemos estudar a evolução de um determinado processo da organização ao longo do tempo. O modelo dimensional deverá ser composto por estruturas autónomas que permitam representar fielmente um processo de negócio da organização – a estas estruturas, tal como já foi atrás referido, dá-se o nome de Data Mart. O conjunto dos vários Data Marts consolidados da organização dá origem ao Data Warehouse. Estas estruturas serão abordadas de forma mais pormenorizada quando abordarmos o subcapítulo Modelação Dimensional e o capítulo O modelo desenvolvido;
- O **acesso a dados** – esta última componente consiste nas ferramentas que são utilizadas para aceder aos dados existentes na área de apresentação. Podem ser ferramentas tão simples como consultas ad-hoc, de dificuldade intermédia como os dashboards ou até ferramentas complexas como

ferramentas de data-mining ou modelos de previsão matemáticos. É nesta componente que o utilizador do Data Warehouse terá contacto com a informação disponibilizada pelo mesmo.

A Fig. 7 apresenta uma representação destas componentes e suas relações.

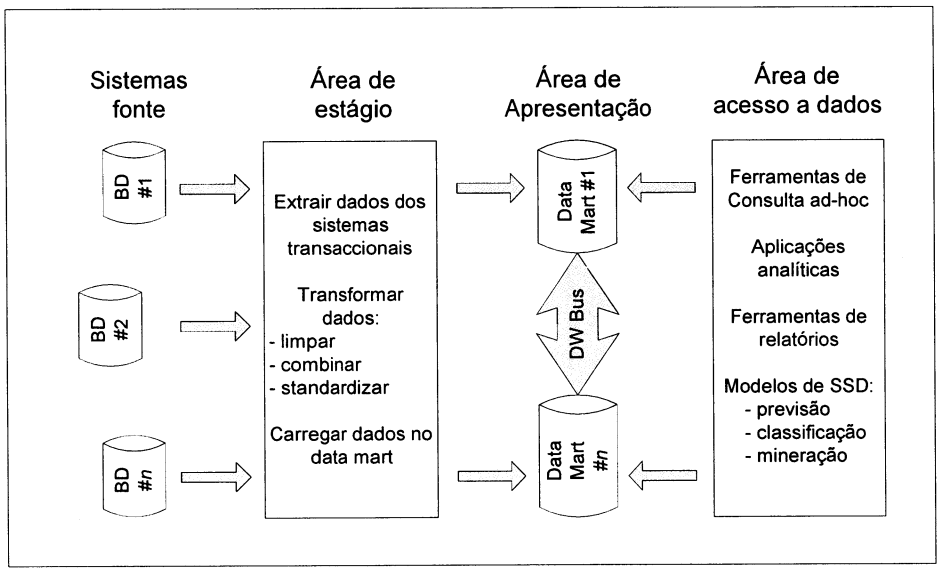


Fig. 7 – Componentes do Data Warehouse

(Kimball & Ross, 2002)

Outro aspecto importante que se deve ter em consideração quando pensamos em sistemas analíticos é a diversidade de questões colocadas pelos analistas, a que o data warehouse deve conseguir dar resposta. Para além das questões colocadas pelos analistas, estes devem poder combinar a informação de uma forma fácil, que permita a descoberta de novos indicadores - em última análise, conseguiremos cativar o analista a utilizar o Data Warehouse mesmo sem existir nenhuma questão *à priori*.

Em suma, com a implementação do Data Warehouse, pretende-se dotar a organização de uma ferramenta que permita responder às seguintes considerações (Kimball & Ross, 2002):

- O Data Warehouse deve tornar a informação da organização facilmente acessível – o utilizador deverá conseguir aceder à informação de uma forma fácil e com resultados rápidos, em que seja possível utilizar ferramentas de acesso aos dados fáceis de utilizar e compreender. A rotulação dos dados deverá obedecer a regras claras, que permitam compreender o que cada elemento significa – de preferência devem utilizar-se normas de atribuição de nomes aos atributos;
- A informação deve ser apresentada de forma consistente, ou seja, devemos ter a garantia de qualidade e credibilidade da informação disponibilizada;

- O Data Warehouse deve ser adaptável a novas realidades e resistir à mudança, condição inevitável e incontornável nas organizações de hoje;
- A segurança deve ser um dado adquirido quando nos referimos a um Data Warehouse, o controlo do acesso à informação deve ser eficaz, pois este contém a informação sensível da organização;
- O Data Warehouse deve ser utilizado como uma ferramenta que aumenta a qualidade das decisões;
- Para ser uma mais valia é imperativo que o data warehouse seja aceite pelos actores da organização: as pessoas que a compõem. Podemos ter a ferramenta tecnologicamente mais avançada do mercado que, se ninguém a utilizar, estará inevitavelmente condenada ao fracasso.

2.4 O Modelo Dimensional

Tal com já foi anteriormente referido, a construção do Data Warehouse assenta no modelo dimensional. Este modelo consiste numa forma de armazenar a informação diferente do modelo entidade-relação utilizado nas bases de dados transaccionais, na medida em que utiliza como estrutura de armazenamento o esquema em estrela, também designado por *star-schema*. O objectivo pretendido com esta forma de estruturar a informação é facilitar a obtenção de três características fundamentais do Data Warehouse: compreensão da informação pelo utilizador, performance das consultas efectuadas e resistência à mudança (Cardoso et al. 2003). Esta arquitectura possui algumas características que a tornam adequada para a obtenção destes objectivos:

- A forma simples e simétrica que possui, e a organização por entidades de negócio (como o aluno, o produto, etc.), facilitam a compreensão por parte dos utilizadores sobre como conseguir extrair a informação que pretendem;
- A melhoria de performance – enquanto os sistemas relacionais, pela sua elevada normalização, estão normalmente optimizados para as tarefas de inserção, actualização e eliminação de dados, o modelo dimensional é também caracterizado por ser altamente desnormalizado (Caldeira, 2008), baseando-se no esquema em estrela que pressupõe a utilização de tabelas de dimensão ligadas directamente a uma tabela de factos central, o que alivia a necessidade de utilização de uniões ou "joins" entre tabelas para obter a informação. Outro aspecto importante é a utilização de informação agregada pré-calculada, o que permite também melhorar bastante o tempo de resposta do sistema (Cortes, 2005);
- A capacidade de actualização – como as suas tabelas possuem uma estrutura muito similar, este modelo permite adicionar novos campos ou tabelas sem implicar alterações profundas ao modelo e que se podem efectuar com uma quantidade de esforço reduzida.

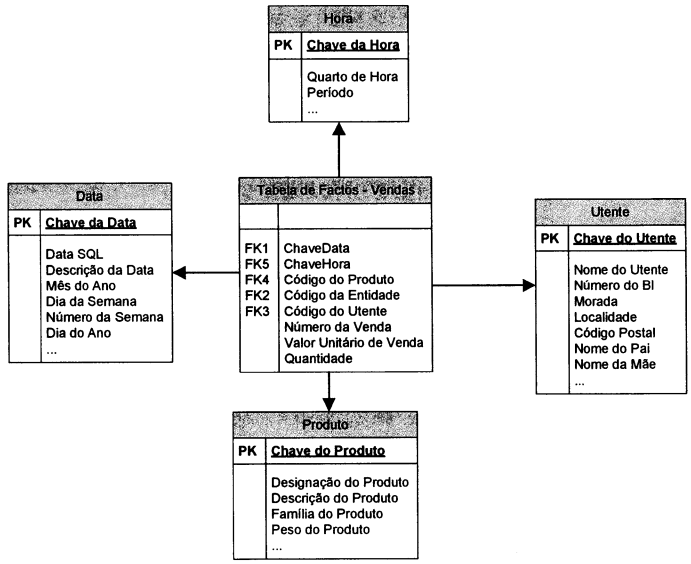


Fig. 8 – Exemplo de um esquema em estrela

O desenho do modelo dimensional assenta num processo que se pode decompor em 4 fases distintas:

- Encontrar os processos de negócio – em primeiro lugar devemos estudar a organização em termos da forma como a mesma realiza o seu negócio, de forma a conseguir identificar qual o processo ou processos de negócio mais importantes. Para nos auxiliar nesta etapa existem ferramentas como a matriz BUS, que iremos explorar em detalhe quando abordarmos o protótipo desenvolvido;
- Definir o grão ou atomicidade da informação – depois de identificar o processo a modelar, devemos passar à identificação da granularidade ou nível de detalhe que os registos devem possuir (ex. registo da nota obtida pelo aluno a uma unidade curricular para um tipo de inscrição). De preferência, devemos utilizar o nível de detalhe mais atómico possível, pois só assim garantimos que o data warehouse terá a sua capacidade de resposta maximizada, ou seja, se possuírmos a informação com o máximo de detalhe podemos não só responder ao máximo de consultas feitas pelos analistas, mas também agregar a informação da forma que pretendermos;
- Definir as tabelas de dimensão – as dimensões são as tabelas que representam as entidades participantes no processo de negócio e são compostas pelos descritivos (normalmente textuais) do negócio. São normalmente tabelas com muitos atributos e poucos tuplos, quando comparadas com as tabelas de factos. Quando uma tabela dimensão é utilizada em mais do que um processo de negócio, mantendo a sua estrutura física denomina-se por Dimensão Conforme, como veremos no sub-capítulo "Tabelas Dimensão";

- Definir as tabelas de factos – as tabelas de factos são o último elemento a ser definido na estrutura do modelo dimensional mas não são por isso menos importantes. Têm como responsabilidade armazenar os atributos mensuráveis do processo de negócio que contribuam para a análise do que estamos a medir e estabelecer a ligação com as tabelas de dimensão. Num modelo em estrela existe uma tabela de factos e várias tabelas de dimensão, sendo este número de dimensões influenciado consoante a granularidade do processo de negócio.

Temos de ter em atenção que todos os factos que identificarmos têm que ser coerentes com o grão utilizado, caso contrário deverão fazer parte de outra tabela de factos. A Fig. 8 apresenta um exemplo de um esquema em estrela.

PROCESSO DE NEGÓCIO

Dada a importância que assume na arquitectura em BUS dimensional, o processo de negócio, já referido anteriormente algumas vezes, deve merecer uma atenção especial nesta dissertação.

Quando pensamos na estrutura de uma organização, normalmente pensamos numa estrutura hierárquica em que no topo temos a área de administração da empresa e depois as várias áreas funcionais que se subdividem em ramos, como o departamento de recursos humanos ou de marketing e publicidade, por exemplo. A organização desta forma permite tirar partido de várias vantagens, como o facto de assim ter as pessoas de determinada área de conhecimento todas agrupadas (todas as pessoas da área de matemática estarão no departamento de matemática, por exemplo), o que permite sinergias em termos da partilha do conhecimento. Outra vantagem é a possibilidade de, como o conhecimento se encontra especializado, quando é necessário noutra área da organização um especialista de uma área diferente, há a possibilidade de utilizar recursos humanos internos (por exemplo, se o departamento de química necessitar de um especialista em estatística pode recorrer ao departamento de matemática). A Fig. 9 apresenta um exemplo de uma estrutura hierárquica.

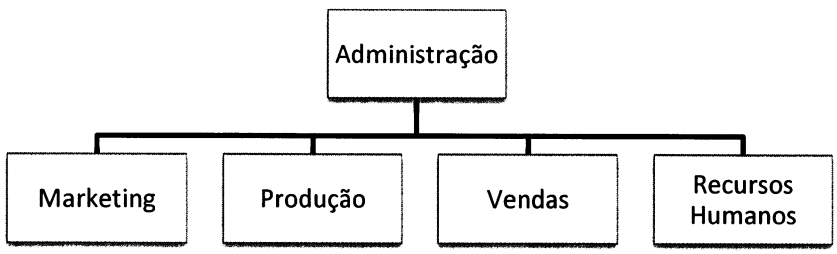


Fig. 9 – Estrutura hierárquica

No entanto, a estrutura hierárquica também possui alguns aspectos menos positivos. É frequente as áreas funcionais armazenarem informação existente e mantida por outras áreas funcionais, criando inconsistências na informação, ou

existirem falhas de comunicação entre áreas. Existe também normalmente a tendência de cada área funcional se fechar sobre si própria, concentrando-se no cumprimento dos seus objectivos em deterimento dos da organização. Mas é quando nos debruçamos sobre o funcionamento da organização em termos da sua actividade e dos produtos e serviços que desenvolve, que mais facilmente percebemos a necessidade de cada área funcional comunicar entre si e de colocar o fruto do seu trabalho como elemento central da sua acção. Pegando num exemplo, se observarmos o processo de venda de uma propina a um aluno, verificamos que esta acção é transversal a várias áreas funcionais da organização, como por exemplo a contabilidade, as vendas e o sector académico, necessitando de recursos de várias áreas. Se nos concentrarmos naquilo que é o objecto do nosso negócio verificamos que a informação não pode ter uma organização exclusivamente hierárquica, mas sim orientada ao conjunto de actividades realizadas pela organização, transversais às várias áreas funcionais, envolvendo diferentes entidades das mesmas.

É neste ponto que entra o processo de negócio. De acordo com Pearlson e Saunders (2009), podemos definir o processo de negócio como o "conjunto de acções ou tarefas sequênciais e inter-relacionadas que transformam os *inputs* em *outputs*" e que inclui o seguinte:

- Um início e um fim;
- Entradas e saídas (inputs e outputs);
- Um conjunto de tarefas que transformam os inputs em outputs;
- Um conjunto de métricas para medir a eficácia.

A construção de um Data Warehouse segundo o modelo dimensional assenta sobre este princípio, devemos identificar os processos de negócio da organização de forma a que nos possamos debruçar sobre cada um e analisar a importância que tem para a organização. Em seguida, construímos um Data Mart que permita conter toda a informação relevante para cada processo de negócio que desejamos estudar. Se a informação sobre uma actividade da empresa estiver localizada num único local e organizada de forma correcta, será possivel estudar a forma de maximizar os indicadores ou métricas que influenciam essa actividade.

Exemplo de um processo de negócio, "A requisição de produtos de economato pelos membros da ESTBarreiro/IPS". Em primeiro lugar, o docente ou funcionário não docente efectua a requisição do material, em seguida o material é entregue e é efectuada a afectação ao membro respectivo.

Fig. 10 – Processo de requisição de material

TABELAS DE FACTOS

As tabelas de factos, como atrás referido, são as tabelas que contêm as métricas que permitem analisar a performance do negócio. São tabelas compostas por um número muito elevado de tuplos ou registos (na ordem dos milhares ou mesmo milhões) e por um número de atributos ou colunas reduzido, ou seja, possuem uma cardinalidade elevada (Pereira, 1998).

Tabela de Factos - Vendas	
PK,FK1	**Chave da Data**
PK,FK5	**Chave da Hora**
PK,FK4	**Código do Produto**
PK,FK2	**Código da Entidade**
PK,FK3	**Código do Utente**
	Número da Venda (DD)
	Valor Unitário de Venda
	Quantidade

Fig. 11 – Exemplo de uma Tabela de Factos

Cada tuplo numa tabela de factos corresponde a uma medida, devendo possuir o mesmo nível de detalhe de todos os outros tuplos dessa mesma tabela. Ao nível de detalhe da tabela de factos damos o nome de grão ou granularidade da tabela de factos. O grão da tabela de factos é definido pelas tabelas dimensão que nela participam, pois cada facto é obtido da intersecção de todas as dimensões. O grão é um elemento bastante importante pois permite-nos definir que tipo de análises e agregações podemos fazer ao dados, ou seja, quanto mais pequeno fôr o grão da tabela de factos maiores as potencialidades de agregação que a mesma poderá ter, considerando que os factos deverão ser, sempre que possível, numéricos e aditivos. O tipo de factos presentes na tabela de factos permite definir que somas e contagens podemos efectuar – a aditividade da tabela de factos.

TIPOS DE FACTOS

A aditividade da tabela de factos é um aspecto bastante importante e central no modelo dimensional, visto que, quando efectuamos análises de informação não estamos interessados no tratamento registo a registo, mas sim na manipulação de de grandes quantidades de registos, analisados por vários prismas de consulta. A maneira mais fácil de reduzir o tamanho da tabela de factos, tornando a consulta de informação mais célere e eficaz é efectuar agregações das linhas da tabela de factos, realizando somas ou contagens sempre que possível. A agregação da informação permite facilitar a consulta da mesma e a obtenção de resultados por vários aspectos de observação, como por exemplo por período de tempo. Permite também, ao reduzir o nº de registos, aumentar a velocidade de exploração do Data Mart e reduzir o espaço ocupado na base de dados.

Outro aspecto importante em termos de performance é que, ao contrário das tabelas de dimensão, as tabelas de factos são altamente normalizadas, existindo muito pouca redundância de informação, o que as torma bastante eficazes.

No que respeita à aditividade, existem essencialmente três tipos de factos:

- Factos aditivos – são os factos que se podem somar para todas as dimensões que participam na tabela de factos, como por exemplo o valor monetário da venda de um determinado produto que, como podemos ver na Fig. 11, qualquer que seja a partição que façamos das dimensões, o facto continua sempre a permitir a sua soma;
- Factos semi-aditivos – são factos que apenas são aditivos face a algumas das dimensões que participam na tabela de factos e não à sua totalidade. Por exemplo, a nota que o aluno obteve a uma Unidade Curricular pode ser considerada um facto semi-aditivo se considerarmos a dimensão Tipo de Exame (normal, melhoria, etc.).
- Factos não aditivos – são medidas normalmente sem caraterísticas aritméticas e que, por exemplo, não se podem somar de todo. O ano curricular em que o aluno está inscrito é um caso de facto não aditivo.

Tabela 4 – Exemplos de aditividade em factos

Facto	Tipo	Descrição
Quantidade vendida do produto x	Aditivo	Nº de elementos vendidos do produto x numa transacção
Número de inscrição a uma unidade curricular	Semi-aditivo	O número de inscrição depende do tipo de inscrição, logo só é aditivo na dimensão Tipo de Inscrição, caso contrário é não aditivo
Preço unitário do produto x	Não aditivo	Preço total em euros do produto x

Voltando ao tópico do grão da tabela de factos, devemos sempre desenhar o Data Warehouse utilizando os Data Marts com o grão mais pequeno possível, também designado por nível atómico. Como normalmente a tabela de factos não se limita a um registo mas sim milhares, o nível atómico não é o mais utilizado para efectuar análises, no entanto permite sempre sumarizar os factos criando agregações ao nível de detalhe que nos interessar. Alguns exemplos de grão de uma tabela de factos podem ser:

Registo transaccional de venda de um produto na ESTBarreiro/IPS

Vendas de propinas realizadas por ano lectivo e por Aluno

Resultados semestrais obtidos por Unidade Curricular

Chaves na tabela de Factos

Como podemos verificar pela Fig. 11, as tabelas de factos possuem, para além das medidas do negócio, um conjunto de atributos chave estrangeira correspondentes às chaves primárias de cada uma das tabelas dimensão que participam no esquema em estrela. A chave primária da tabela de factos consiste num subconjunto ou na totalidade destas chaves estrangeiras. A utilização de chaves artificiais na tabela de factos (ou fact surrogate keys) não é uma prática muito comum, pois para além de tornar a tabela maior, não tem qualquer significado para o negócio.

No entanto, existem algumas características nas chaves artificiais que as revestem de interesse. A sua utilização é sugerida em algumas situações (Kimball, 2009):

- As chaves artificiais permitem identificar um registo da tabela de factos de forma unívoca e imediata;
- Como a sua atribuição é sequencial, uma tarefa de inserção de registos fará com que os mesmos fiquem organizados de forma sequencial;
- A chave artificial permite utilizar operações de inserção e eliminação (insert e delete) em vez de actualizações (updates);
- A chave artificial pode ainda tornar-se numa chave estrangeira numa tabela de factos com um nível de granularidade mais baixa.

Tipos de tabelas de Factos

As tabelas de factos podem classificadas em três tipos distintos (Kimball & Ross, 2002):

- Tabelas transaccionais;
- Tabelas de sumarização periódica;
- Tabelas de sumarização acumulada.

As tabelas de factos transaccionais, como o próprio nome indica, têm um nível de detalhe correspondente a um registo por cada transacção efectuada no sistema transaccional. É útil, por exemplo para medir a frequência com que ocorre determinada actividade na organização ou o período em que ocorre. É o único tipo de tabela que permite responder a questões como:

"Quantas inscrições de alunos ocorrem por dia?"
"As inscrições em exame mensal ocorrem em que dias do período de inscrições?"
"Os pagamentos de propina ocorrem mais no período diurno ou nocturno?"

Características:

- Armazenam um registo por cada transacção efectuada no sistema operacional;
- A dimensão data (e hora, se existir) possui um nível de detalhe mais elevado;
- A granularidade possui o nível mais reduzido;

- O tamanho da tabela de factos é o mais elevado dos três tipos.

Por sua vez, as tabelas de sumarização periódica são tabelas em que os registos são sumarizados de acordo com um determinado período importante para a organização, como por exemplo "*os resultados de avaliação obtidos pelos alunos semestralmente às unidades curriculares*". Permitem um acesso mais rápido aos dados e facilitam a visualização dos resultados por período. Uma das funcionalidades para que é normalmente utilizada é a análise de tendência de evolução ao longo de um período.

Características:

- Armazenam um registo para cada período de tempo definido com os dados agregados;
- A dimensão data (e hora, se existir) passa o seu nível de detalhe para o período de tempo definido;
- É uma tabela de tamanho mais reduzido.

Por último, as tabelas de sumarização acumulada destinam-se a registar o ciclo de vida de uma actividade ao longo das várias etapas que essa mesma actividade decorre. Por exemplo, se quisermos controlar o processo de obtenção de uma nota por aluno a uma unidade curricular podíamos ter uma tabela de factos como a apresentada na Fig. 12, em que cada registo da tabela de factos regista as várias etapas que o processo envolve.

Factos Inscrição - Sumarização Acumulada	
PK,FK5	**Chave da Data de inscrição**
PK,FK4	**Chave da Data de avaliação**
PK,FK3	**Chave da Data de lançamento da nota**
PK,FK7	**Chave da Data de revisão de prova**
PK,FK1	**Chave da Data de validação da nota**
PK,FK2	**Chave do Aluno**
PK,FK6	**Chave da Unidade Curricular**
PK,FK8	**Chave da Data da última actualização**
	Nota Obtida Número de Inscrições

Fig. 12 – Tabela de Factos de sumarização acumulada

Para além das chaves das várias dimensões temporais que compõem a duração da actividade, existe normalmente mais uma dimensão temporal para registar a data da última alteração. Como nem todas as chaves de data são conhecidas quando o processo começa, são utilizadas chaves fictícias ou *dummy* para representar os valores desconhecidos.

Características:

- Apresentam em cada registo todas as etapas de duração de uma determinada actividade;
- Não é tão comum como os outros dois tipos de tabelas de factos;
- O tempo de preenchimento de um registo é indeterminado;
- Os factos e os valores das chaves estrangeiras só podem ser alterados até a actividade terminar;
- Possuem vários atributos com indicadores temporais, um para cada etapa da actividade.

TABELAS DE FACTOS SEM FACTOS

Como vimos ao longo deste sub-capítulo, a estrutura das tabelas de factos é normalmente composta por duas áreas: as colunas com as medidas de negócio e as chaves estrangeiras que ligam a tabela de factos às dimensões que nela participam. No entanto, podem existir situações em que tenhamos tabelas de factos desprovidas de qualquer facto mensurável – designadas por tabelas de facto sem factos (ou como também são conhecidas em inglês, *factless fact table*). Estas tabelas de factos, que são compostas apenas pelas chaves primárias das tabelas dimensão, possuem o objectivo de registar as ocorrências de determinado evento na organização, como por exemplo, o registo de inscrição de um aluno num ano lectivo, como podemos observar na Fig. 13. Como podemos verificar, embora fosse possível realizar as contagens de ocorrências utilizando qualquer atributo que compõe a chave primária da tabela de factos, a inclusão de um atributo adicional com o valor 1 permite tornar mais legíveis as consultas que se efectuem sobre a tabela (Kimball & Ross, 2002).

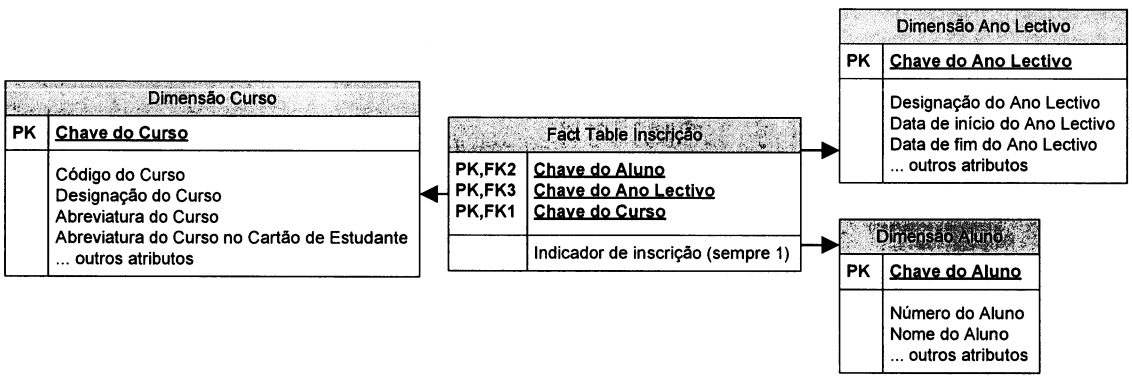


Fig. 13 – Tabela de factos sem factos (para registo de ocorrências)

Outro aspecto importante da utilização de tabelas de factos sem factos é a possibilidade de registarem "não ocorrências", ou seja, permitir o estudo de determinada situação, se esta ocorrer mas também se não ocorrer – são as chamadas *coverage factless fact table* (ou "tabelas de factos de cobertura total"). Por exemplo, se

quiséssemos analisar a não inscrição dos alunos num ano lectivo, podíamos considerar o atributo "indicador de inscrição" da tabela de factos da Fig. 13 com um domínio {1, 0} para indicar, respectivamente, a inscrição ou não do aluno no ano lectivo. Esta solução é possível para o caso apresentado, partindo do pressuposto que o número de alunos que não realizem inscrição não represente uma percentagem muito grande da totalidade de alunos existentes e que, portanto, tal não se traduza num aumento do tamanho da tabela de factos que possa comprometer a sua performance. Se, por exemplo, estivéssemos perante um Data Mart "registo de venda de produtos por transacção" e o objectivo fosse registar os produtos que não obtiveram qualquer venda no final de cada dia, como o número de registos seria potencialmente bastante elevado, a solução passaria por uma *coverage factless fact table* exclusiva para tratar essas ocorrências (Kimball & Ross, 2002).

TABELAS DIMENSÃO

As tabelas dimensão são responsáveis por conter os descritivos textuais que caracterizam as medidas existentes na tabela de factos. As dimensões são caracterizadas por ter, normalmente, um número de colunas bastante elevado e um número de tuplos relativamente reduzido face às tabelas de factos. Designam-se por dimensões conformes as tabelas de dimensão que participem em mais do que um processo de negócio, partilhando atributos com o mesmo nome e conteúdo. Este conceito de dimensão conforme está na base da arquitectura BUS, já apresentada anteriormente.

A ligação entre as tabelas de dimensão e a tabela de factos é efectuada mediante a utilização de chaves sem significado para o negócio, ou *surrogate keys,* que constituem a chave primária da tabela dimensão. Estas chaves são valores únicos, normalmente inteiros, atribuídos a cada linha da tabela dimensão e que permitem as seguintes vantagens:

- Independência face às chaves do sistema transaccional – mesmo que a chave primária no sistema transaccional se altere (mudança de SI, por exemplo), essa alteração não influencia o Data Warehouse;
- Facilidade de integração de fontes de dados heterogéneas – na utilização de diferentes sistemas transaccionais pode acontecer que ambos armazenem informação relacionada com o mesmo assunto, utilizando chaves primárias diferentes;
- Introdução de registos que não existam no sistema transaccional – se por exemplo, o aluno não estiver inscrito numa turma (lapso no sistema transaccional), podemos introduzir um registo *dummy* "sem turma atribuída" utilizando a chave fictícia;

- Alteração de registos – quando necessitamos de alterar um registo, podemos ter de adicionar um novo registo, como iremos ver de seguida. Com a chave fictícia, conseguimos garantir a unicidade da chave;
- Performance – por utilizarmos um número inteiro para a chave primária, a performance é sempre maximizada e o espaço em disco minimizado.

Produto	
PK	**Chave do Produto**
	Designação do Produto Descrição do Produto Família do Produto Peso do Produto Altura do Produto Largura do Produto Profundidade do Produto Marca do Produto ... (outros)

Fig. 14 – Exemplo de uma tabela Dimensão

HIERARQUIAS

As hierarquias são relações que se podem estabelecer entre os atributos de uma tabela dimensão e que permitem agrupar a informação de forma a que possamos efectuar as análises partindo de um atributo pai para um atributo filho, sumarizando a informação de forma automática. Estas relações entre os atributos ocorrem devido à desnormalização das tabelas dimensão, ao conter informação na mesma tabela que não se relacione apenas com a chave primária da mesma, leva a que seja possível criar relações hierárquicas entre os atributos. A consulta destas hierarquias é efectuada mediante a utilização de operações de consulta detalhada – também conhecidas por operações de *Drill Down* – ou resumos da informação generalizada – também designados por operações *Roll Up* (Cortes, 2005).

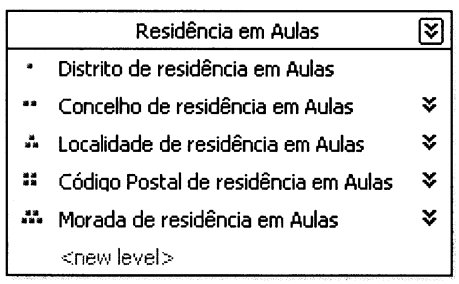


Fig. 15 – Estrutura hierárquica com profundidade fixa

Vejamos o exemplo da Fig. 15: podemos começar por consultar a informação dos alunos por distrito e efectuar uma operação de *drill-down* para os concelhos e por aí adiante até chegar ao nível específico desejado, ou trabalhar no sentido contrário e ver

a informação mais agregada à medida que subimos de nível. A designação "nível" é a indicação que se dá a cada atributo presente na hierarquia.

Existem alguns aspectos que devemos ter em consideração quando criamos hierarquias (Mundy, 2008):

- Podemos ter várias hierarquias na mesma tabela dimensão – ter mais do que uma hierarquia numa dimensão permite enriquecer as pesquisas na mesma. No entanto, devemos trabalhar com os utilizadores do Data Warehouse na definição dos nomes dos atributos e das hierarquias para que o seu significado seja facilmente entendido;
- Deve existir integridade referencial de muitos-para-um entre níveis – uma localidade pertence a um (e apenas a um) concelho, um concelho apenas a um distrito e assim sucessivamente, à medida que se efectua uma operação de *roll-up*;
- Estudar com cuidado a fase de desenho – a reutilização de atributos em mais do que uma hierarquia deve ser analisada com cuidado, em algumas situações podemos ter interesse em efectuar intersecções entre as hierarquias, no entanto temos que garantir que os dados suportam essa intersecção. Um exemplo desta situação é ilustrado pela Fig. 16, com as hierarquias *Semana* e *Calendário*, da dimensão Data. Numa primeira análise o atributo "ano" poderia ser partilhado por ambas as hierarquias. No entanto, se a organização definir como início da semana a segunda-feira, então a semana número 1 pode começar depois do início do ano, terminando nesse caso a última semana do ano já no ano seguinte.

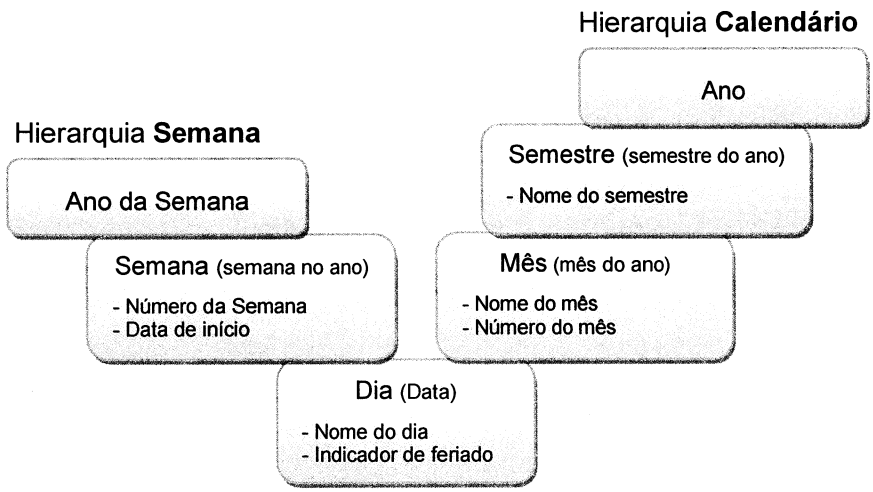


Fig. 16 – Exemplo de hierarquias para a dimensão Data

A fonte de dados utilizada para carregar as tabelas dimensão e respectivas hierarquias representa um papel bastante importante no desempenho destas ao longo do tempo. Se utilizarmos uma fonte de dados com a informação correcta e bem

estruturada (como por exemplo proveniente de um sistema relacional normalizado) a manutenção do correcto funcionamento da hierarquia torna-se bastante mais fácil. Se, por outro lado, for utilizada uma fonte não estruturada, com informação imperfeita (em que o caso da folha de cálculo é um bom exemplo) então a hierarquia não terá o comportamento desejado, ou terá pelo menos uma manutenção bastante difícil (Mundy, 2008).

Sempre que possível, o ideal é efectuar a manutenção da estrutura hierárquica directamente no sistema transaccional quando este assenta numa base de dados relacional normalizada, pois normalmente as estruturas hierárquicas são facilmente mantidas com tabelas normalizadas, como apresenta a Fig. 17.

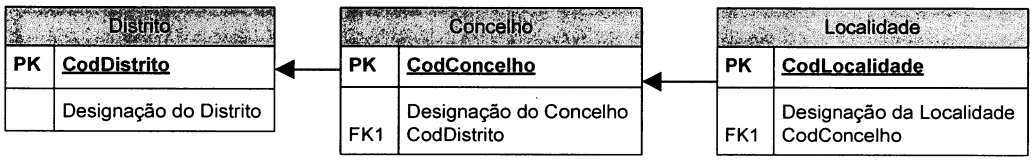


Fig. 17 – Hierarquia no modelo relacional (normalizada)

Como vimos, as hierarquias de formato fixo são relativamente fáceis de desenvolver, pois este desenvolvimento corresponde a um conjunto de atributos de uma dimensão. Não deve, no entanto, utilizar-se um número de níveis muito grande, o que poderia confundir o utilizador – no máximo não devemos exceder os 5 níveis (Caldeira, 2008). Um exemplo duma hierarquia de formato fixo é a hierarquia "Residência em Aulas" que vimos na Fig. 15, em que podemos apresentar a informação sumariada por Distrito, depois por Concelho e por aí adiante, utilizado a operação de *drill-down*.

Para além das hierarquias com um número fixo de níveis, existem também hierarquias de formato variável. Estas hierarquias não são tão fáceis de executar (Caldeira, 2008), pois exigem um processo recursivo para o seu desenvolvimento. Um exemplo deste problema são as unidades curriculares com precedências, em que podemos ter vários níveis, como nos mostra a Fig. 18.

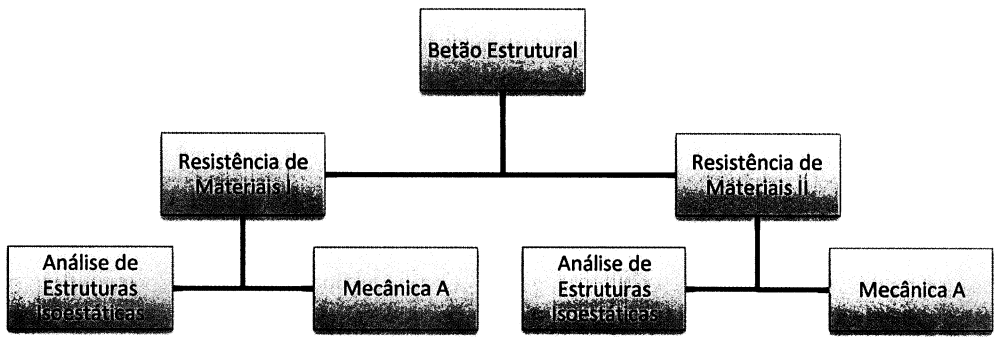


Fig. 18 – Estrutura hierárquica com profundidade variável

Como o SQL não é muito eficaz a percorrer estruturas recursivas, a solução para estes casos passa pela utilização de uma tabela ponte, que iremos abordar no capítupo destinado ao modelo desenvolvido.

DIMENSÕES DE ALTERAÇÃO LENTA

Embora teoricamente a informação contida na tabela de dimensão não devesse sofrer alterações ao longo do tempo, a verdade é que por vezes é necessário efectuar alterações aos registos destas tabelas. Podem acontecer alterações devido às seguintes situações:

- Existe um erro na informação – neste caso o valor no sistema transaccional é corrigido e as análises efectuadas a partir desse momento deverão utilizar o valor corrigido;
- Existe uma actualização devido a um factor externo – se por exemplo um funcionário da instituição resolver casar, terá de alterar o seu estado civil no sistema e, em alguns casos, mudar o apelido do nome;
- Existe uma alteração funcional na organização – se houver uma alteração no plano de estudos de uma unidade curricular, por exemplo ao nível do seu acrónimo, essa alteração tem de ser reflectida na dimensão Unidade Curricular.

Embora existentes, não são alterações que ocorram com grande frequência, daí a sua designação de dimensões de alteração lenta ou *slowly changing dimensions*, em inglês (SCD, como também são conhecidas).

As SCD podem ser classificadas em três tipos diferentes, que iremos abordar em seguida. Estes tipos de alteração podem diferir de atributo para atributo dentro da mesma tabela dimensão, consoante a importância do mesmo.

Alterações Tipo 1 – as alterações do tipo 1 são as mais fáceis de implementar pois consistem em simplesmente alterar o valor antigo pelo novo. Têm a desvantagem de eliminar a existência de histórico, pois perdemos o valor antigo. Normalmente utilizam-se quando a alteração consiste em alterar um valor anteriormente omisso ou quando a alteração não tem impacto ou especial relevo para a organização. Esta importância deverá ser discutida com os utilizadores do Data Warehouse. Por exemplo, no caso de um erro no valor do atributo "E-Mail #1" da dimensão Aluno, como não se trata de um atributo com relevância e necessidade de histórico, podemos utilizar uma alteração do tipo 1. A Tabela 5 mostra uma alteração do tipo I, antes e após a alteração.

Tabela 5 – Exemplo da Dimensão Funcionário com uma alteração Tipo 1

Chave do Funcionário	Nome do Funcionário	E-Mail	Outros atributos	Data Início	Data Fim
97	Manuel Silva Franco	me@myemail.com	...	01-01-1999	
...	...	...	...	...	...

Antes

Chave do Funcionário	Nome do Funcionário	E-Mail	Outros atributos	Data Início	Data Fim
97	Manuel Silva Franco	msf@nova_vida.com	...	01-01-1999	
...	...	...	...	...	...

Depois

Alterações Tipo 2 – as alterações do tipo 2 consistem em adicionar um novo registo à tabela com o novo valor no atributo respectivo e invalidar o registo antigo. Este processo de invalidação do registo antigo é efectuado à custa de um atributo temporal que permite controlar a validade dos registos na tabela dimensão. Com este tipo de alteração mantém-se um registo perfeito do histórico da informação. Um exemplo de atributo que pode utilizar este tipo de alteração é o número de dependentes do funcionário, atributo que pode ter implicações no cálculo dos descontos no ordenado e que portanto necessita de manutenção de informação histórica.

Tabela 6 – Exemplo da Dimensão Funcionário com uma alteração Tipo 2

Chave do Funcionário	Nome do Funcionário	Número de Dependentes	Outros atributos	Data Início	Data Fim
97	Manuel Silva Franco	0	...	01-01-1999	20-11-2009
98	Joana Rito e Sousa	2	...	01-01-1999	
...	...	...	...	...	...
201	Manuel Silva Franco	1	...	21-11-2009	

Alterações Tipo 3 – as alterações do tipo 3 não são muito frequentes pois implicam alterações na estrutura da própria tabela dimensão. Ocorrem quando se pretende analisar cenários alternativos, simulando as duas situações: com o valor alterado e com o valor original. A alteração consiste em adicionar uma nova coluna para armazenar o novo valor sem alterar o antigo. Utilizando o exemplo da Tabela 6, se fosse utilizada uma alteração do tipo 2, quando fizessemos uma pesquisa pelas Unidades Curriculares do tipo "Matérias Complementares", não iríamos ver os registos da Unidade Curricular Química I anteriores à alteração. Se o objectivo fôr a análise comparativa estudando a situação como se a alteração não tivesse ocorrido, então teríamos de utilizar uma alteração do tipo 3, como apresenta a Tabela 7, que representa a situação em que uma Unidade Curricular muda de classificação (no caso de uma reestruturação curricular, por exemplo).

Tabela 7 – Exemplo da Dimensão Unidade Curricular com uma alteração Tipo 3

Chave da Unidade Curricular	Designação da Unidade Curricular	Outros Atributos	Tipo da Unidade Curricular	Novo Tipo da Unidade Curricular
123	Química I	...	Matérias Complementares	Ciências de Base
...	...	...	...	...

MINIDIMENSÕES

Quando uma dimensão atinge um número de atributos elevado, as alterações do tipo 2 tornam-se bastante difíceis de implementar (Kimball & Ross, 2002), pois para além destas dimensões possuirem muitos atributos, as suas alterações são mais frequentes, dada a importância que representam para o negócio. Isto leva ao aumento desnecessário do tamanho da base de dados pois a redundância com a alteração do tipo 2 cria um novo registo quase idêntico, alterando apenas o valor de um dos atributos do registo.

O objectivo das minidimensões passa por tentar dividir a dimensão retirando-lhe os conjuntos de atributos relacionados que permitem agrupar ou classificar os restantes atributos. O exemplo de uma minidimensão é o apresentado na Fig. 19.

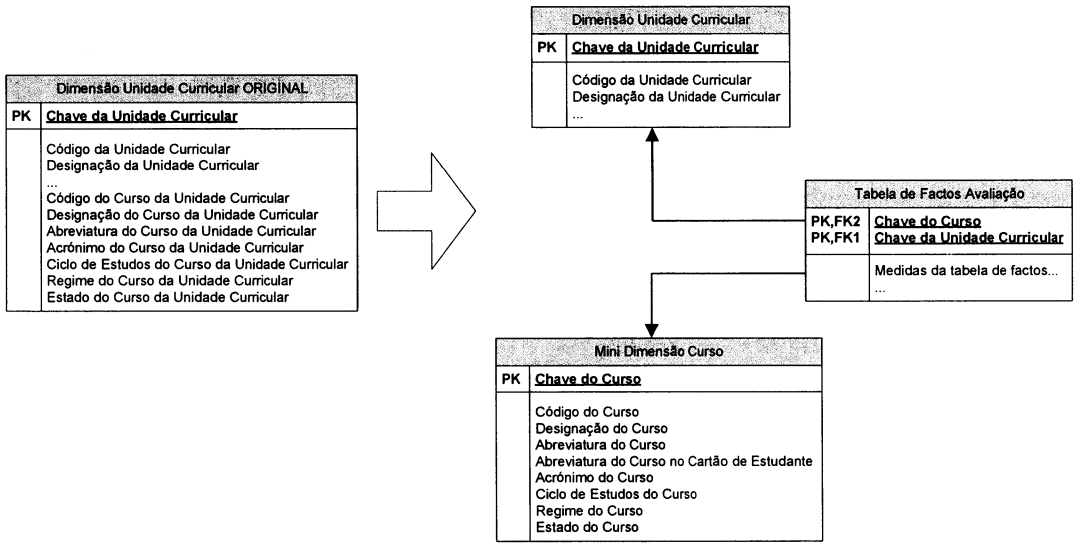


Fig. 19 – Exemplo de uma Minidimensão

A DIMENSÃO DATA

A construção de qualquer Data Mart terá sempre na dimensão Data um objecto fundamental, indispensável para dotar de capacidade de análise temporal todas as medidas desse mesmo Data Mart. É normalmente das primeiras dimensões a ser implementada e, ao contrário da maioria das dimensões, o seu preenchimento é efectuado sem recorrer a informação presente nos sistemas operacionais da organização. Este preenchimento pode possuir uma frequência de actualização diferente do restante sistema, na medida em que ocorre normalmente na primeira vez em que o sistema é carregado e inclui a introdução de informação por antecipação, ou seja, podemos carregar a dimensão não só com registos de suporte à informação já ocorrida – registos históricos, como também com informação para as operações que

ainda se irão realizar (podemos carregar os registos de data para os próximos 5 a 10 anos, por exemplo).

Embora o grão da dimensão Data seja o dia, pode existir a necessidade de existir um nível de grão mais fino, como por exemplo a hora ou mesmo o minuto ou o segundo. Nestes casos deve ser adicionada ao modelo uma nova dimensão, que se poderá designar por exemplo, dimensão Hora. No caso do modelo apresentado, no processo de venda é utilizado a dimensão hora com uma granularidade de 15 minutos ou, de forma mais corrente, um quarto de hora.

O conjunto de atributos com as características da data deve ser pensado de forma cuidadosa, de forma a garantir que o utilizador consegue analisar a informação de todos os ângulos que necessita. É comum fazerem parte desta dimensão atributos genéricos como "Descrição completa da data", "Indicador de Feriado" ou "Dia do ano", bem como atributos específicos do negócio que estamos a modelar, como "Ano Lectivo" ou "Momento académico". A Tabela 8 apresenta alguns exemplos destes atributos que permitem a sua contextualização.

Tabela 8 – Exemplo de registos para a Dimensão Data

Chave da data	Descrição completa da Data	Dia do Ano	Indicador de Feriado	...	Ano Lectivo	Momento Académico
120091	1 de Janeiro de 2009	1	Feriado		2009/2010	Férias de Natal
220091	2 de Janeiro de 2009	2	Não Feriado		2009/2010	Férias de Natal
...	...	...	...	...	...	...
119001	Data não disponível					
31999912	Ainda não aconteceu					

A chave primária, à semelhança das chaves das outras dimensões, corresponde à utilização de uma chave artificial. No entanto, como podemos verificar na Tabela 8, no caso da dimensão Data esta não necessita de ser um número sequencial numérico sem significado, pois não é de esperar a necessidade de alterações do Tipo 2 ou a integração da informação na dimensão Data proveniente de vários sistemas operacionais (Kimball, Mundy, & Thornthwaite, 2006). Assim, podemos utilizar uma mnemónica para a chave primária, composta pelo dia, o ano e o mês da data (DDAAAAMM), que facilita a identificação dos registos da tabela de factos utilizando a chave da Data, o que pode auxiliar em algumas tarefas de administração do sistema.

Ainda na Tabela 8, podemos verificar que os dois últimos registos possuem características especiais, funcionando, respectivamente, para indicar uma data em falta no sistema, ou que determinada data ainda não ocorreu.

Esta dimensão, pelo conjunto de diferentes características que possui e pela importância que revela no Data Warehouse – as análises ao longo duma linha temporal são as mais frequentes (Caldeira, 2008) – é uma forte candidata à utilização de hierarquias, como por exemplo as que vimos na Fig. 16.

DIMENSÕES "ROLE-PLAYING"

As dimensões *role playing* são as tabelas dimensão que podem desempenhar vários papéis dentro do Data Warehouse, consoante o contexto em que sejam utilizadas. A dimensão Data é um bom exemplo de uma dimensão *role playing*: é frequente existirem datas com significados diferentes, como por exemplo a data de inscrição do aluno numa Unidade Curricular e a data de avaliação a essa mesma Unidade Curricular. A Fig. 20 apresenta um exemplo da utilização role-playing para a dimensão Data.

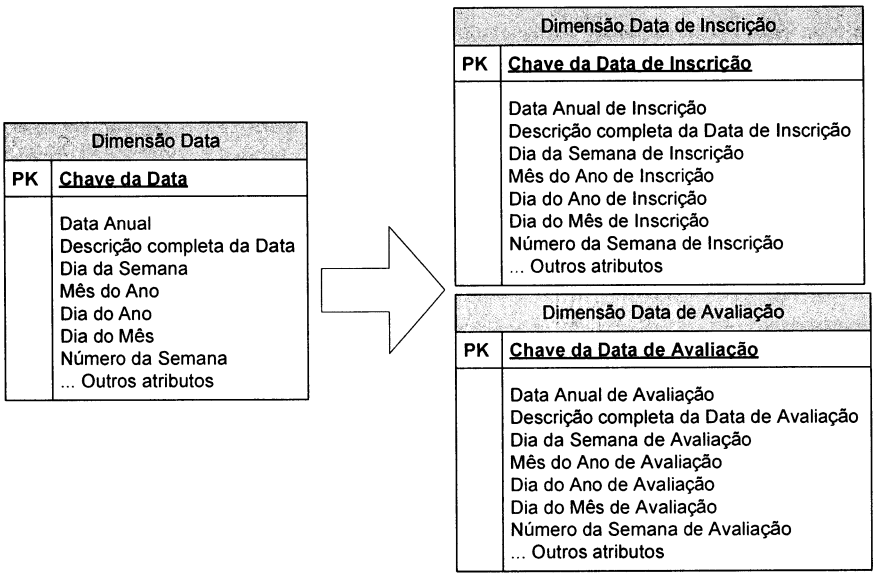


Fig. 20 – Exemplo de dimensão Role Playing

Outro exemplo é a dimensão Professor. No processo de percurso académico do aluno, a dimensão Professor pode assumir papéis distintos: o de professor que lecciona a Unidade Curricular ou o de professor responsável pela mesma. No entanto, a dimensão professor pode ser utilizada para as duas situações sem qualquer alteração. As dimensões *role playing* são construídas com a utilização de vistas, em que criamos uma nova dimensão, alterando apenas as designações dos vários atributos para o novo papel, tal como podemos verificar no exemplo que se apresenta em seguida.

```
CREATE VIEW DimV_Responsável
AS
SELECT
[Chave do Professor] as [Chave do Responsável da UC],
    [Código do Professor] as [Código do Responsável da UC],
    [Nome do Professor] as [Nome do Responsável da UC],
    [Morada do Professor] as [Morada do Responsável da UC],
    [Localidade do Professor] as [Localidade do Responsável da UC],
    [Código Postal do Professor] as [Código Postal do Responsável da UC],
    [Número de BI do Professor] as [Número de BI do Responsável da UC],
    [Estado Civil do Professor] as [Estado Civil do Responsável da UC],
    .... (outros campos)
FROM DimProfessor
```

DIMENSÕES DEGENERADAS

As dimensões degeneradas são dimensões desprovidas de atributos, ou seja, os atributos destas dimensões já estão atribuídos a outras dimensões. Correspondem normalmente às chaves do sistema transaccional, como por exemplo o número do aluno, que é a chave primária no sistema transaccional para a tabela Aluno e corresponde a uma dimensão degenerada no modelo dimensional. São adicionados como atributos na tabela de factos e identificam-se com a colocação da abreviatura "DD" a seguir ao nome do atributo (ver Fig. 11).

Permitem efectuar a verificação rectroactiva dos registos na tabela de factos com os registos no sistema transaccional, pois passa a existir na tabela de factos um atributo que permite efectuar essa ligação.

DIMENSÃO FLOCO DE NEVE (SNOWFLAKING, EM INGLÊS)

A técnica de *snowflaking* contrasta com a utilização do esquema em estrela e consiste na normalização da tabelas dimensão na tentativa de eliminar alguns grupos de atributos repetidos e assim melhorar a manutenção e reduzir a ocupação de espaço. Esta técnica é desaconselhada por vários autores, (Caldeira, 2008; Kimball & Ross, 2002), na medida em que a sua utilização leva a uma degradação na performance do sistema e a um aumento da complicação do modelo – que é precisamente o contrário das características que deve possuir um Data Warehouse. Esta técnica afasta-nos da simplicidade do modelo em estrela e das suas características de performance.

O caso mais extremo de snowflaking irá resultar no modelo normalizado na 3ª forma normal. Na Fig. 21 podemos verificar um exemplo de uma parte de um modelo em *snowflake*, neste caso para o distrito e concelho da dimensão Aluno.

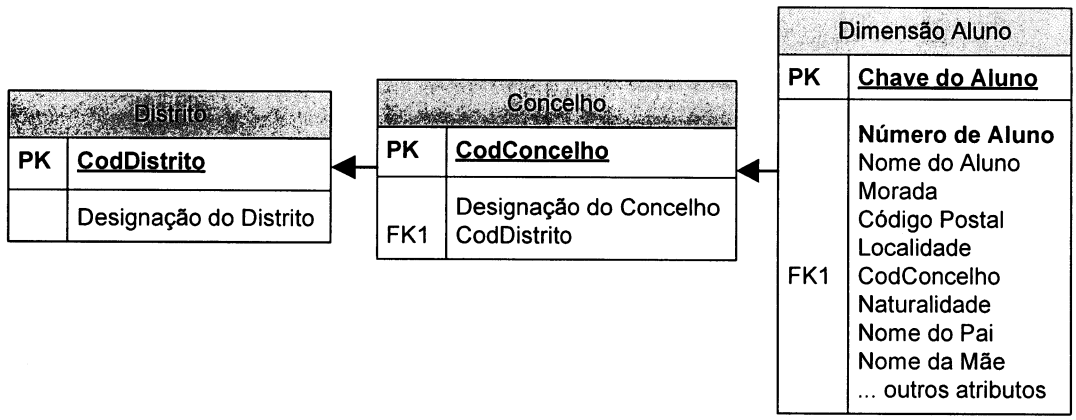


Fig. 21 – Exemplo de dimensão em floco de neve

Existem no entanto algumas situações em que podemos ter a necessidade de utilizar a técnica de snowflake, quando queremos criar agrupamentos que permitam

classificar os valores de uma dimensão, como por exemplo a utilização de uma classificação por região NUTS II[6] dos alunos da dimensão Aluno, como nos mostra a Fig. 22. A estas dimensões dá-se o nome de tabela de suporte ou *Outrigger*.

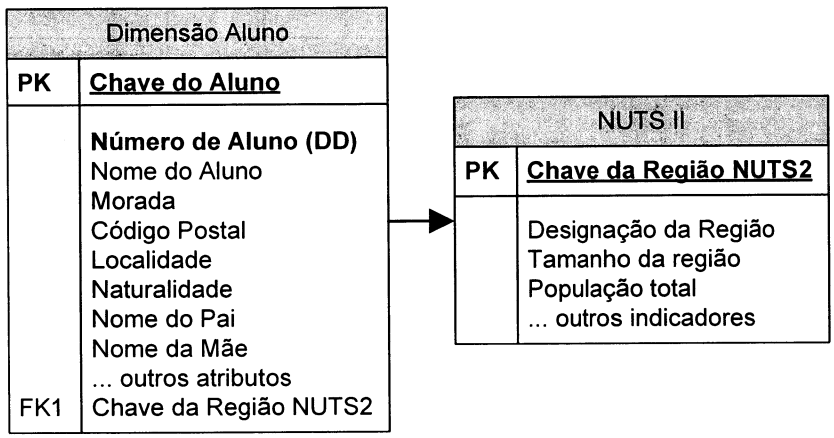


Fig. 22 – Dimensão *outrigger* "região NUTS II" para a Dimensão Aluno

AS DATAS NO MODELO DIMENSIONAL

Como vimos anteriormente, a dimensão Data desempenha um papel muito importante num Data Warehouse. Mas não é apenas nesta dimensão que podemos encontrar datas num Data Warehouse. Existem várias situações em que são utilizadas datas que funcionam como marcas temporais nos dados, como por exemplo no processo de negócio de Percurso Académico, em que temos a data de entrada na instituição ou a data de nascimento do aluno.

Estas datas possuem três formas de ser tratadas (Caldeira, 2008):

- Introdução na tabela de dimensão como atributo – se a função da data fôr apenas informativa, então será mais um atributo da tabela dimensão a que respeita;
- Introdução de atributos adicionais – se as análises efectuadas exigirem agregações por mês, por ano ou de outro tipo, então deverão ser incluídos atributos adicionais para permitirem as análises pretendidas;
- Utilização de uma dimensão *outrigger* – no caso de se tratar de uma análise mais complexa, em que sejam necessárias várias componentes da data, então deverá ser criada uma ligação entre a dimensão que possui o atributo data e a própria dimensão data, funcionando esta última como um *outrigger* da primeira.

[6] NUTS é a abreviatura de *Nomenclature of Territorial Units for Statistics* ou, em português, Nomenclatura Comum das Unidades Territoriais Estatísticas. Representa uma classificação territorial europeia com base em critérios geográficos e populacionais, utilizada para fins estatísticos.

3. O Problema

3.1. Motivação

As Instituições de Ensino Superior geram, na sua actividade do dia-a-dia, grandes quantidades de informação, de origem bastante diversa. Esta diversidade surge, por exemplo, quando identificamos os vários intervenientes na organização, pois existe informação sobre alunos, funcionários (docentes e não docentes), recursos materiais, produtos e serviços transaccionados ou informação sobre os seus fornecedores. Esta informação encontra-se normalmente organizada pelos vários departamentos ou áreas funcionais da organização, como o sector académico ou o departamento de contabilidade. Mas também podemos distinguir a informação quando estamos a tratar um interveniente específico. Vejamos por exemplo o caso dos alunos: podemos ter informação relacionada com o historial académico, com a forma de ingresso na instituição ou com os registos de aquisições/pagamentos efectuados, entre outros.

O armazenamento/gestão dessa informação e a sua disponibilização aos diversos actores da instituição de uma forma diferenciada e concisa, mas ao mesmo tempo completa e robusta, de forma a providenciar a capacidade de análise e consequente tomada de decisões não é o propósito do sistema transaccional que, como já anteriormente foi referido, se destina a registar e gerir as operações do dia-a-dia da organização.

Com os desafios que se colocam às instituições nos tempos que correm torna-se essencial que consigamos tratar as fontes de informação ao nosso dispor no sentido de melhorar a qualidade das decisões tomadas. Um desses desafios é relação da instituição com o mundo empresarial e a comunidade em que se insere, procurando não só sinergias que permitam a divulgação do conhecimento que possui mostrando as suas valias e a qualidade do serviço que oferece, mas também criando parcerias que permitam evoluir técnica e cientificamente os seus diplomados e também os seus recursos humanos – docentes e não docentes – aumentando assim a qualidade do serviço prestado. Outro desafio, relacionado com o anterior, é a procura de fontes de financiamento externo. Uma das formas que temos de atrair esse investimento passa por conseguir medir de forma eficaz a qualidade do nosso serviço, melhorando-o dia após dia.

A diminuição de alunos candidatos ao ensino superior também é um desafio importante, bem como a tentativa constante de atrair os melhores alunos. Esta tarefa, embora muito difícil, tem de ser perseguida e a melhor forma de o fazer é aumentar a qualidade do ensino prestado.

Em termos das limitações com que os utentes da instituição se deparam quando necessitam de utilizar a informação de que dispõem no seu sistema de informação, podemos destacar os seguintes:

- Acesso limitado à informação – o acesso à informação disponível no sistema transaccional só pode ser efectuado por uma de duas formas: ou existe uma interface implementada pelos arquitectos do sistema de informação (a *API – Application Program Interface*), o que torna o utilizador completamente dependente dos técnicos de informática ou então o utilizador possui conhecimentos não só de informática mas também da estrutura de dados específica do sistema de informação, que lhe permitam extrair a informação desejada através de linguagens estruturadas de dados (como seja por exemplo o *SQL*). Mesmo que o acesso aos dados não seja um problema, analisar a informação que se encontra num registo transaccional limita ou dificulta bastante a análise, normalmente efectuada tendo interesse em períodos agregados no tempo;
- Dificuldade de integração – é muitas vezes necessário integrar informação distribuída por fontes de dados independentes, lógica e físicamente. A informação a disponibilizar a um determinado sector da organização pode encontrar-se armazenada em diferentes formatos (bases de dados relacionais, folhas de cálculo, informação externa à organização) ou em diferentes servidores, e ser necessário combinar esta informação de forma a responder às solicitações dos decisores. No caso da instituição em estudo, o Instituto Politécnico de Setúbal, como o objecto desta tese assenta principalmente sobre a realidade de uma das suas Unidades Orgânicas (a Escola Superior de Tecnologia do Barreiro), a opção de desenvolver o SI à medida das necessidades da organização permite limitar até certo ponto este problema, pelo menos ao nível da comunicação entre os SI de área funcional, e à usual heterogeneidade do tratamento da informação entre os diferentes sistemas. De qualquer forma, é um problema continua a existir pois existem fontes externas de informação, como ficheiros de folha de cálculo, por exemplo. A solução apresentada permite no entanto solcionar todos este problemas, quando o sistema evoluir para toda a Instituição;
- Análise de informação histórica e agregada – os sistemas transaccionais registam apenas a informação mais actual, não mantendo o histórico da informação na base de dados, pelo menos de uma forma que permita a consulta directa. Isto leva a que seja extremamente difícil analisar a evolução histórica da informação, como por exemplo, a evolução do preço dos produtos e serviços transaccionados pela escola ao longo de um período. Por outro lado, a análise de informação é uma tarefa que exige a interpretação dos dados sob diferentes níveis de detalhe ou agregação, como seja, por exemplo o estudo da

assiduidade de uma determinada turma: podemos querer estudar a evolução semanal, anual, por unidade curricular, etc.

Torna-se então necessário explorar outras técnicas que, ao mesmo tempo que permitam um acesso simples e directo à informação, com o nível de detalhe necessário, garantam independência da infra-estrutura tecnológica implementada, não exigindo aos seus utilizadores quaisquer conhecimentos tecnológicos adicionais para aceder a essa informação.

A solução estudada é o Data Warehouse, que irá permitir dotar a instituição de uma ferramenta analítica para auxiliar os seus gestores no processo de tomada de decisão e permitir o acesso à informação de uma forma mais rápida. Um outro aspecto importante é a possibilidade desta ferramenta, ao integrar a informação de várias fontes de dados e proceder à sua normalização, permitir disponibilizar uma única versão da realidade.

Na tabela abaixo apresentada podemos verificar o volume de vendas em milhões de dólares, nos anos de 2005 e 2006.

Tabela 9 – Vendas mundiais de plataformas de Data Warehouse por fabricante

IDC (2007)

	Receitas (milhões de dólares)		Quota (%)		Crescimento (%)
	2005	2006	2005	2006	2005-2006
Oracle	1.653,2	1.869,9	32,6	32,8	13,1
IBM	1.112,2	1.245,1	21,9	21,8	12,0
Microsoft	630,1	775,1	12,4	13,6	23,0
SAS	457,3	498,7	9,0	8,7	9,0
NCR Teradata	400,6	418,8	7,9	7,3	4,6
Informatica	222,8	248,0	4,4	4,3	11,3
Sybase	88,0	94,1	1,7	1,6	7,0
Business Objects	51,1	57,9	1,0	1,0	13,2
Fujitsu	34,5	35,3	0,7	0,6	2,5
Netezza	22,8	32,3	0,4	0,6	42,0
Subtotal	4.672,6	5.275,2	92,0	92,3	12,9
Outros	399,2	428,1	8,0	7,7	8,7
Total	5.071,8	5.703,3	100,0	100,0	12,5

3.2. CARACTERIZAÇÃO DO SISTEMA EXISTENTE

A Escola Superior de Tecnologia do Barreiro, do Instituto Politécnico de Setúbal, adiante designada por ESTBarreiro/IPS, possui um Sistema de Informação – o

SIESTBarreiro/IPS – desenvolvido à medida das suas necessidades pelo Núcleo de Informática da escola e que tem como missão proporcionar à instituição um recurso integrado de gestão da informação. Este sistema de informação garante o tratamento informático das tarefas operacionais realizadas pelas várias àreas de funcionamento da escola e registo no sistema transaccional da informação produzida.

A opção de desenvolver internamente o sistema de informação (também normalmente designado por desenvolvimento *in-house*) resultou da especificidade e frequência de alterações e ajustes a que uma solução deste tipo está sujeita e ao consequente benefício em termos de domínio não só da plataforma desenvolvida bem como da tecnologia utilizada para a desenvolver, o que permite um tempo de resposta mais eficaz pelo Núcleo de Informática às solicitações apresentadas.

O SIESTBarreiro/IPS possui uma arquitectura modular baseada nos sistemas Enterprise Resurces Planning (ERP) anteriormente apresentados e compreende as seguintes componentes:

- **Gestão de Serviços Académicos** – módulo desenvolvido para os serviços académicos, permite a gestão da informação académica do aluno e dos currículos disponibilizados pela Escola;
- **Gestão de Produtos/Serviços** – gestão das receitas geradas pela escola, é o módulo utilizado pela contabilidade e por todos os serviços que gerem receitas. Permite, por exemplo, controlar o pagamento de propinas, emolumentos devidos pelos vários serviços prestados (como os certificados e certidões), ou aquisição de produtos da reprografia. A informação produzida por este módulo é posteriormente tratada pela contabilidade (mediante um back-end de acesso ao sistema e a utilização de relatórios) e integrada na aplicação de contabilização analítica existente nos serviços centrais do IPS, baseada no software Primavera;
- **Gestão de Recursos Humanos** – módulo que permite controlar a informação sobre os recursos humanos da Escola, nomeadamente, dados pessoais, currículo, informação contratual, docência e trabalho realizado dentro e fora da instituição. Neste módulo não existe a componente de processamento de salários, visto essa tarefa ser executada a um nível central nos serviços da presidência do IPS, por uma aplicação externa que está neste momento a sofrer algumas aterações, nomeadamente, a mudança de plataforma;
- **Gestão de Serviço Docente** – permite gerir toda a informação relacionada com o serviço dos docentes da escola, como por exemplo: o serviço lectivo ou docência, os sumários, o preenchimento dos programas e cargas horárias, etc.

A implementação técnica do sistema de informação, no que diz respeito às ferramentas de interface com o utilizador, assenta em aplicações distintas: um portal web para funcionar como *front-end* para o utilizador e uma ferramenta *Windows-*

based, que funciona como *back-end* para o funcionamento dos serviços. É com estas duas ferramentas que toda a informação produzida na instituição é recolhida e colocada no sistema de armazenamento. Para além destas ferramentas, existem mais algumas fontes de informação dispersas pela instituição, algumas em bases de dados não ligadas ao sistema de informação e outras em formatos de folha de cálculo. Existem também fontes de informação externa à escola, nomeadamente, proveniente do ministério da educação e que utiliza bases de dados em Microsoft Access. É o caso por exemplo dos resultados de acesso anual à ESTBarreiro/IPS e que, para além destes resultados, fornecem também informação estatística útil para estudar o perfil do candidato à escola.

No que concerne ao modelo de dados desenvolvido para suporte às aplicações, o SIESTBarreiro/IPS está dividido em várias bases de dados transaccionais. Estas bases de dados estão implementadas num servidor central único, acessível apenas pelos membros do Núcleo de Informática e pelos serviços que utilizam as aplicações de acesso ao SI. A Fig. 23 apresenta um diagrama conceptual da arquitectura do sistema, onde podemos observar o servidor principal de bases de dados, apenas acessível pelas ferramentas de back-end e respectivos utilizadores.

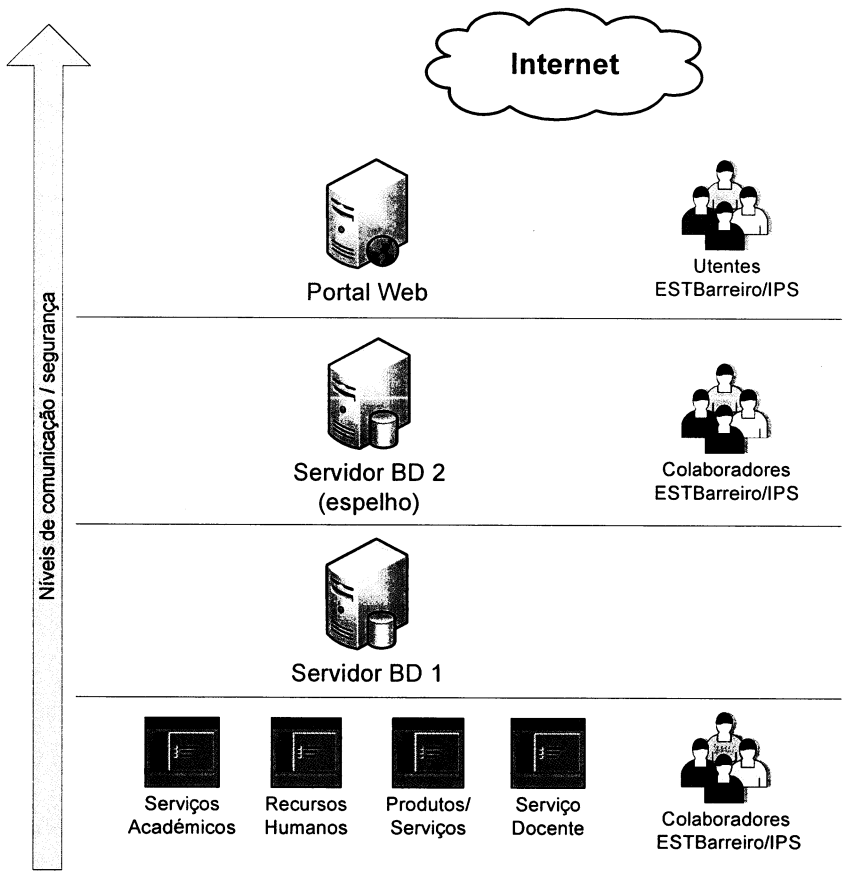


Fig. 23 – Diagrama conceptual da arquitectura do SI

Este servidor replica a informação para o servidor 2 que possibilita o acesso à informação a partir dos portais do estudante e do funcionário e também do portal público. Por questões de segurança, a comunicação apenas é permitida no sentido da seta da esquerda.

No entanto, não deixa de ser possível actualizar a informação existente no servidor DB 1 através de módulos desenvolvidos para funcionamento nos portais, desde que a comunicação se efectue no sentido servidor 1 → servidor 2, ou seja, para que exista comunicação, a mesma deverá ser iniciada pelo servidor 1. Para o efeito, consoante a importância dos dados e a frequência de actualização dos mesmos, existem agentes de software que periodicamente verificam se existe informação nova introduzida nos portais e, após efectuar alguns testes de qualidade, carregam essa informação nas bases de dados do Servidor BD 1.

As bases de dados estão implementadas de acordo com o modelo relacional de Codd (1970), normalizadas até à terceira forma normal (3FN). A comunicação entre as mesmas é efectuada, sempre que necessário, através de triggers e sistemas de replicação para garantir a consistência e integridade dos dados. O motor de base de dados utilizado é o Microsoft SQL Server, na versão 2008, visto a instituição possuir acesso facilitado na aquisição de ferramentas Microsoft, através do *Microsoft Campus Agreement*[7].

COMPONENTES DO SISTEMA DE INFORMAÇÃO

O modelo transaccional de bases de dados desenvolvido para o SIESTBarreiro/IPS é composto por 4 bases de dados relacionais, uma para cada módulo do sistema de informação. A estrutura completa da base de dados pode ser consultada no Anexo III desta dissertação.

Em termos das funcionalidades disponibilizadas pelo sistema de informação, podemos consultar na Tabela 10 um resumo das principais funcionalidades. Esta tabela foi construída utilizando como base a linguagem de modelação UML[8] (Nunes & O'Neill, 2004), e apresenta as funcionalidades agrupadas por actor do sistema/respectivo *Use Case*.

---

[7] O *Microsoft Campus Agreement* é um programa de licenciamento destinado às instituições de ensino superior e que permite obter software Microsoft a preços bastante reduzidos face aos valores de mercado. O software adquirido ao abrigo deste programa não pode ser utilizado para fins comerciais.

[8] UML – *Unified Modelling Language* é uma linguagem de notação gráfica orientada a objectos, desenvolvida pela empresa Rational Software (www.rational.com), que se tornou um *standard* de-facto para a especificação, visualização, desenvolvimento e manutenção de software (Glassey, 2008).

Tabela 10 – Identificação de Use Cases por Actor

ACTOR	USE CASE
Funcionário	• Registar/validar inscrição em exame • Validar Inscrição no ano lectivo • Emitir certificados e certidões • Validar lançamento de pautas • Alterar os dados do aluno • Disponibilizar pautas de exame • Produzir listagens diversas • Registar pagamento de produto/serviço • Requisitar ou adquirir produto
Professor	• Efectuar lançamento de notas • Consultar lista de inscritos • Inserir sumários • Consultar serviços de docente • Gerir curriculum • Actualizar dados pessoais • Consultar informação do aluno • Definir Programa e Carga Horária (PCH) das Unidades Curriculares de que é responsável • Requisitar ou adquirir produto
Director	• Definir responsáveis dos serviços • Inserir/actualizar dados pessoais do funcionário/professor • Registar/actualizar o contrato do funcionário/professor • Definir as necessidades de serviço docente • Consultar listagens de informação
Aluno	• Consultar os sumários das unidades curriculares • Efectuar inscrição/mudança de turma; • Efectuar pré-inscrição em exame • Consultar informação pessoal • Efectuar pré-inscrição no ano lectivo • Consultar informação académica • Efectuar aquisição de produto/serviço

As Fig. 19 a 22 apresentam a modelação UML dos *Use Cases* da tabela anterior, aplicada aos quatro módulos do Sistema de Informação. A Fig. 24 é referente ao módulo de Recursos Humanos e, como podemos verificar, neste módulo participam três actores: o Director, o Funcionário (tipicamente do secretariado da direcção) e um actor que representa a base de dados "espelho" que vimos na Fig. 23. Esta base de dados contém um espelho da informação introduzida na base de dados principal do SI

e é actualizada sempre que a informação na BD principal sofre alterações em informação disponibilizada aos utentes através do portal web. Para o efeito são utilizados *triggers,* que são mecanismos que, como o próprio nome indica, funcionam como "gatilhos" que dispoletam acções quando determinado evento ocorre na base de dados.

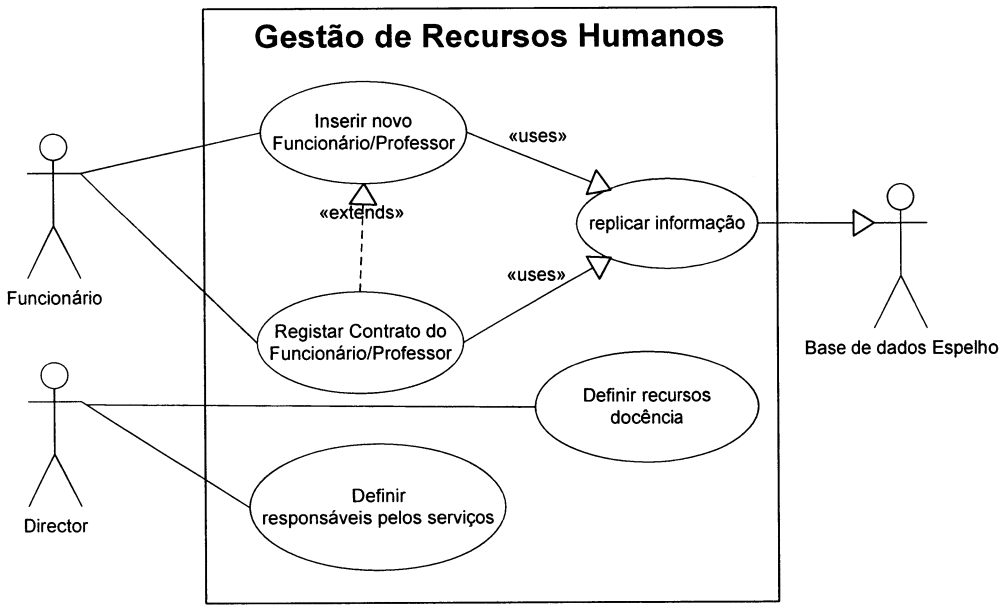


Fig. 24 – Diagrama de Use Cases do módulo Gestão de Recursos Humanos

De seguida temos os *use cases* do módulo Gestão Académica. Este módulo é o mais complexo de todo o sistema, e é também o mais importante para a instituição: para além de representar as principais actividades de funcionamento, é também, ou por consequência, o módulo que mais actores envolve. Este factor reveste-o de especial importância em termos de disponibilidade, pois qualquer eventual falha neste sistema coloca em causa a imagem de todo o sistema.

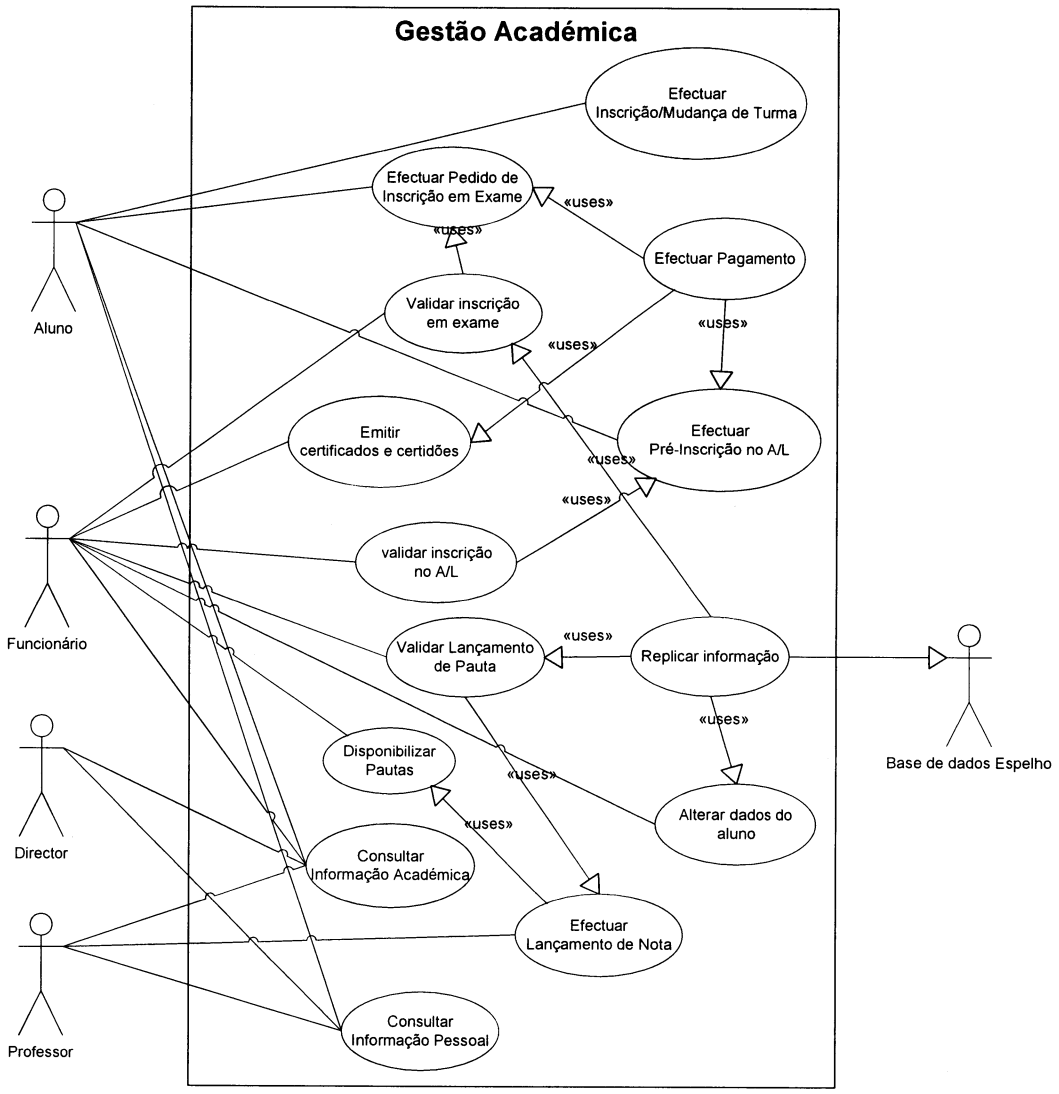


Fig. 25 – Diagrama de Use Cases do módulo Gestão Académica

Por último temos os módulos de gestão de produtos e serviços e o de gestão do serviço docente. De notar que no segundo diagrama, o actor professor tem uma generalização do responsável da unidade curricular e do responsável da área científica, ambos com atribuições diferentes.

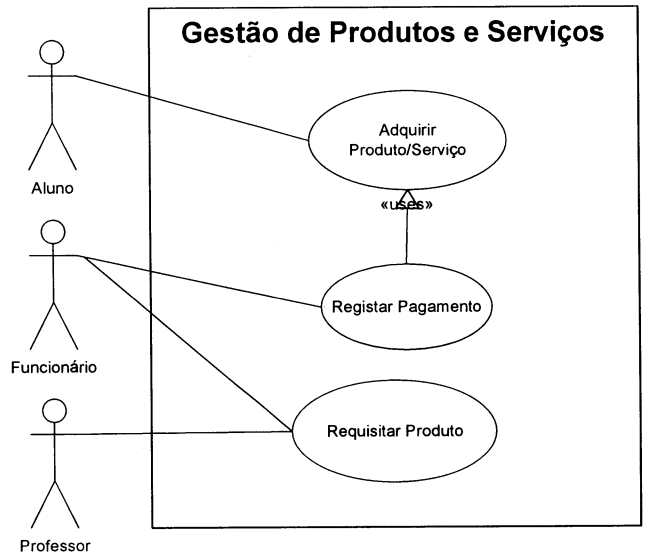


Fig. 26 – Diagrama de Use Cases do módulo Produtos e Serviços

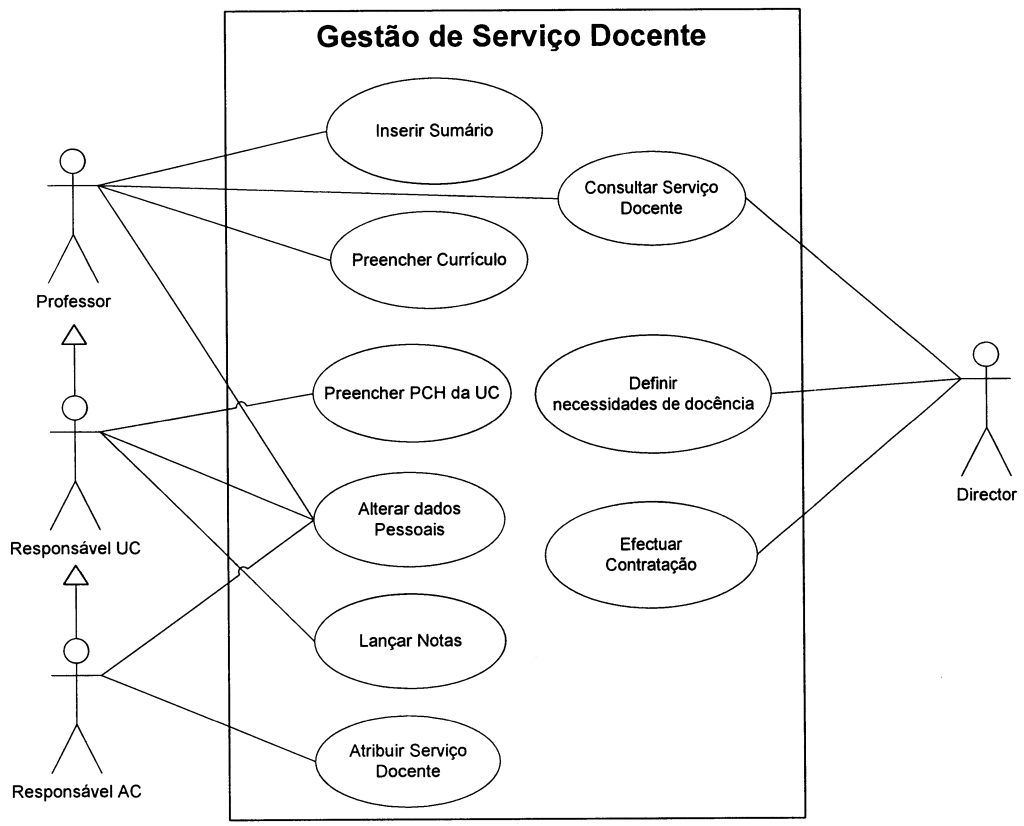


Fig. 27 – Diagrama de Use Cases do módulo Serviço Docente

4. O Modelo Desenvolvido

De forma a dotar a instituição IPS e, numa primeira fase, a sua unidade orgânica ESTBarreiro/IPS de uma ferramenta que pemita a análise de informação de acordo com as características referidas nos capítulos anteriores, foi desenvolvido um Data Warehouse, baseado no modelo dimensional de Ralph Kimball (Kimball & Ross, 2002), tendo em conta os diferentes fluxos de informação e os desafios que se colocam às Instituições de Ensino Superior anteriormente abordados.

Este modelo organizacional vai permitir disponibilizar a informação aos diferentes actores da instituição de acordo com as suas necessidades, autonomizando-os em relação à tecnologia, dotando-os de uma potente ferramenta de análise. Assim, qualquer utilizador poderá aceder à informação de que necessita e tratá-la da forma que mais lhe convier.

O Data Warehouse será dimensionado de forma a abranger a informação que necessite de análise, incluída no sistema de informação existente, e relacionar essa informação com a forma de "fazer negócio" da instituição, acrescentando toda a informação relacionada com os processos identificados e que ainda não esteja contemplada no SI ou que provenha de fontes externas de dados.

A facilidade que a abordagem de Ralph Kimball nos permite em termos do faseado desenvolvimento da solução de Data Warehouse foi outro aspecto importante para a escolha da mesma. O facto de ser possível construir o Data Warehouse de acordo com os processos de negócio da instituição permite não só a possibilidade de fasearmos o seu desenvolvimento processo a processo como também é ideal no caso em que os departamentos de TI têm dificuldades em termos de recursos humanos. Este é normalmente o caso das Instituições de Ensino Superior, principalmente as de Ensino Superior Politécnico. Não só existe dificuldade em conseguir contratar profissionais como também é muito complicada a sua manutenção no serviço (melhor remuneração no sector privado é uma das razões mais comuns para a saída da instituição).

Assim, o modelo proposto irá definir os processos de negócio mais importantes em termos de análise para a organização e será implementado um protótipo funcional com um desses processos. Esse protótipo será validado pelos seus utilizadores, nomeadamente, o director da instituição e alguns responsáveis das áreas científicas de conhecimento da ESTBarreiro/IPS. Posteriormente o acesso será alargado aos restantes actores envolvidos: os funcionários responsáveis pelos vários serviços envolvidos.

A implementação com sucesso de um Data Warehouse depende de um conjunto alargado de factores, não é suficiente conseguir identificar correctamente os processos de negócio, desenvolver o modelo dimensional, ter uma ferramenta de

construção de Data Warehouses tecnologicamente muito avançada ou mesmo desenhar um processo ETL eficiente e eficaz para conseguir construir com sucesso um Data Warehouse. É uma combinação destes e de outros factores que permitem uma implementação bem sucedida.

A principal preocupação de quem constrói um Data Warehouse deve ser o foco no valor do negócio da organização – é este o factor mais determinante de sucesso a longo prazo (Kimball, Mundy, & Thornthwaite, 2006). Devemos "despir" a pele de técnicos e preocuparmo-nos com as necessidades do negócio, pois um Data Warehouse centrado apenas na tecnologia pode acarretar o risco de trazer pouco valor acrescentado para a organização, não sendo por isso adoptado pela mesma. A Fig. 28 apresenta o ciclo de vida do desenvolvimento dimensional da organização.

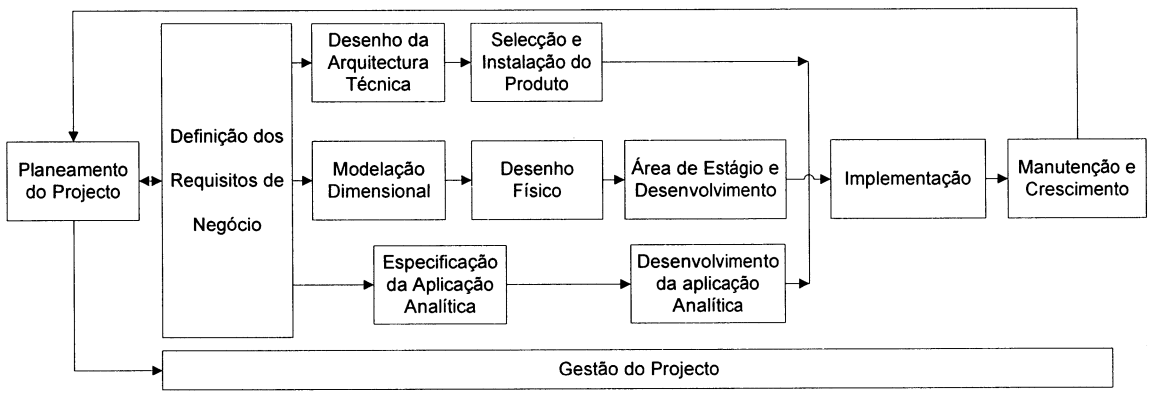


Fig. 28 – Ciclo de vida do desenvolvimento dimensional da organização (Kimball & Ross, 2002)

Segundo Ralph Kimball, este diagrama permite identificar as principais actividades necessárias para o correcto desenvolvimento e manutenção de um Data Warehouse. É possível identificar uma sequência de desenvolvimento do projecto, a dependência entre tarefas e as tarefas que podem ocorrer simultâneamente. Como podemos também observar pela Fig. 28, o processo de criação de um Data Warehouse é um processo que se pode repetir sempre que se necessitar de aumentar a sua abrangência no interior da organização, ou seja, sempre que fôr necessário abranger mais um determinado processo de negócio, voltamos a iniciar o ciclo de vida dimensional.

4.1. Construção do Modelo Dimensional

Em seguida vamos abordar as várias actividades necessárias para definir os requisitos da organização, aplicando cada uma, sempre que assim se justifique, ao modelo desenvolvido.

PLANEAMENTO E GESTÃO DO PROJECTO

O projecto tem início, como seria de esperar, com a fase do planeamento. Nesta fase é estudada a capacidade da organização para implementar o projecto, definem-se os pressupostos para o mesmo e estuda-se a aplicabilidade da solução proposta ao problema, ou seja, verifica-se se o problema existente tem como melhor solução o desenvolvimento de uma ferramenta de Data Warehouse e se existem condições na organização para dar início ao projecto. Existem cinco factores decisivos para o sucesso de um projecto de Data Warehouse (por ordem decrescente de importância) (Fichter, 2009; Kimball & Ross, 2002):

- A existência de um patrocínio forte – deverá existir na organização um apoio forte e a convicção de que o Data Warehouse é uma ferramenta indispensável para a organização. Embora por um lado seja importante que este apoio tenha um rosto visível, alguém facilmente identificável como importante para a organização, não convém que se centre apenas em uma pessoa, dado o ponto crítico em que o mesmo se tornaria para o sucesso do projecto. No caso do objecto desta dissertação, foi possível obter o melhor dos patrocínios – o director da Escola, que inclusivé foi recentemente nomeado Pró-Presidente do IPS, com o sistema de informação IPS como competência directa. Para além do director, que também possui uma importância chave na àrea científica de Mecânica e Estruturas, existem também alguns responsáveis de áreas científicas que se envolveram activamente no projecto (como as áreas de Química e de Arquitectura Urbanismo e Vias de Comunicação), e que representam áreas fundamentais da organização da escola;
- Existir um motivo forte para a construção do Data Warehouse – o Data Warehouse não deve ser abordado como mais uma ferramenta à disposição da organização mas sim como a solução para problemas críticos para o negócio que o tornem uma ferramenta essencial. Neste caso, não só a análise de dados incentivada no interior da Instituição com vista à melhoria dos resultados, como as próprias obrigações perante a tutela (o Ministério da Ciência e Ensino Superior) em termos da disponibilização e envio de informação e análise de indicadores (como por exemplo o DIMAS e o REBIDES) são fortes motivações a que o sistema actual tem dificuldade em dar resposta e que o Data Warehouse permite responder de forma eficaz, como veremos no capítulo destinado à análise de resultados;
- Possibilidade de execução – o projecto deverá ter as condições necessárias à sua implementação, quer em termos de tecnologia, de recursos humanos e, principalmente, em termos de dados disponíveis. Sem dados que alimentem o Data Warehouse, não será possivel o sucesso do projecto. Estes dados deverão, no entanto, também possuir a qualidade que permita acrescentar valor às decisões (Dijcks, 2004);

- Relação entre o negócio e o sector informático – é importante que os membros do sector informático conheçam e respeitem o negócio e vice-versa. Ao terem de trabalhar em conjunto (convêm lembrar que podem ser aplicações de vários fabricantes, com suporte técnico diverso), o sucesso de implementação dependerá, embora não sendo um factor crítico, da relação entre estas duas estruturas. No caso em estudo, como o desenvolvimento é feito internamente e à medida, este problema encontra-se solucionado – existe uma forte ligação e conhecimento entre o Núcleo de Informática e, neste caso, os vários sectores da escola, particularmente com a direcção da mesma. Para além deste relacionamento, existe também uma boa percepção mútua do que envolve a realização das tarefas de cada uma destas estruturas;
- Cultura analítica – a forma como a organização está habituada a suportar a tomada de decisões é também um factor importante. Se as decisões habitualmente forem tomadas por instinto ou percepção do funcionamento do negócio, a introdução de uma ferramenta analítica pode não ser tão fácil. Mais uma vez, no caso em estudo e até pela natureza científica de que se reveste a instituição em causa, este aspecto não se revelou ser um problema.

Em termos da actividade de gestão do projecto, esta fase consiste no acompanhamento do projecto, garantindo que existem os seguintes recursos (Cardoso et al, 2003):

- Um plano de execução do projecto e documentação do mesmo, que identifique todas as tarefas a realizar;
- A gestão da abrangência do projecto, que identifique quais os processos prioritários e que, portanto, serão abordados numa primeira fase;
- Um plano de comunicação do projecto à comunidade, que permita, por um lado explicar o funcionamento da solução desenvolvida e, por outro, cativar os diferentes utilizadores para o projecto.

O plano de execução do projecto para o Data Warehouse em estudo é apresentado na Tabela 11. A duração indicada corresponde ao tempo médio previsto para cada tarefa, não refectindo uma execução cronológica correspondente ao ano civil. O objectivo traçado de implementação do protótipo foi de aproximadamente 6 meses, calculando uma média de 4 horas diárias, 5 dias por semana. Os espaços em branco representam actividades com uma duração média inferior a dois dias, ficando normalmente englobados a par de outras tarefas.

As fases de desenvolvimento posteriores ao ponto 6 correspondem ao desenvolvimento do protótipo, que consiste na implementação do Data Mart correspondente ao processo de negócio "Percurso Académico" e não a todo o Data Warehouse. Em termos dos actores envolvidos são apresentados os principais actores que participaram na elaboração dos requisitos. O Director representou o duplo papel

de Director da instituição e também de Presidente do Conselho Científico. O responsável da área científica representa também o responsável pela Comissão de avaliação da qualidade.

Tabela 11 – Plano de execução do projecto

Nº	Tarefa	Duração (semanas)	Actores envolvidos
1	Reunião inicial, definição dos requisitos de negócio. Reuniões individuais com cada elemento	1	Director; Presidente do Conselho Pedagógico; Responsável área científica; NucInf[9]
2	Análise dos dados internos e externos geridos pela instituição	4	NucInf
3	Desenvolvimento do modelo dimensional e análise da adequabilidade aos requisitos iniciais	4	NucInf
4	Apresentação do modelo desenvolvido		Director, NucInf
5	Análise da plataforma a utilizar para implementar a solução desenvolvida	1	NucInf
6	Preparação dos recursos técnicos e implementação da solução	8	NucInf
7	Implementação do front-end	4	NucInf
8	Apresentação do portal e explicação do seu funcionamento.		Director; Presidente do Conselho Pedagógico; Responsável área científica;
9	Correcção de erros e implementação de sugestões	1	NucInf
10	Teste da aplicação pelos utilizadores	2	Director; Presidente do Conselho Pedagógico; Responsável área científica;

A partir da segunda metade da etapa 6 foram realizadas reuniões informais com o director, com uma periodicidade semanal e uma duração de cerca de 1 hora, para efectuar eventuais acertos na informação e no processo ETL realizado.

Quanto à abrangência do projecto, um dos aspectos chave da abordagem de Ralph Kimball preconiza que o desenvolvimento deve acontecer de uma forma faseada, (Kimball & Ross, 2002) concentrando-se esse desenvolvimento num processo de negócio de cada vez. Assim, embora o modelo desenvolvido apresente vários processos de negócio, tentando modelar os processos mais importantes para a organização, o protótipo desenvolvido concentra-se num único processo de negócio que, como veremos adiante, é o que mais importância revela para a organização – o Percurso Académico.

[9] NucInf – Núcleo de Informática da ESTBarreiro/IPS

No que respeita ao plano de comunicação, foram efectuadas uma sessão de apresentação do projecto com os vários intervenientes no mesmo (ponto 8 da tabela anterior) e algumas reuniões individuais de esclarecimento de dúvidas (no decorrer do ponto 10). Nestas foi possível não só apresentar a solução desenvolvida e a forma de trabalhar com as ferramentas de acesso aos dados como também identificar indicadores de medida (factos) em falta e corrigir algumas definições.

DEFINIÇÃO DOS REQUISITOS DO NEGÓCIO

Em seguida temos a fase de definição de requisitos de negócio. Esta fase evolui em comunicação com a fase anterior de planeamento do projecto, como podemos verificar pela seta bidireccional que une as duas actividades (ver Fig. 28) e consiste, como o nome indica, em perceber de forma clara quais são os requisitos do negócio. Estas primeiras fases têm sido alvo de atenção ao longo dos últimos dois capítulos e resultam do contacto com os intervenientes no negócio, efectuado através de contactos informais, do trabalho directo no sistema de informação existente com o Núcleo de Informática da Instituição e de várias reuniões com os vários actores principais participantes nas tomadas de decisão.

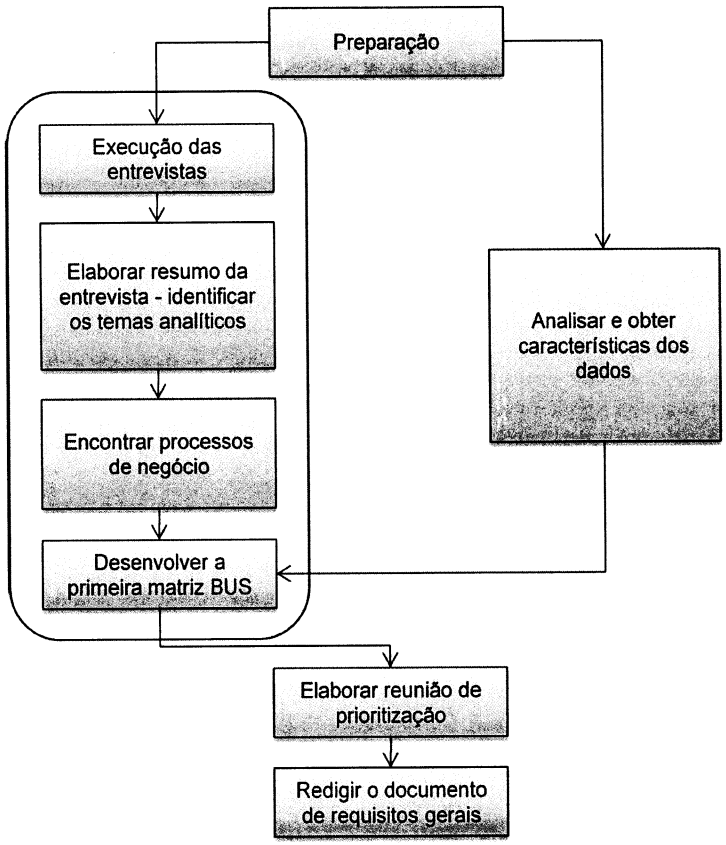


Fig. 29 – Fluxograma da definição de requisitos de negócio

Fonte: (Kimball, Mundy, & Thornthwaite, 2006)

A Fig. 29 apresenta um fluxograma que permite decompor a actividade de definição dos requisitos de negócio em várias sub-actividades. No lado esquerdo do fluxograma as quatro actividades encontram-se agrupadas para indicar que, embora estejam identificadas como actividades independentes, normalmente ocorrem de forma encadeada para cada entrevista, ou seja, após realizar uma entrevista deve-se logo elaborar um resumo da mesma, identificando todos as aspectos de análise mencionados pelo entrevistado e os processos de negócio que os suportam, colocando os resultados numa matriz que se vai construindo à medida que se realiza cada entrevista.

A primeira actividade que deve ocorrer é a fase de preparação em que não só identificamos os indivíduos que importa entrevistar na organização como estruturamos a forma como vamos conduzir as entrevistas e adaptamos as perguntas a cada um dos entrevistados. Para perceber quem devemos incluir na lista de entrevistas, uma ferramenta muito útil é o organograma da instituição. O organograma da ESTBarreiro/IPS é apresentado na Fig. 30.

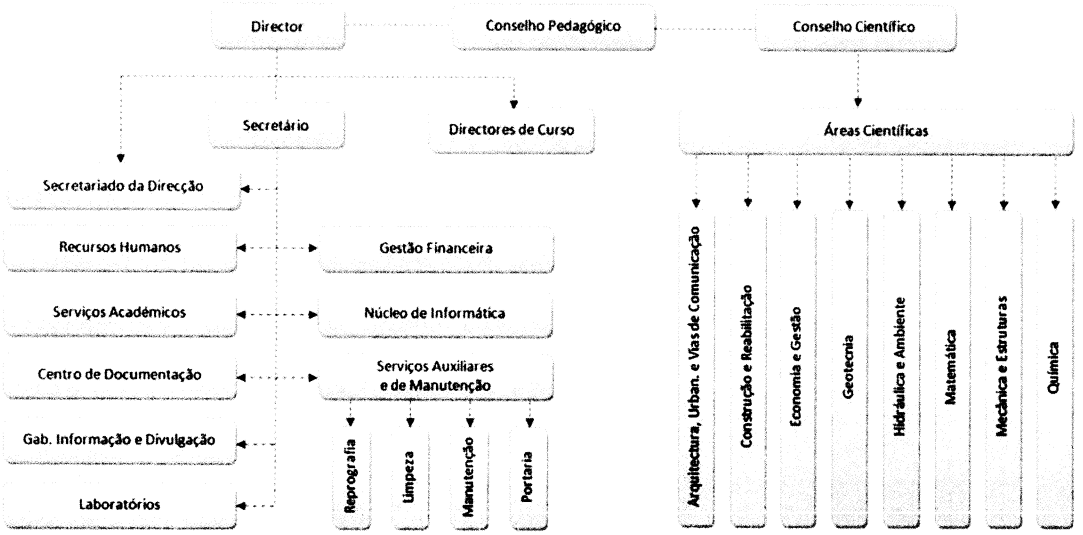


Fig. 30 – Organograma da ESTBarreiro/IPS

Um aspecto muito importante na preparação destas reuniões consiste em tentar comunicar na linguagem do entrevistado e não conduzir a reunião entrando pelo campo dos aspectos técnicos. Devem ser evitadas perguntas como "*o que pretende ver implementado com o Data Warehouse?*" pois a resposta muito certamente será "*tudo!*" (Kimball, Mundy, & Thornthwaite, 2006). Devemos deixar que seja o entrevistado a falar sobre os aspectos do seu trabalho, indicando por exemplo quais as suas responsabilidades e competências que o cargo exige e não pedir para enumerar os indicadores que analisa na tomada de decisões de modo a que consigamos construir o modelo logo na entrevista.

Desta forma, não condicionamos o entrevistado e conseguimos extrair mais informação acerca da forma de funcionamento do seu cargo. Outro aspecto importante a ter em consideração na realização dos contactos com os vários actores é conseguir absorver a terminologia utilizada (Cardoso et al, 2003): é muito importante que se perceba a terminologia utilizada pelos entrevistados pois nem sempre se utilizam os mesmos termos para designar a mesma realidade. Isto realça a importância da utilização de normas nos dados sempre que possível, como iremos abordar no subcapítulo dedicado à área de estágio e desenvolvimento.

Tabela 12 – Lista de principais entrevistas para desenvolver o modelo

Elemento entrevistado	Outras funções desempenhadas
Director	Presidente Conselho Científico
Presidente Conselho Pedagógico	Responsável de Unidade Curricular Responsável pelos Serviços Académicos
Presidente Comissão de avaliação da qualidade	Responsável de Unidade Curricular Responsável Área Científica
Responsável Financeiro	-

A Tabela 12 apresenta os intervenientes na organização identificados como importantes para a definição de requisitos de negócio e aos quais foram efectuadas entrevistas. Não estão incluídos na tabela os contactos efectuados com elementos técnicos do sistema de informação e outros participantes indirectos no desenvolvimento do Data Warehouse. Em termos de escalonamento, é preferível que se iniciem as entrevistas por ordem crescente de importância dos entrevistados na organização, começando por estabelecer contactos horizontalmente na hierarquia (Cardoso et al, 2003), começando com elementos das áreas operacionais e, sempre que possível, conhecidos e de confiança. O resultado destes contactos aplicados ao desenvolvimento do Data Warehouse da ESTBarreiro/IPS permite (Kimball, Mundy, & Thornthwaite, 2006):

- Preparar melhor as reuniões a efectuar com os elementos de chefia;
- Ter uma visão mais abrangente da organização;
- Perceber o vocabulário utilizado.

É importante referir que os contactos com as pessoas responsáveis técnicos pelo sistema de informação, deverão ser intercalados com estas entrevistas para conseguir perceber que dados existem e de que forma se encontram organizados. Como o Data Warehouse possui um desenvolvimento faseado é importante que tenhamos, sempre que desenvolvemos uma fase, a noção do seu enquadramento na organização como um todo para não correr o risco de criar estruturas fechadas que restrinjam a circulação da informação (Cardoso et al, 2003). Assim, ao mesmo tempo que

conduzimos as reuniões devemos também estar a analisar os dados e verificar as caracteristicas e limitações dos mesmos – é a actividade do ramo da direita na Fig. 29.

Tabela 13 – Resumo dos temas de análise identificados

Elemento entrevistado	Temas analíticos encontrados
Director	Proceder a análise da composição do corpo docente da ESTBarreiro/IPS: • Estudo por ETI's, por categoria, por habilitação, etc; • Informação para o REBIDES; Analisar a evolução das avaliações por: • Ano lectivo ou semestre; • Unidade curricular, área científica ou curso; • Estudante; Verificar se existe relação entre as inscrições (número, tipo, curso, etc.) em Unidades Curriculares num semestre e o insucesso nesse semestre Analisar a evolução do sucesso escolar por UC Decidir sobre o número de turmas a colocar em funcionamento em cada semestre
Presidente Conselho Pedagógico	Avaliar o desempenho pedagógico do corpo docente e respectiva evolução através da análise dos inquéritos pedagógicos Analisar a assiduidade às diferentes UC's Analisar os resultados escolares dos estudantes Analisar problemas de sucesso escolar e compreender as suas origens Analisar as inscrições em UC's, para estabelecimento dos calendários escolares
Presidente Comissão de avaliação da qualidade	Elaborar relatório dos resultados de avaliação por semestre e ano lectivo, analisando o desempenho escolar e diversas vertentes por curso, por tipo de inscrição, por contingente, etc. Acompanhar o cumprimento dos programas das UC's
Responsável Financeiro	Estudar a evolução das vendas por período; Obter informação sobre os produtos a adquirir no final do ano com o orçamento restante Analisar o andamento do pagamento de propinas
Responsáveis das AC's (Áreas Científicas)	Conhecer as disponibilidades docentes da AC, em cada semestre, tendo como objectivo a planificação do serviço docente Conhecer as necessidades de serviço docente da AC (análise às inscrições em UC's da AC) Conhecer (gerir) as verbas atribuídas à AC
Director de Curso	Avaliar a opinião dos estudantes relativamente ao desenrolar das actividades pedagógicas nas UC's, resultante dos inquéritos pedagógicos Analisar o desempenho escolar dos estudantes do curso Analisar a assiduidade às UC's do Curso Acompanhar o desenvolvimento das actividades lectivas das UC's com o respectivo programa Elaborar Relatório de Bolonha
Presidente Científico	Acompanhar processos de creditação de competências Decidir sobre número de vagas a atribuir a cada curso Desencadear procedimentos de renovação de contratos de docentes Decidir relativamente à distribuição de serviço docente por semestre

Após cada entrevista deve-se elaborar um resumo da mesma identificando os temas de análise (ou em Inglês, *analytic themes*), chegando ao final de todas as entrevistas com uma lista dos temas de análise mais comuns, também conhecidos em Inglês por *common analytic themes.* Na Tabela 13 é apresentado um resumo dos temas analíticos mais importantes identificados nas entrevistas realizadas. A última linha, "Outros", engloba todos os temas analíticos encontrados no decurso de contactos informais e entrevistas realizadas com os restantes membros da organização.

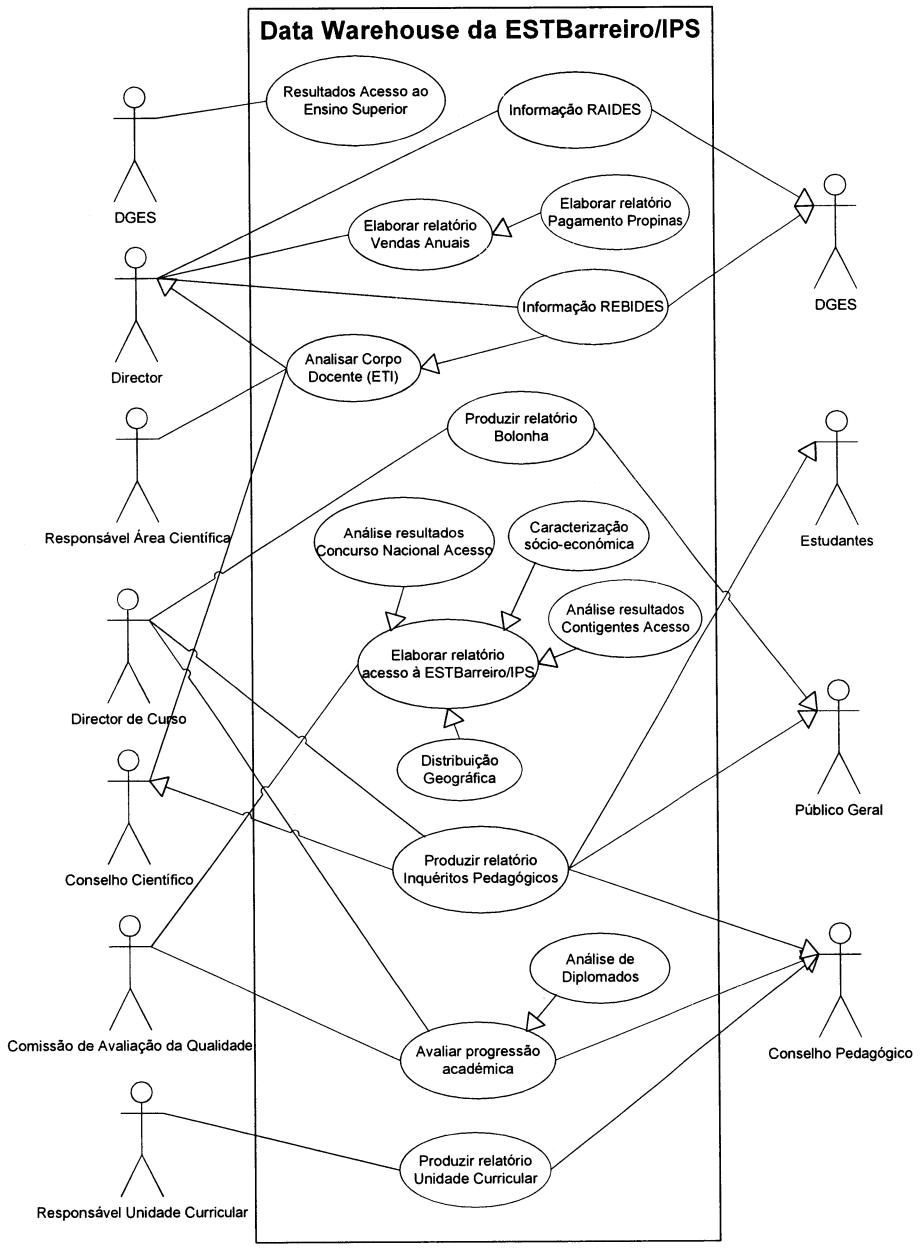


Fig. 31 – Diagrama UML com Use Cases para o Data Warehouse

No final desta ronda de entrevistas e contactos informais devemos conseguir definir quem serão os utilizadores do Data Warehouse e de que forma tomam as suas

decisões. O diagrama da Fig. 31 apresenta a modelação UML de *use cases* com a descrição de requisitos do projecto, obtidos nessas entrevistas.

É utilizada a linguagem de modelação visual UML devido à simplicidade que possui para representar um sistema de informação (Nunes & O'Neill, 2004) e à utilização de símbolos padrão que permitem a comunicação com as várias equipas envolvidas no desenvolvimento da solução. Esta linguagem é também aquela em que os membros do Núcleo de Informática possuem mais experiência de modelação e que utilizam para as várias necessidades de modelação que existem nos sistemas de informação desenvolvidos actualmente.

Tabela 14 – Identificação dos Processos de negócio

Tema analítico	Análises necessárias	Processo de negócio
Resultados de acesso ao ensino superior	Elaborar relatório de acesso ao ensino superior. Efectuar seguintes análises: • resultados Concurso Nacional de Acesso; • caracterização sócio-económica; • resultados por contigente de acesso; • análise por distribuição geográfica.	Ingresso na instituição
Informação RAIDES	Inquérito ao Registo de Alunos Inscritos e Diplomados do Ensino Superior Fonte: (GPEARI)	Ingresso na Instituição Percurso Académico
Informação REBIDES	Relatório com informação obrigatória a enviar para o Ministério	Análise Corpo Docente
Análise Corpo Docente	Actividade lectiva; Estudo de ETI	Análise Corpo Docente
Percurso académico	Análise dos resultados obtidos por semestre e por ano lectivo; Análise do abandono escolar	Percurso Académico
Análise de diplomados	Diplomados por curso e ano lectivo; Análise do percurso académico do diplomado	Percurso Académico
Relatório Unidade Curricular	Análise dos resultados obtidos por semestre e anuais por unidade curricular; Resultados dos alunos em 1ª inscrição; Resultados dos alunos em outras inscrições	Percurso Académico
Análise de vendas	Evolução das vendas por período; Vendas por categoria	Vendas
Estudo de propinas	Relatório de propinas pagas por ano lectivo; Propinas pagas por aluno	Vendas
Relatório de Bolonha	Análise dos inscritos por curso e por ano	Ingresso na Instituição Percurso Académico
Inquéritos por Unidade Curricular	Média das respostas : • ao inquérito da unidade curricular; • ao professor, pelo tipo de aula leccionada.	Percurso Académico

Após a identificação de todos os temas de análises comuns e antes de passar à etapa seguinte é importante validar os resultados obtidos apresentando-os aos entrevistados e discutindo eventuais correcções e/ou melhorias a efectuar.

Terminada esta validação, podemos passar à identificação dos processos de negócio utilizados para suportar as análises que necessitamos de efectuar. A Tabela 14 apresenta um resumo dos processos de negócio identificados.

A MATRIZ BUS INICIAL

Terminada a fase de descoberta dos processos de negócio, vamos inseri-los numa matriz que permite funcionar como o mapa de dados da organização. Esta matriz vai permitir ligar os processos de negócio às entidades e objectos que participam nesses mesmos processos. A matriz BUS inicial para a ESTBarreiro/IPS é apresentada na Tabela 15. Nesta matriz, as linhas correspondem aos processos de negócio apresentados, enquanto as colunas apresentam as várias entidades que nelas participam – são as dimensões que vão ser construidas no modelo dimensional. Às dimensões que participam em mais do que um processo de negócio e que partilham uma estrutura de atributos idêntica damos o nome de dimensões confome (serão abordadas em maior detalhe no subcapítulo modelação dimensional).

Tabela 15 – Matriz BUS Inicial

Processos de Negócio	Dimensões																								
	Data	Hora	Aluno	Professor	Formação Académica	Funções de Gestão	Regime Contratual	Área Científica	Curso	Unidade Curricular	Entidade	Estado da Avaliação	Ano Curricular	Assiduidade	Sumário	Turma	Regime Estatutário	Tipo de Inscrição (exame)	Regime de Ingresso	Actividade Lectiva	Programa e Carga Horária	Tipo de Aula	Informação de Acesso	Produto	Fornecedor	TOTAL
Percurso Académico	x		x	x				x	x	x		x	x	x		x	x	x				x				13
Análise do Corpo Docente	x			x	x		x							x	x					x	x					8
Inscrição / Acesso	x		x						x										x				x			5
Vendas / Requisições	x	x	x					x	x		x		x						x				x	x	x	11

Finalmente, devemos efectuar uma última reunião que envolva os actores da Tabela 12, a equipa de desenvolvimento do projecto e o(s) patrocinador(es) do mesmo, denominada reunião de prioritização (Kimball, Mundy, & Thornthwaite, 2006),

para definir qual a ordem de importância dos processos de negócio para a organização. Devem ser apresentados e explicados os processos de negócio encontrados e os temas analíticos que os mesmos permitem abordar para que todos os participantes na reunião percebam a solução apresentada.

Deverá ser utilizada uma matriz de duas linhas por duas colunas como podemos verificar na Fig. 32, em que no eixo horizontal se indica a facilidade de implementação do processo de negócio e no eixo vertical se representa o valor ou importância que cada processo representa na organização. Esta matriz resulta de uma outra que provém da área de gestão, designada por matriz BCG (abreviatura de Boston Consulting Group), e que permite analisar o portefólio de negócios das empresas, em termos do crescimento/participação no mercado das áreas de negócio envolvidas (Londrevie et al., 1996).

Para definir a prioritização dos processos, cada elemento presente na reunião deverá distribuir os processos de negócio pelas quatro áreas existentes na matriz até que se chegue a um consenso sobre a localização de todos os processos existentes. A Fig. 32 representa o resultado dessa reunião aplicada ao desenvolvimento do Data Warehouse da ESTBarreiro/IPS.

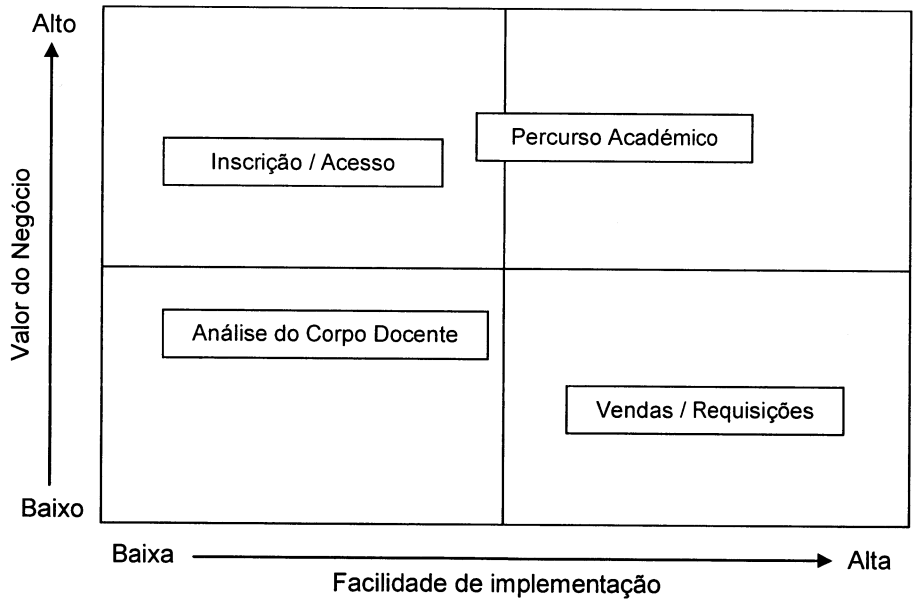


Fig. 32 – Matriz de prioritização

Da análise desta matriz podemos concluir, por exemplo, que o processo de maior valor acrescentado para a organização é o processo "Percurso Académico", sendo no entanto de dificuldade moderada em termos da sua implementação. Por outro lado, o processo de vendas é mais exequivel mas tem um impacto no negócio menos importante. É importante que os técnicos que vão desenvolver o Data Warehouse liderem a atribuição do nível de dificuldade de execução a cada um dos processos,

pois são estes que melhor percebem as dificuldades existentes a nível técnico (qualidade e disponibilidade dos dados, por exemplo) com que se vão deparar. (Kimball, Mundy, & Thornthwaite, 2006). A dificuldade atribuída aos processos de negócio "Inscrição/Acesso" e "Análise do Corpo Docente" está relacionada precisamente com o problema apresentado no exemplo anterior: a qualidade e disponibilidade dos dados existente dificultam o seu desenvolvimento.

Após esta análise com todos os futuros utilizadores do Data Warehouse, foi decidido iniciar o desenvolvimento do mesmo pelo processo "Percurso Académico" pois existe na instituição uma necessidade urgente em analizar os seus resultados. O facto de começar o desenvolvimento do Data Warehouse por um processo de negócio com elevado valor para o negócio pode acarretar um risco mais elevado (Kimball & Ross, 2002), pois uma eventual falha terá uma visibilidade elevada e o crédito atribuído à tecnologia pode ficar afectado. No entanto, após apresentado o risco à direcção, como se trata de uma iniciativa no âmbito de um trabalho académico e possui o patrocínio forte do próprio director da instituição, manteve-se a opção por este processo de negócio.

Terminada a fase de definição dos requisitos do negócio, o desenvolvimento do projecto de Data Warehouse passa a desenvolver-se em três eixos de execução paralela:

- Eixo tecnológico – composto pelos componentes desenho da arquitectura técnica e selecção e instalação do produto, que vão servir de plataforma ao desenvolvimento do Data Warehouse;
- Eixo de dados – composto pela modelação dimensional, o desenho físico e a área de estágio e desenvolvimento;
- Eixo analítico – composto pelas actividades de especificação das aplicações analíticas e pelo seu respectivo desenvolvimento;

EIXO DE DADOS

Esta é das componentes mais importantes do desenvolvimento do Data Warehouse e que normalmente consome mais tempo a implementar. Para o desenvolvimento do modelo dimensional, a actividade principal é, como o próprio nome indica, a Modelação Dimensional. Assim, vamos abordar esta actividade de forma detalhada, de forma a construir o modelo. As restantes actividades serão abordadas com o desenvolvimento do protótipo, no sub-capítulo 4.2.

MODELAÇÃO DIMENSIONAL

A primeira actividade a abordar no eixo de dados é a elaboração do modelo dimensional do Data Warehouse. Esta actividade envolve a utilização dos conceitos teóricos abordados no sub-capítulo 2.4 do Capítulo 2, para construir a estrutura de

cada processo de negócio identificado na fase anterior. Assim, vamos aplicar as 4 etapas de desenvolvimento do modelo dimensional para construir o modelo.

A primeira etapa consiste em definir o processo de negócio a modelar. No caso do Data Warehouse da ESTBarreiro/IPS (que passará a ser, a partir deste capítulo, abreviado por DW ESTBarreiro/IPS), com a matriz BUS da Tabela 15, obtida na fase de definição de requisitos de negócio, foram identificados 4 processos de negócio principais, dos quais foi estabelecida uma prioridade que podemos consultar na matriz de prioritização da Fig. 32. Assim, o processo de negócio a modelar será o "Percurso Académico" dos alunos na ESTBarreiro/IPS.

Normalmente, se a definição dos requisitos de negócio tiver sido cuidadosamente elaborada e estudada com os intervenientes da organização, então o processo de negócio por onde se vai iniciar o Data Warehouse deverá estar correctamente identificado. É para isso de extrema importância que se façam estes dois instrumentos de trabalho fundamentais: a matriz BUS e a matriz de prioridades.

A segunda etapa do desenho dimensional consiste na definição do nível de detalhe ou grão a aplicar ao modelo desenvolvido. Neste caso, o nível de detalhe do processo "Percurso Académico" identificado foi:

"O registo de avaliação obtido por um aluno a uma inscrição num tipo de avaliação (ou exame) de uma Unidade Curricular"

Com este nível de detalhe será possível aos utilizadores realizar todas as consultas que os dados existentes na organização permitem, ou seja, não haverá qualquer restrição ao nível de detalhe disponível ao utilizador. Nesta fase deve-se tentar encontrar sempre o maior nível de detalhe possível – que permita, por um lado, responder às necessidades dos utilizadores e, por outro, que seja possível obter, tendo em vista os sistemas transaccionais existentes e a forma como estes armazenam a informação. Se a identificação do nível de detalhe inicial de cada processo de negócio a modelar cumprir a regra anterior torna-se sempre possível analisar a informação com a precisão que cada utilizador do sistema necessitar.

Na terceira etapa vão ser identificadas as dimensões que participam no processo de negócio a modelar. Na fase de definição dos requisitos de negócio já foi efectuada uma primeira abordagem às dimensões, utilizando a matriz BUS, e que permitiu identificar as dimensões mais comuns. Algumas destas dimensões também são logo identificadas a partir do nível de detalhe escolhido na etapa anterior.

Mas, como o desenvolvimento do DW é feito de uma forma faseada, um Data Mart de cada vez, é uma boa prática que se efectue uma sessão de brainstorming para confirmar as dimensões que vão participar no Data Mart, pois o período de tempo entre a definição de requisitos e a construção de cada Data Mart pode ser suficiente

para surgir nova informação ou a necessidade de ajustes nos requisitos do negócio para o Data Mart a desenvolver.

Por outro lado, esta é a fase em que começamos realmente a "meter as mãos na massa", ou seja, que começamos a verificar na prática a implementação do Data Warehouse. As dimensões identificadas na matriz BUS podem nesta etapa sofrer algumas alterações, como dar origem a várias dimensões ou mesmo deixarem de ser dimensões por se chegar à conclusão de não fazerem sentido na aplicação ao modelo. A sessão de brainstorming, juntamente com a verificação da aplicabilidade de cada dimensão e respectivo teste de funcionamento vai permitir identificar correctamente todas as dimensões necessárias. Assim, as dimensões identificadas para o processo de negócio "Percurso Académico", para o nível de granularidade definido na etapa anterior foram as seguintes:

Data

A dimensão Data possui duas realidades de utilização diferentes neste processo de negócio: a data de inscrição e a data de avaliação do aluno a um exame da Unidade Curricular. Assim, esta dimensão irá desempenhar dois papéis neste processo de negócio o que fará com que, ao nível da implementação do Data Mart esta dimensão se transforme numa dimensão *role playing*, como veremos no próximo sub-capítulo de forma detalhada, dedicado ao protótipo desenvolvido.

Semestre Lectivo

O Semestre Lectivo é uma das dimensões que não constava da Matriz BUS inicial e que foi identificada na sessão de brainstorming realizada nesta etapa da modelação dimensional. Numa primeira análise, o Semestre Lectivo seria obtido através de uma hierarquia construída a partir de atributos presentes na dimensão Data, "Ano Lectivo → Semestre Lectivo". No entanto, após analisar a especificidade do negócio, concluiu-se que teria de ser criada uma dimensão própria, pois a dimensão Data não permite identificar o semestre lectivo respectivo à avaliação. Ou seja, a data de avaliação do aluno a um exame numa Unidade Curricular pode, em termos do calendário escolar, ser efectuada num semestre ou mesmo ano lectivo diferente daquele a que realmente a avaliação diz respeito. Por exemplo, se na dimensão Data o calendário escolar do ano lectivo 2008/2009 tiver início a 15/09/2009 e fim a 14/09/2010, uma avaliação realizada dentro deste intervalo seria automaticamente classificada como pertencente a esse ano lectivo. No entanto, devido a questões de calendarização interna da instituição, é frequente que ocorram, principalmente nos casos de épocas especiais de exame, avaliações fora destes períodos limite do ano lectivo, o que, se fosse utilizada a dimensão Data, erradamente classificaria a avaliação no ano lectivo seguinte e constituía um erro do ponto de vista da análise da informação na organização. Por outro lado, também existem situações em que são abertas excepções para casos especiais de inscrição (atestado médico, etc.) em que as datas de inscrição e

avaliação não coincidem com o ano lectivo e/ou semestre em que a avaliação terá de ser considerada. Por estes factores, decidiu-se criar uma dimensão própria para o tratamento do calendário lectivo de avaliação.

Aluno

Esta dimensão foi relativamente fácil de identificar, visto se tratar de uma entidade fundamental no processo de negócio – o aluno. A composição desta dimensão será semelhante ao que já acontece com o sistema transaccional, existindo um registo para cada aluno com todos os atributos que o caracterizam. Esta é uma dimensão que será utilizada por vários processos de negócio, tratando-se por isso de uma dimensão conforme. Vai conter várias alterações do tipo 2, como iremos analisar no desenvolvimento do protótipo, no sub-capítulo 4.2.

Unidade Curricular

A dimensão Unidade Curricular é outra dimensão de grande importância para o processo de negócio em estudo e que vai ser utilizada com vários níveis de agregação, como veremos na actividade seguinte, "configuração e desenho físico". Todos os atributos que caracterizam cada Unidade Curricular leccionada na escola serão incluídos nesta dimensão. Outro aspecto importante desta dimensão é a hierarquia precedências da Unidade Curricular. Como se trata de uma hierarquia de profundidade ou número de níveis variável, a solução passa pela implementação de uma tabela ponte.

Curso

A dimensão Curso foi alvo de alguma discussão com os vários participantes no processo de *brainstorming*. A questão que se colocava era se a dimensão Curso deveria realmente existir como uma dimensão ou se a informação sobre o curso deveria fazer parte da dimensão Unidade Curricular e existir nesta uma hierarquia que permitisse efectuar operações de roll-up e drill-down.

A decisão final recaiu para a criação da dimensão Curso como uma entidade própria devido essencialmente a dois aspectos:

- Em primeiro lugar o curso representa, em termos do negócio da organização, uma entidade autónoma, sobre a qual se efectuam várias análises, independentes da sua estrutura curricular, e que é utilizada em vários processos de negócio;
- Por outro lado, a dimensão Unidade Curricular é uma dimensão fundamental para o negócio da organização, e sobre a qual incidem frequentemente alterações do Tipo 2, o que conduziu à tomada da decisão de criar a dimensão Curso como uma minidimensão;

Área Científica

A Área Científica é uma dimensão que também evoluiu de uma forma semelhante à dimensão Curso e que permite o registo de todas as características relacionadas com as áreas científicas da ESTBarreiro/IPS.

Como vimos no organograma da ESTBarreiro/IPS da Fig. 30, a organização interna não obedece a uma estrutura departamental, mas antes é efectuada em áreas científicas, que por sua vez "prestam serviços" aos vários cursos. Esses serviços correspondem essencialmente às várias Unidades Curriculares ministradas na Instituição. A Fig. 33 apresenta uma representação esquemática em matriz da organização académica da Instituição.

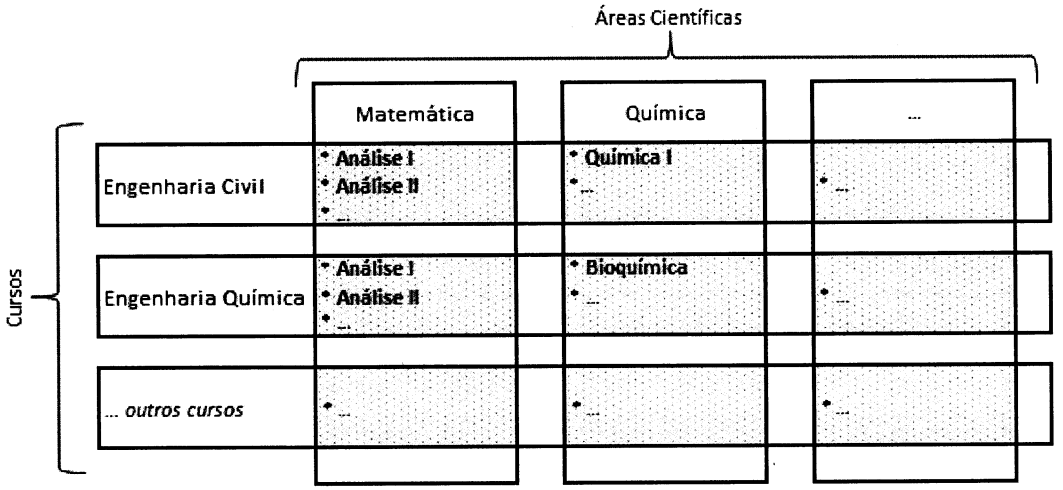


Fig. 33 – Estrutura organizativa por Áreas Científicas

Professor

A dimensão Professor é outra dimensão base do Data Mart, que permite conter todos os atributos descritivos da entidade professor da Unidade Curricular. É uma dimensão que, à semelhança da dimensão Data, também representa dois papéis distintos no processo de negócio em estudo: o de professor responsável pela unidade curricular e o de professor que leccionou essa mesma unidade curricular. Assim, também nesta dimensão existirá a aplicação do conceito de dimensão *role-playing*.

Corpo Docente

Esta dimensão corpo docente é onde irá participar a dimensão Professor com um dos papéis que desempenha: o de Professor que lecciona a Unidade Curricular. No entanto, possui uma característica particular: estabelece uma relação *muitos-para-muitos* (também designadas relações *m:n*) com a tabela de factos do Data Mart "Percurso Académico". Quer isto dizer que, para um registo da tabela de factos existem vários registos da tabela de dimensões e, do mesmo modo, a cada registo da tabela de dimensões correspondem vários registos da tabela de factos – esta última

constitui a situação normal na relação entre as tabelas de dimensão e as tabelas de factos: uma relação de *um-para-muitos* (também designadas relações *1:n*). Traduzido para uma situação concreta, a docência de uma mesma Unidade Curricular pode ser assegurada por vários docentes e esse grupo de docentes pode leccionar várias Unidades Curriculares.

Assim, para resolver esta situação procedeu-se à criação de uma tabela "ponte" que irá permitir transformar as relações de *m:n* em relações de *1:n*.

Turma

A dimensão turma permite dotar o processo de negócio de informação relacionada com a turma a que o aluno pertence em cada registo de um tipo/época de avaliação numa Unidade Curricular. Esta dimensão também mereceu uma atenção especial devido às divergências que levantou na sessão de *brainstorming* e na análise das fontes de dados com o pessoal técnico.

Surgiram essencialmente dois problemas relacionados com a informação existente relacionada com as turmas. Por um lado, a informação no sistema operacional encontra-se bastante incompleta, principalmente nos primeiros anos da instituição, não existindo informação em suporte digital. Depois, verificou-se na reunião de *brainstorming* que a informação existente também não possui a qualidade necessária para permitir análises fidedignas: embora exista um registo dos alunos inscritos por turmas, não existe controlo por meios informáticos que permita controlar a que turmas os alunos realmente assistem, e a informação transmitida pelos presentes na reunião não foi animadora: os alunos mudam frequentemente de turma a que assistem mediante um acordo verbal com o professor da mesma.

Após esta análise de funcionamento das turmas no sistema operacional da organização e da reunião de brainstorming, foi possível chegar à conclusão que não era de todo possível aproveitar a informação das turmas, pelo menos na versão inicial do Data Warehouse da Universidade do Oriente!

Tipo de Inscrição e Estado de Avaliação

Por último foram identificadas, na matriz BUS inicial, duas entidades que também participam no processo "Percurso Académico": o Tipo e o Estado de Avaliação do aluno a uma Unidade Curricular. No entanto, após a sessão de *brainstorming* efectuada, foram analisadas as características destas dimensões e foi possível verificar que a utilização das mesmas necessitava de alguma revisão.

A dimensão Tipo de Inscrição possui as características relacionadas com o tipo de inscrição efectuada a que o aluno se submeteu para avaliação a uma determinada Unidade Curricular, como por exemplo, a avaliação por Exame Mensal, por Época Especial, etc. Possui um número de atributos reduzido e de hipóteses também limitado.

Quanto ao Estado de Avaliação, este não constitui uma verdadeira dimensão, na medida em que representa apenas um atributo que indica o estado final da avaliação, como por exemplo "Aprovado", "Não Avaliado", etc. e um outro atributo com a descrição da avaliação. Visto que os atributos não pertencem a nenhuma outra dimensão já identificada, estamos perante uma dimensão "Lixo" ou *Junk dimension*. As dimensões com esta designação surgem quando, após terem sido identificadas as dimensões que participam no Data Mart e definidos os respectivos atributos, se verifica que alguns atributos ficam "órfãos" de dimensão, ou seja, não se adequam a nenhuma das dimensões definidas. Nestes casos existem três situações possíveis (Kimbal & Ross, 2002; Caldeira, 2008):

- O atributo é adicionado à tabela de factos como uma medida – esta solução, para além de aumentar significativamente o tamanho da tabela de factos, não traz vantagens para o modelo dimensional a inclusão de atributos textuais na tabela de factos;
- É criada uma dimensão para o atributo – esta situação pode dar origem a um número elevado de tabelas de dimensão;
- É criada uma dimensão "lixo" para os atributos – esta situação permite manter o modelo dimensional com um nível de complexidade reduzida. Devemos, no entanto, ter em atenção quando agrupamos vários atributos que não se relacionam: se a dimensão se tornar muito grande, temos de considerar a necessidade de ter mais do que uma dimensão "Lixo";

Após esta análise das duas estruturas que haviam inicialmente sido pensadas com duas dimensões, foi considerado útil a sua união em apenas uma dimensão "lixo", com a designação "Tipo de Inscrição e Estado de Avaliação", visto que a dimensão resultante não implicaria um número elevado de registos e permitiria simplificar o modelo dimensional desenvolvido.

A Tabela 16 apresenta um exemplo da estrutura final da dimensão "Tipo de Inscrição e Estado de Avaliação".

Tabela 16 – Atributos da dimensão "Tipo de Inscrição e Estado de Avaliação"

Chave do Tipo de Inscrição e Estado de Avaliação	Designação do Tipo de avaliação	Descrição do Tipo de avaliação	...	Designação do Estado de Inscrição
1	Época Normal			Aprovado
2	Exame Mensal			Aprovado
...	...	...	...	...
n	Época Normal			Reprovado
...	...			...

Antes de passarmos à quarta e última etapa da modelação dimensional, a Tabela 17 apresenta um resumo de todas as dimensões utilizadas no processo de negócio "Percurso Académico", no seu nível de detalhe mais elevado.

Tabela 17 – Lista de Dimensões do Processo de negócio "Percurso Académico"

PERCURSO ACADÉMICO	
Dimensão	Caracterização resumida
Data	Permite registar a data de inscrição e a data de avaliação do aluno num exame da Unidade Curricular
Semestre Lectivo	Centrar as medidas em análise no ano lectivo e semestre correspondente
Aluno	Características dos alunos que realizaram as avaliações
Unidade Curricular	Características das Unidades Curriculares
Curso	Dimensão com a informação sobre os Cursos Ministrados na Instituição e turmas disponíveis
Área Científica	Características da área científica
Corpo Docente	Corpo docente que lecciona a Unidade Curricular
Tipo e Época de Avaliação	Dimensão "Lixo" que agrega os atributos isolados, desprovidos de dimensão própria
Professor	Dimensão que representa dois papéis: o de Professor responsável e o de Professor que lecciona a Unidade Curricular Regista a informação pessoal sobre o Professor

A última etapa de desenho do modelo dimensional, após terem sido identificadas todas as dimensões, é a identificação dos factos gerados pelo processo de negócio. Normalmente, os factos estão directamente relacionados com o processo de negócio a modelar (Kimball, Mundy, & Thornthwaite, 2006). Neste caso, em que o processo a modelar está relacionado com o percurso académico dos alunos, existem algumas medidas que se identificam de forma quase imediata, como a nota ou classificação obtida num registo de avaliação a uma unidade curricular num determinado tipo de avaliação. Outros não são de identificação tão imediata, como o ano curricular em que o aluno se encontrava no momento em que obteve a respectiva classificação.

Tabela 18 – Factos do processo de negócio "Percurso Académico"

Indicador de medida (facto)	Descrição
Nota	Classificação obtida pelo aluno à Unidade Curricular no tipo de avaliação respectivo
Créditos ECTS[10]	Número de créditos atribuídos ao aluno pela obtenção de nota positiva na Unidade Curricular
Peso	Peso da Unidade Curricular na média final
Número de Inscrição	Número da inscrição efectuada pelo aluno à Unidade Curricular
Ano Curricular	Ano curricular do curso que o aluno frequentava no momento em que obteve a classificação.

[10] ECTS – *European Credit Transfer and Accumulation System*, ou Sistema Europeu de Acumulação e Transferência de Créditos, em Português (European Commission, 2009)

Dada a importância e centralidade que as medidas assumem no Data Warehouse, é muito importante que invistamos algum tempo na sua identificação. Os factos identificados para o processo "Percurso Académico" são apresentados na Tabela 18.

Estes factos são os fundamentais para medir a performance do processo de negócio. Existem algumas combinações e cálculos que podem ser efectuados a partir destes factos e que permitem efectuar outras análises, que iremos abordar na próxima actividade do Eixo de Dados da modelação dimensional – o Desenho Físico.

No sub-capítulo destinado ao protótipo, também iremos verificar em detalhe a aplicação destas medidas calculadas.

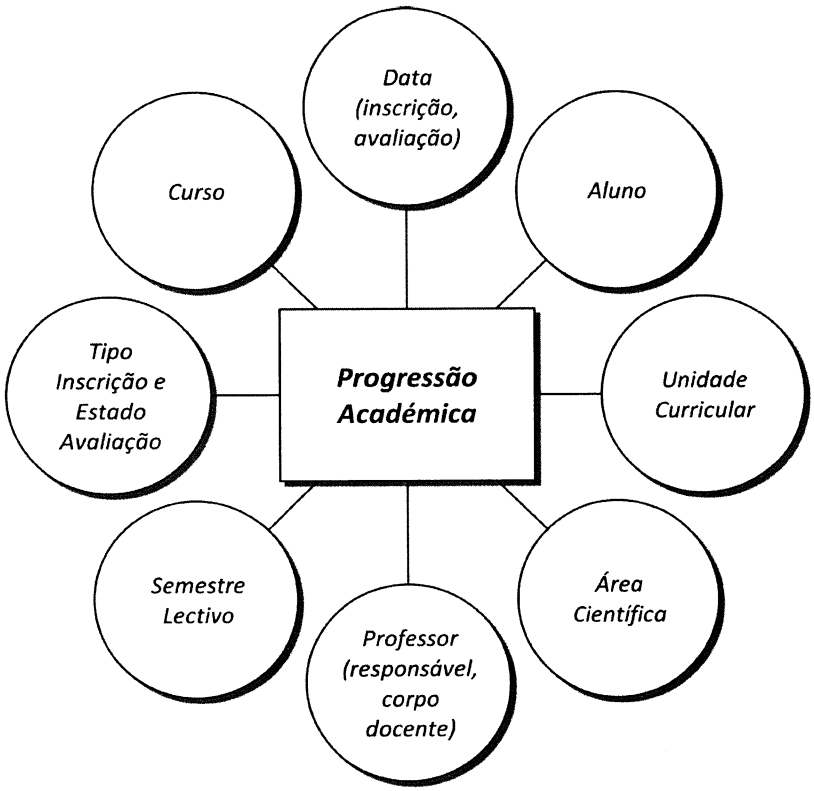


Fig. 34 – Modelo dimensional inicial do processo "Percurso Académico"

Com esta última etapa termina a fase inicial de modelação dimensional, e podemos sumarizar esta fase do modelo num esquema que é apresentado na Fig. 34, também designado por *bubble chart*.

De seguida, podemos iniciar o preenchimento das tabelas dimensão e de factos com a lista de atributos que caracterizam cada uma.

Os atributos identificados foram obtidos através do cruzamento de duas fontes: a reunião de *brainstorming* com os os vários intervenientes que vão utilizar o modelo desenvolvido e a análise realizada em conjunto com os técnicos do Núcleo de Informática que permitiu identificar as fontes de dados disponíveis.

Dimensão Data

A dimensão Data irá conter os atributos constantes na Tabela 19. Esta dimensão terá dois papéis, Data de Inscrição e Data de Avaliação que serão descriminados no sub-capítulo 4.2, "Protótipo Desenvolvido".

Tabela 19 – Atributos da Dimensão Data

Nome do Atributo	Descrição	Exemplo
Data	Data de Inscrição e de Avaliação	01/01/2010
Descrição Completa da Data	Descrição por extenso da data	1 de Janeiro de 2010
Dia da Semana	Dia da semana por extenso	Sexta-Feira
Dia do Mês	Ordinal do mês	1
Dia do Ano	Ordinal do Ano Civil (1 a 366)	1
Número do Mês	Ordinal do Mês (1 a 12)	1
Número da Semana	Ordinal da semana do ano Norma ISO8601: 1ª Semana → 1ª Segunda-feira do Ano	53
Indicador de Feriado	Dia feriado ou Não feriado	Feriado
Indicador de Fim-de-semana	Dia de Semana ou fim-de semana	Dia de Semana
Ano Civil	Ordinal do ano do calendário civil	2010
Semestre Civil	Descrição do semestre da data	1º Semestre
Trimestre Civil	Descrição do trimestre da data	1º Trimestre
Mês do Ano	Extenso do mês da data	Janeiro
Momento Lectivo do Ano	Evento do calendário académico da data	Férias de Natal

Não foram incluídos nesta dimensão os atributos relacionados com datas fiscais, pois a Instituição IPS, com o novo Regime Jurídico das Instituições de Ensino Superior (RGIES), por questões de organização e funcionamento interno, encontra-se numa fase de centralização dessa informação nos Serviços da Presidência. Esta matéria será alvo de atenção no Capítulo "Conclusão e estudos futuros".

Também não são incluídos os atributos que permitem localizar o registo de avaliação no ano e semestre lectivo pelas razões indicadas na identificação da dimensão Semestre no início deste sub-capítulo.

Dimensão Semestre Lectivo

Para a dimensão Semestre Lectivo foram identificados os atributos da Tabela 20.

Tabela 20 – Atributos da Dimensão Semestre

Nome do Atributo	Descrição	Exemplo
Designação do Semestre	Designação do semestre em que se realizou a avaliação	ÍMPAR
Ano Lectivo	Ano Lectivo a que o semestre respeita	2009/2010
Data de Início do Semestre	Data em que o semestre se inicia	20/09/2009
Dia de Fim do Semestre	Data em que o Semestre Termina	27/02/2010

Esta dimensão, embora possua poucos atributos, é bastante importante para o processo de negócio modelado pois é muito solicitada nas análises efectuadas pelos

utilizadores, como veremos no capítulo destinado à apresentação de resultados. Permite ter um funcionamento semelhante como se tratasse de uma hierarquia da dimensão Data. As agregações efectuadas, que veremos na próxima actividade do eixo de dados, "Desenho físico", também tiram partido desta dimensão.

Dimensão Professor

A Tabela 21 apresenta a lista de atributos identificados para a Dimensão Professor. Esta dimensão vai desempenhar os papéis "Responsável da Unidade Curricular" e "Corpo Docente da Unidade Curricular", como veremos no Protótipo Desenvolvido.

Tabela 21 – Atributos da Dimensão Professor

Nome do Atributo	Descrição	Exemplo
Código do Professor	Dimensão degenerada do sistema operacional	712
Nome do Professor	Nome completo do professor	António Pereira
Data de Nascimento	Data de nascimento do professor	01/09/1978
Localidade de Residência	Localidade do professor, de acordo com a norma dos CTT	Lisboa
Código Postal	Código Postal do professor	1500-169
Morada	Morada do professor, de acordo com a norma dos CTT	Rua Bocage, Nº1
Número de BI	Número do Bilhete de Identidade	11339999
Data de Validade do BI	Data de validade do Bilhete de Identidade	01/09/2012
Estado Civil	Estado civil do Professor	Solteiro(a)
Número de Contribuinte	Número de Identificação Fiscal	219999999
Bairro Fiscal	Bairro Fiscal do Número de Identificação Fiscal	Lisboa
Código da Repartição de Finanças	Código da repartição de Finanças	2178
Número de Telefone	Número de telefone fixo do Professor	Não Disponível
Número de Telemóvel	Número de telefone móvel do Professor	969999999
Número de ADSE	Número alfanumérico da ADSE	017399999OA
NIB	Número de Identificação Bancária	003509990009999360048
E-Mail #1	Conta de e-mail institucional	Antonio.pereira@estbarreiro.ips.pt
E-Mail #2	Conta de e-mail pessoal	A_pereira@hotmail.com
Dependentes	Número de dependentes	0
Sexo	Masculino	Masculino
Nome do Pai	Nome completo do Pai	João Pereira
Nome da Mãe	Nome completo da Mãe	Edite Pereira
UserName	Nome de utilizador do Professor	Antonio.pereira
Número do Gabinete	Número do Gabinete	Gab. 2.04
Extensão telefónica	Extensão telefónica do Gabinete do professor	2203

Dimensão Aluno

A dimensão Aluno é a que possui um maior conjunto de atributos. Estes atributos identificados para a dimensão Aluno são apresentados na Tabela 22.

Tabela 22 – Atributos da Dimensão Aluno

Nome do Atributo	Descrição	Exemplo
Número do Aluno	Dimensão degenerada do sistema operacional. Representa o número do aluno	7
Nome do Aluno	Nome completo do aluno	Rui Silva
Data de Nascimento	Data de nascimento do aluno no BI	06/10/1978
Número do BI	Número do Bilhete de Identidade	11339999
Data de Validade do BI	Data de validade do BI (DD/MM/AAAA)	23/10/2012
Local de Emissão do BI	Local de emisão do BI	Setúbal
Distrito – Aulas	Distrito de Residência, em tempo de aulas	Setúbal
Concelho – Aulas	Concelho de residência, em tempo de aulas	Moita
Localidade – Aulas	Localidade de residência, em tempo de aulas	Moita
Código Postal – Aulas	Código postal em tempo de aulas	2864-007
Morada – Aulas	Morada, em tempo de aulas	Rua das Flores, Nº 3
Distrito – Férias	Distrito de residência, em férias	Faro
Concelho – Férias	Concelho de residência, em férias	Portimão
Localidade – Férias	Localidade de residência, em férias	Portimão
Código Postal – Férias	Código postal de residência em férias	8500-081
Morada – Férias	Morada, em tempo de férias	Rua das Naus, Nº1
Nome do Pai	Nome completo do Pai	João Silva
Nome da Mãe	Nome completo da Mãe	Maria Silva
Sexo	Sexo do Aluno	Masculino
Número de Contribuinte	Número de contribuinte do Aluno	119036999
Data de Entrada	Data em que o aluno ingressou na ESTBarreiro/IPS	15/09/1999
Nacionalidade	Nacionalidade no BI do aluno	Portuguesa
Naturalidade	Naturalidade no BI do aluno	Lisboa
Telefone	Contacto telefónico fixo do aluno	265523999
Telemóvel	Contacto telefónico móvel do aluno	969899999
E-Mail #1	E-Mail institucional	Rui.silva@estbarreiro.ips.pt
E-Mail #2	E-Mail externo	My_mail@mail.pt
Tipo de Inscrição	Tipo da inscrição	Maiores de 23 anos
Ano Lectivo de Entrada	Ano lectivo em que o aluno ingressou, pela primeira vez na ESTBarreiro/IPS	1999/2000

A dimensão Aluno é uma das que será utilizada em vários processos de negócio, constituíndo assim uma dimensão conforme do Data Warehouse da ESTBarreiro/IPS.

Dimensão Curso

Os atributos identificados para a dimensão Curso são apresentados na Tabela 23. Nesta dimensão, para além dos atributos específicos do curso estão também incluídos os atributos que caracterizam as turmas disponíveis.

Tabela 23 – Atributos da Dimensão Curso

Nome do Atributo	Descrição	Exemplo
Código do Curso	Dimensão degenerada do sistema operacional que permite identificar o curso (definida pela DGES)	9089
Designação do Curso	Designação completa do curso (aprovada pela DGES)	Licenciatura em Engenharia Civil
Abreviatura do Curso	Designação abreviada do curso	Engenharia Civil – Diurno
Acrónimo do Curso	Designação de três caracteres que permite identificar o curso	CVD
Regime do Curso	Regime de funcionamento do curso: pode ser diurno ou nocturno	Diurno
Ciclo de Estudos do Curso	Designação do ciclo de estudos a que o curso pertence	1º Ciclo de Estudos
Abreviatura do Curso para o Cartão	Abreviatura do curso para o cartão de estudante	Eng. Civil-D
Estado do Curso	Indicador de estado do curso	OPEN

Dimensão Tipo e Estado Avaliação

Os atributos identificados para a Dimensão "Tipo de Inscrição e Estado de Avaliação" são apresentados na Tabela 24.

Tabela 24 – Atributos da Dimensão Tipo de Inscrição e Estado de Avaliação

Nome do Atributo	Descrição	Exemplo
Código do Tipo de Avaliação	Dimensão degenerada do sistema operacional para o tipo de avaliação	1
Designação do Tipo de Avaliação	Designação atribuída ao tipo de avaliação possível	Normal
Descrição do Tipo de Avaliação	Descrição textual do tipo de Avaliação	Avaliação da Inscrição no início do semestre
Participa na média final	Atributo que indica se o tipo de avaliação participa na média final	Participa na média final
Altera o número de inscrição		Altera o número de Inscrição
Designação do Estado de Avaliação	Estado final de avaliação do aluno à Unidade Curricular	Aprovado
Descrição do Estado de Avaliação	Descrição textual do estado final de avaliação do Aluno à Unidade Curricular	O aluno foi aprovado à Unidade Curricular

Esta dimensão inclui os atributos órfãos de dimensão "Designação do Estado da Avaliação" e "Descrição do Estado da Avaliação". Insere-se, portanto, no conceito de *Junk Dimension*.

Dimensão Área Científica

Quanto à dimensão Área Científica, os atributos identificados são os apresentados na Tabela 25. Esta dimensão possui algumas alterações do tipo 2, como iremos verificar no próximo sub-capítulo, Protótipo desenvolvido.

Tabela 25 – Atributos da Dimensão Área Científica

Nome do Atributo	Descrição	Exemplo
Código da Área Científica	Dimensão degenerada do sistema operacional que permite identificar o curso	1
Designação da área Científica	Designação atribuída à área científica pelo Conselho Científico	Matemática
Abreviatura da Área Científica	Designação abreviada da área científica	MAT
Piso da Área Científica	Piso do edifício atribuído à área científica	Piso 2
Telefone Geral da Área Científica	Extensão geral interna da área científica	2000
Responsável pela Área Científica	Nome do responsável pela área científica. Tem de ser um docente da instituição	Ana Andrade
Telefone do Responsável da Área Científica	Telefone do gabinete do responsável da área científica	2204
E-Mail do Responsável da Área Científica	Contacto de e-mail institucional do responsável pela área científica	Ana.andrade@estbarreiro.ips.pt

A dimensão Funcionário podia desempenhar a função de dimensão *outrigger* para a dimensão Área Científica, pois o responsável da Área Científica é designado de entre os Professores da Instituição. No entanto, como o número de atributos que importa reter do responsável da Área Científica é reduzido, e as análises necessárias não são muito frequentes, a opção tomada passou por incluir esses atributos na própria dimensão da Área Científica e não recorrer à técnica de dimensão *Outrigger*.

Dimensão Unidade Curricular

Sempre que possível, recorreu-se à utilização de normas (internas da instituição, nacionais ou internacionais) para formatar as designações e os valores dos atributos definidos.

Tabela 26 – Atributos da Dimensão Unidade Curricular

Nome do Atributo	Descrição	Exemplo
Código da Unidade Curricular	Código utilizado no sistema operacional para identificar a Unidade Curricular (dimensão degenerada)	CVD1005
Designação da Unidade Curricular	Nome da Unidade Curricular no Plano de Estudos (nomenclatura definida pelo Conselho Científico)	Mecânica A
Abreviatura da Unidade Curricular	Abreviatura definida pelo Conselho Científico para a Unidade Curricular	Mec A
Tipo da Unidade Curricular	Tipo da Unidade Curricular, de acordo com catalogação do MCTES	Ciências de Base
Ano Curricular	Ano do plano de estudos da Unidade Curricular	1º Ano Curricular
Semestre Curricular	Semestre ordinal do Plano de Estudos da Unidade Curricular	1º Semestre
Unidade Curricular Opcional	Indicador de obrigatoriedade ou não da Unidade Curricular no Plano de Estudos	Obrigatória
Unidade Curricular Funciona em ambos os Semestres	Indicador de funcionamento da Unidade Curricular no Plano de Estudos	Funciona em ambos os Semestres

No caso da Dimensão Unidade Curricular, como podemos verificar pela tabela, as normas utilizadas são, quase na sua totalidade, definidas pelo Conselho Científico da instituição ou pelo MCTES – Ministério da Ciência, Tecnologia e Ensino Superior.

A dimensão Unidade Curricular possui os atributos apresentados na Tabela 26.

Modelo Dimensional Final

Após a identificação de todos os atributos das dimensões que participam no processo de negócio "Percurso Académico", e da definição da respectiva tabela de factos para o grão encontrado, vamos construir o modelo dimensional final.

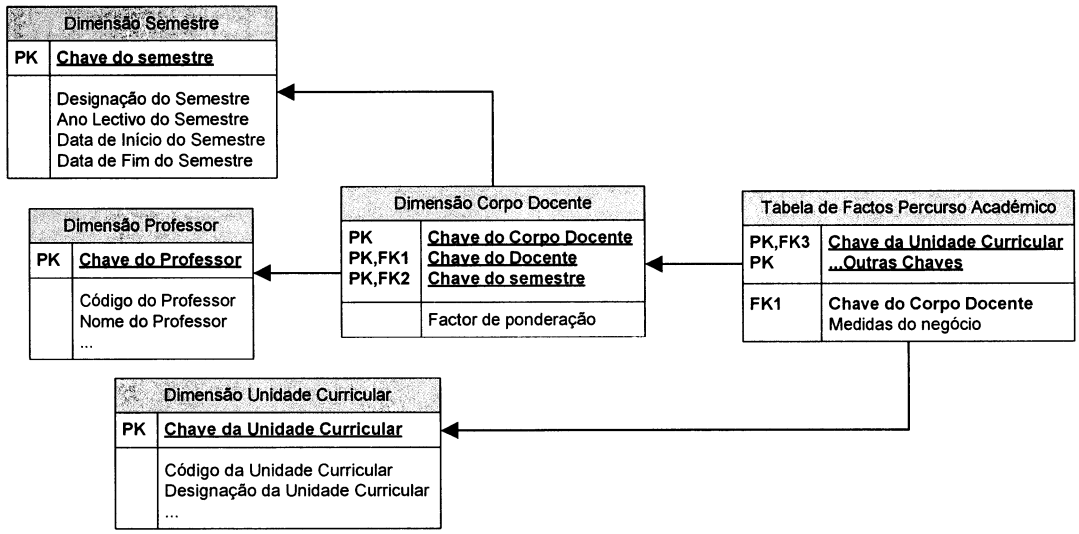


Fig. 35 – Tabela Ponte para a dimensão Corpo Docente

Para isso temos ainda que abordar o desenvolvimento da dimensão Corpo Docente que, como já foi referido, representa uma relação *muitos-para-muitos* com a tabela de factos e portanto terá de ser transformada numa relação *um-para-muitos* através do recurso a uma tabela ponte. A Fig. 35 apresenta uma estrutura de tabela ponte para a dimensão Corpo Docente. O factor de ponderação é um rácio efectuado com base no número de professores que participam na leccionação da Unidade Curricular. Assim, se por exemplo, existirem dois docentes a leccionar uma determinada Unidade Curricular num semestre, então existirá um factor de ponderação de 0,5 para cada um.

Outro aspecto importante é a utilização da dimensão semestre como factor temporal de associação e dissolução de um qualquer corpo docente de uma Unidade Curricular, em vez da utilização de uma data para assinalar o início da criação do corpo docente e uma data para o eventual fim desse mesmo corpo docente. Esta decisão surgiu após a análise da informação armazenada nos sistemas operacionais e e da introdução deste tema na reunião de *brainstorming* efectuada. O funcionamento operacional da instituição em termos da definição do corpo docente possui uma lógica de organização semestral, ou seja, o corpo docente é definido e válido por um

semestre completo. Assim sendo, antes do início de cada semestre, é efectuada uma revisão das necessidades e definido o novo corpo docente para esse semestre (mesmo que o corpo docente se mantenha, o sistema operacional define um novo registo que permite estabelecer a duração de funcionamento desse corpo docente.

O modelo dimensional final para o processo de negócio "Percurso Académico" é apresentado na Fig. 36.

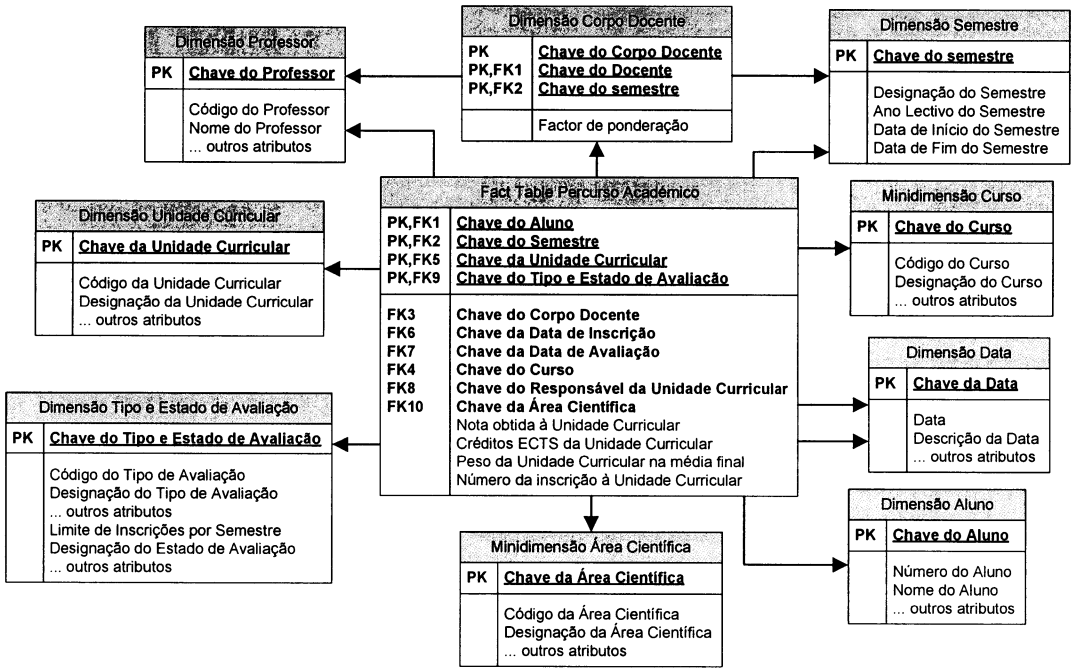


Fig. 36 – Modelo Dimensional para o processo de negócio "Percurso Académico"

A dimensão Data e a dimensão Professor são representadas neste modelo dimensional pelas tabelas físicas. No desenvolvimento do protótipo, vão ser utilizadas na representação de cada papel que desempenham, analisados na abordagem anterior de cada dimensão.

4.2. CONSTRUÇÃO DO PROTÓTIPO

Com a primeira actividade do Eixo de Dados obtivémos o Modelo Dimensional para o processo de negócio "Percurso Académico". As etapas seguintes do eixo de dados vão permitir implementar um protótipo funcional para o modelo identificado que, juntamente com os restantes eixos e actividades de desenvolvimento do ciclo de vida dimensional da organização que vimos na Fig. 28, vão permitir a construção do Data Warehouse.

EIXO DE DADOS

CONFIGURAÇÃO E DESENHO FÍSICO

A actividade de configuração e desenho físico divide-se, como o próprio nome sugere, em duas áreas distintas: numa primeira fase, quando se implementa uma solução de Data Warehouse (ou qualquer outro projecto que possa ter um impacto semelhante numa organização) deve planear-se com cuidado a definição de requisitos de hardware do equipamento em que vai ser instalada a solução preconizada (o tipo de servidor ou servidores, processador(es), a memória necessária, o espaço em disco, etc.) – resumindo, a configuração física do sistema que irá suportar o Data Warehouse. Ainda nesta fase deverá também ser abordada a instalação das ferramentas de software em que a solução irá ser desenvolvida, ou seja, deve abordar-se a instalação do servidor de bases de dados, da plataforma de desenvolvimento das ferramentas de BI e de todo o software que permita dar suporte à solução.

No caso específico desta dissertação, como esta fase introduz algum distanciamento face ao objectivo principal de desenvolvimento da solução conceptual e técnica do Data Warehouse, entrando na área de hardware e software, o seu desenvolvimento não irá merecer uma atenção pormenorizada neste capítulo. Será, no entanto, incluído um anexo com as considerações de instalação de hardware e software para o desenvolvimento do protótipo e de boas práticas no desenvolvimento de soluções de Data Warehouse.

Na segunda fase desta actividade, devemos abordar a forma como iremos passar do modelo lógico de dados desenvolvido nas etapas anteriores para a implementação física do mesmo. Esta abordagem vai ser efectuada particularmente ao nível da construção da base de dados no sistema relacional que suporta o modelo dimensional do Data Warehouse, como por exemplo a forma de construção das tabelas, a utilização de vistas ou *views*, os índices, chaves fictícias, etc. É também nesta fase que se estuda a utilização de agregações, hierarquias e tabelas de estágio.

Construção das Tabelas

Antes de iniciar a construção das tabelas físicas que vão suportar as tabelas de dimensão e de factos identificadas na definição do modelo, devemos ter a seguinte informação definida (Kimball, Mundy, & Thornthwaite, 2006):

- Nomenclaturas – os nomes das tabelas e respectivos atributos foram definidos na construção do modelo no sub-capítulo anterior. Em termos de tabelas físicas e respectivos atributos devemos manter a mesma nomenclatura, pois irá facilitar a construção da base de dados OLAP;
- Tipo de dados e respectivo tamanho – devemos garantir que todas as colunas têm o tipo de dados certo e o tamanho suficiente para armazenar os registos;

- Tipo de tratamento às alterações em atributos – identificar as alterações do tipo 1, tipo 2 e, eventualmente, também do tipo 3;
- Definição de hierarquias – devem ser identificadas as hierarquias existentes em cada tabela de dimensão;
- Definição de índices – sempre que as dimensões possuírem um número elevado de atributos deve analisar-se a necessidade de adicionar índices para melhorar o desempenho de pesquisa nas tabelas. No caso do protótipo desenvolvido, os índices fornecidos pelas chaves primárias das tabelas dimensão são suficientes, pois as dimensões não possuem um tamanho muito grande (após o carregamento inicial, a dimensão com mais registos tinha cerca de 4000 registos).

Em termos de chaves primárias das tabelas, à excepção da dimensão Corpo Docente, todas as tabelas implementadas utilizam como chave primária uma chave fictícia, indicada a negrito na primeira linha de cada tabela.

O tratamento de registos com valores nulos também deve ser alvo de atenção nesta fase. Sempre que possível devemos evitar a ocorrência de valores nulos num Data Warehouse, pois podem confundir o utilizador (Caldeira, 2008): uma pergunta legítima poderá ser "*O atributo está vazio porque aconteceu um erro informático de carregamento de dados ou será que a informação não existe mesmo e está em falta na organização?*". Assim, sempre que um determinado atributo não tenha valor disponível no momento do carregamento para o Data Warehouse – a fase de ETL – devemos adicionar um rótulo que o caracterize, como por exemplo "Informação não disponível". Sempre que um atributo não possa permitir valores vazios, podemos efectuar essa verificação logo ao nível físico da tabela, definindo a propriedade *NOT NULL* para os atributos desejados no momento de criação da tabela.

Um outro aspecto importante na abordagem aos valores nulos está na forma de tratamento de nulos nas relações entre a tabela de factos e as tabelas de dimensão, como por exemplo, quando existe um registo de Avaliação de um Aluno a uma Unidade Curricular que não tem o respectivo Corpo Docente definido. Neste caso, como devemos tratar o registo de chave estrangeira na tabela de factos? Uma metodologia eficaz passa pela introdução de um registo *Dummy* na tabela de dimensão, com a chave primária com o valor -1 para permitir a associação com a tabela de factos. Neste protótipo, a introdução do registo *Dummy* será efectuada directamente pelo processo de ETL.

Para caracterizar cada uma das tabelas a nível físico, vamos percorrer as dimensões do modelo dimensional, identificando cada um destes elementos. A abreviatura "DAL" nas várias tabelas representa o tipo de alteração em cada atributo nas Dimensões de Alteração Lenta. Em seguida abordamos a existência de agregações e caracterizamos as tabelas de factos.

Dimensão Data

A chave primária da dimensão Data, embora se trate de uma uma chave fictícia, composta pelo atributo "Chave da Data", não se trata de uma chave fictícia "normal", sem qualquer significado (que inicie na sequência 1,2,3...) mas sim uma chave que utiliza uma pequena mnemómica na sua construção, nomeadamente, DDAAAAMM. Isto permite que a chave, não deixando de ser um valor numérico, mesmo assim tenha algum significado perceptível para o programador, o que facilita o trabalho de fundo em termos de análises, testes e eventuais correcções à informação efectuada directamente no sistema OLAP. Permite também tratar os valores omissos, como já foi abordado no sub-capítulo 2.4.

Possui a hierarquia "Calendário Civil", composta pelos seguintes atributos:

Ano Civil → Semestre Civil → Trimestre Civil → Mês do Ano

Para o modelo dimensional desenvolvido no sub-capítulo anterior, esta dimensão vai assumir dois papéis distintos: o de Data de Inscrição e o de Data de Avaliação. Uma forma de resolver esta situação poderia ser recorrer à criação de uma vista da dimensão Data para cada um dos papéis que esta terá que desempenhar.

No entanto, a ferramenta que vai ser utilizada para desenvolver a base de dados OLAP do protótipo, o SQL Server Analysis Services, permite tratar directamente as dimensões *role-playing*. Assim, esta solução será implementada na próxima actividade, quando for efectuado o processo ETL.

Tabela 27 – Dimensão Data: tabela física

Data de Inscrição	Data de Avaliação	Tipo de Dados
Chave da Data de Inscrição	**Chave da Data de Avaliação**	**Int**
Data de Inscrição	Data de Avaliação	Smalldatetime
Descrição Completa da Data de Inscrição	Descrição Completa da Data de Avaliação	Varchar(100)
Dia da Semana de Inscrição	Dia da Semana de Avaliação	Varchar(50)
Dia do Mês de Inscrição	Dia do Mês de Avaliação	Int
Dia do Ano de Inscrição	Dia do Ano de Avaliação	Int
Número do Mês de Inscrição	Número do Mês de Avaliação	Int
Número da Semana de Inscrição	Número da Semana de Avaliação	Int
Indicador de Feriado	Indicador de Feriado	Varchar(50)
Indicador de Fim-de-semana	Indicador de Fim-de-semana	Varchar(50)
Ano Civil de Inscrição	Ano Civil de Avaliação	Varchar(4)
Semestre Civil de Inscrição	Semestre Civil de Avaliação	Varchar(20)
Trimestre Civil de Inscrição	Trimestre Civil de Avaliação	Varchar(20)
Mês do Ano de Inscrição	Mês do Ano de Avaliação	Varchar(50)
Momento Lectivo de Inscrição	Momento Lectivo de Avaliação	Varchar(100)

Como o nome dos atributos na dimensão Data já foi definido no Sub-Capítulo 4.2, na Tabela 27, para além do tipo de dados de cada atributo, é apresentado o respectivo nome para cada dimensão *role playing*. Como na dimensão Data não é habitual

ocorrerem alterações de Tipo 2, não existe a necessidade de registar as datas de início e de fim relacionadas com esses acontecimentos.

Dimensão Área Científica

A dimensão Área Científica não possui nenhuma hierarquia. O tipo de dados dos atributos e a gestão de alteração de valores são os apresentados na Tabela 28.

Tabela 28 – Dimensão Área Científica: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave da Área Científica	**Int**	**-**
Código da Área Científica	Varchar(10)	
Designação da Área Científica	Varchar(100)	2
Abreviatura da Área Científica	Varchar(50)	2
Piso da Área Científica	Varchar(15)	
Telefone Geral da Área Científica	Varchar(20)	1
Responsável pela Área Científica	Varchar(100)	2
Telefone do Responsável da Área Científica	Varchar(20)	1
E-Mail do Responsável da Área Científica	Varchar(20)	1
DataInício	Smalldatetime	-
DataFim	Smalldatetime	-

Dimensão Semestre Lectivo

A dimensão Semestre Lectivo é uma dimensão muito particular, que resulta de uma especificidade do negócio abordada no modelo dimensional. Assim, em vez de resultar de uma agregação da informação da dimensão Data, é uma dimensão autónoma, carregada directamente dos sistemas fonte. Possui uma hierarquia, "Calendário Escolar", baseada nos seguintes atributos:

Ano Lectivo → Semestre Lectivo

A Tabela 29 apresenta os tipos de dados da dimensão Semestre Lectivo. Neste caso também não temos alterações de Tipo 2, embora existam algumas de Tipo 1.

Tabela 29 – Dimensão Semestre: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave do Semestre	**int**	**-**
Designação do Semestre	Varchar(25)	
Ano Lectivo do Semestre	Varchar(9)	
Data de Início do Semestre	Smalldatetime	1
Dia de Fim do Semestre	Smalldatetime	1

Dimensão Unidade Curricular

A tabela de dimensão Unidade Curricular possui vários atributos para os quais é necessário guardar histórico (alterações de Tipo 2) e, por isso, tem duas colunas

adicionais, *DataInicio* e *DataFim* para marcar, respectivamente, quando é que o registo entra em vigor e quando é que o mesmo expira de validade.

Tabela 30 – Dimensão Unidade Curricular: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave da Unidade Curricular	**Int**	-
Código da Unidade Curricular (DD)	Varchar(7)	
Designação da Unidade Curricular	Varchar(100)	2
Abreviatura da Unidade Curricular	Varchar(50)	1
Tipo da Unidade Curricular	Varchar(30)	2
Abreviatura do Tipo da Unidade Curricular	Varchar(10)	
Ano Curricular	Varchar(30)	2
Semestre Curricular	Varchar(30)	2
Unidade Curricular Opcional	Varchar(40)	2
Unidade Curricular Funciona em ambos os Semestres	Varchar(40)	2
Data de Início	Smalldatetime	-
Data de Fim	Smalldatetime	-

No que respeita a hierarquias, possui duas hierarquias: uma que permite organizar as Unidades Curriculares por Ano Curricular e Semestre e outra para agrupar as Unidades Curriculares por Tipo.

Ano Curricular → Semestre Curricular → Abreviatura da Unidade Curricular
Tipo de Unidade Curricular → Abreviatura da Unidade Curricular

Dimensão Tipo de Inscrição e Estado de Avaliação

Esta dimensão resulta da junção dos atributos órfãos identificados no desenvolvimento do modelo no sub-capítulo anterior. Podemos identificar a seguinte hierarquia:

Designação do Tipo de Inscrição → Designação do Tipo de Avaliação

Tabela 31 – Dimensão Tipo Inscrição e Estado de Avaliação: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave do Tipo de Inscrição e Estado de Avaliação	Int	-
Designação do Tipo de Inscrição	Varchar(50)	1
Descrição do Tipo de Inscrição	Varchar(200)	1
Participa na média final	Varchar(50)	2
Altera o número de inscrição	Varchar(50)	2
Designação do Estado de Avaliação	Varchar(50)	1
Descrição do Estado de Avaliação	Varchar(200)	1
DataInício	Smalldatetime	-
DataFim	Smalldatetime	-

Dimensão Curso

O tipo de dados dos atributos e o tratamento das alterações na tabela de dimensão Curso é apresentado na Tabela 32.

Tabela 32 – MiniDimensão Curso: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave do Curso	**Int**	**-**
Código do Curso	Varchar(4)	
Designação do Curso	Varchar(100)	2
Abreviatura do Curso	Varchar(50)	2
Acrónimo do Curso	Varchar(15)	1
Regime do Curso	Varchar(15)	2
Ciclo de Estudos do Curso	Varchar(15)	2
Abreviatura do Curso para o Cartão	Varchar(16)	1
Estado do Curso	Varchar(15)	2
DataInício	Smalldatetime	-
DataFim	Smalldatetime	-

Dimensão Aluno

O tipo de dados dos atributos da dimensão Aluno e respectiva forma de gestão das alterações aos atributos pode ser consultado na Tabela 33. Apenas devido ao elevado número de atributos, optou-se por dividir a tabela em duas partes, para facilitar a respectiva consulta.

Tabela 33 – Dimensão Aluno: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave do Aluno	**int**	
Número do Aluno	Int	
Nome do Aluno	Varchar(150)	2
Data de Nascimento	Varchar(15)	1
Número do BI	Varchar(15)	1
Data de Validade do BI	Varchar(15)	1
Local de Emissão do BI	Varchar(50)	1
Distrito – Aulas	Varchar(100)	2
Concelho – Aulas	Varchar(100)	2
Localidade – Aulas	Varchar(50)	2
Código Postal – Aulas	Varchar(15)	2
Morada – Aulas	Varchar(200)	
Distrito – Férias	Varchar(100)	1
Concelho – Férias	Varchar(100)	1
Localidade – Férias	Varchar(50)	1
Código Postal – Férias	Varchar(15)	1

Nome do Atributo	Tipo de Dados	DAL
Morada – Férias	Varchar(150)	1
Nome do Pai	Varchar(150)	1
Nome da Mãe	Varchar(150)	1
Sexo	Varchar(15)	1
Número de Contribuinte	Varchar(15)	2
Data de Entrada	Varchar(15)	1
Nacionalidade	Varchar(50)	1
Naturalidade	Varchar(50)	2
Telefone	Varchar(15)	1
Telemóvel	Varchar(15)	1
E-Mail #1	Varchar(100)	1
E-Mail #2	Varchar(100)	1
Tipo de Inscrição	Varchar(60)	2
Ano Lectivo de Entrada	Varchar(9)	2
DataInício	Smalldatetime	-
DataFim	Smalldatetime	-

Esta dimensão possui algumas hierarquias, nomeadamente, as hierarquias "Residência em férias" e "Residência em aulas", bem como as hierarquias "Inscrição" e "Entrada na Instituição":

Distrito – Férias → Concelho – Férias → Localidade – Férias → Código Postal – Férias
Distrito – Aulas → Concelho – Aulas → Localidade → Aulas → Código Postal – Aulas
Tipo de Inscrição → Nome do Aluno
Ano Lectivo de Entrada → Nome do Aluno

Dimensão Professor

A dimensão Professor é outra dimensão *role playing*, pois representa o Professor Responsável pela Unidade Curricular a que o Aluno efectuou o registo de avaliação. A designação dos atributos presentes na Tabela 34 representa o papel desempenhado pela dimensão no processo de negócio desenvolvido.

Tabela 34 – Dimensão Professor: tabela física

Nome do Atributo	Tipo de Dados	DAL
Chave do Responsável da UC	**Int**	-
Código do Responsável da UC	Varchar(15)	2
Nome do Responsável da UC	Varchar(100)	2
Data de Nascimento do Responsável da UC	Varchar(15)	1
Localidade de Residência do Responsável da UC	Varchar(30)	1
Código Postal do Responsável da UC	Varchar(8)	1
Morada do Responsável da UC	Varchar(200)	1
Número de BI do Responsável da UC	Varchar(15)	-
Data de Validade do BI do Responsável da UC	Varchar(15)	1
Estado Civil do Responsável da UC	Varchar(15)	2
Número de Contribuinte doResponsável da UC	Varchar(15)	2
Bairro Fiscal do Responsável da UC	Varchar(50)	1
Código da Repartição de Finanças do Responsável da UC	Varchar(15)	1
Número de Telefone do Responsável da UC	Varchar(15)	1
Número de Telemóvel do Responsável da UC	Varchar(15)	1

Nome do Atributo	Tipo de Dados	DAL
Número de ADSE do Responsável da UC	Varchar(15)	2
NIB do Responsável da UC	Varchar(21)	2
E-Mail #1 do Responsável da UC	Varchar(100)	1
E-Mail #2 do Responsável da UC	Varchar(100)	1
Dependentes do Responsável da UC	Varchar(15)	2
Sexo do Responsável da UC	Varchar(15)	1
Nome do Pai do Responsável da UC	Varchar(100)	1
Nome da Mãe do Responsável da UC	Varchar(100)	1
UserName do Responsável da UC	Varchar(50)	-
Número do Gabinete do Responsável da UC	Varchar(30)	2
Extensão telefónica do Responsável da UC	Varchar(15)	1
Piso do Gabinete do Responsável da UC	Varchar(15)	1
Área Científica do Responsável da Unidade Curricular	Varchar(30)	2
DataInício	Smalldatetime	-
DataFim	Smalldatetime	-

Dimensão Corpo Docente

A dimensão Corpo Docente é implementada pela tabela ponte que permite o acesso aos dados das tabelas com ela relacionadas. Após o carregamento da tabela na fase de ETL, a ferramenta Analysis Services permite introduzir no modelo dimensional a tabela ponte de forma transparente para o utilizador do Data Warehouse. A chave da dimensão Unidade Curricular foi adicionada no desenho da tabela ponte Corpo Docente de forma a permitir a ligação com integridade referencial entre esta e a tabela de factos pelo atributo "Chave do Corpo Docente". A outra hipótese de solução seria criar mais uma tabela com um único atributo – a Chave do Corpo Docente – para efectuar a ligação.

Tabela 35 – Dimensão Corpo Docente: tabela física

Nome do Atributo	Tipo de Dados
Chave do Corpo Docente	**Int**
Chave do Docente	**int**
Chave do Semestre	**Int**
Chave da Unidade Curricular	**Int**
Factor de Ponderação	Decimal(2,1)

Tabela de factos

A tabela de factos do modelo desenvolvido é caracterizada na Tabela 36. Esta tabela permite identificar a chave primária (PK) da tabela de factos, as chaves estrangeiras (FK), a aditividade e o tipo de dados de cada atributo.

Tabela 36 – Tabela de Factos do modelo desenvolvido

Nome do Atributo	Chave	Aditividade	Tipo de Dados
Chave da Unidade Curricular	**PK, FK**		**Int**
Chave do Aluno	**PK, FK**		**Int**
Chave do Semestre	**PK, FK**		**Int**
Chave do Tipo de Inscrição e Estado de Avaliação	**PK, FK**		**Int**
Chave do Corpo Docente	FK		Int
Chave do Responsável da Unidade Curricular	FK		Int
Chave do Curso	FK		Int
Chave da Área Científica	FK		Int
Chave da Data de Inscrição	FK		Int
Chave da Data de Avaliação	FK		Int
Nota obtida à Unidade Curricular		Semi-aditivo	TinyInt
Créditos ECTS		Semi-aditivo	Decimal(3,1)
Peso da Unidade Curricular na média Final		Semi-aditivo	Decimal(2,1)
Número de Inscrição à Unidade Curricular		Não aditivo	TinyInt
Ano Curricular de Inscrição do Aluno		Não Aditivo	TinyInt

No que respeita à aditividade, como o grão do processo de negócio é muito pequeno (o nível de detalhe é elevado), não existe nenhum facto aditivo, ou seja, que se possa somar para todas as dimensões. No entanto, com a utilização de agregações conseguimos facilmente obter factos aditivos. Se retirarmos do processo de negócio a dimensão "Tipo de Avaliação e Estado de Inscrição", todas as medidas semi-aditivas passam a ser aditivas.

Agregações

A utilização de agregações é uma técnica fundamental nos projectos de Data Warehouse. São uma ferramenta que permite sumarizar a informação em diferentes níveis de granularidade e sob diferentes ângulos de visualização, de forma a melhorar a velocidade de exploração dos dados e a performance do sistema.

As agregações podem ser construídas de duas formas: ou retiramos dimensões do processo de negócio ou então reduzimos a sua dimensão utilizando as hierarquias existentes. As dimensões resultantes deste processo deverão ser versões condensadas da versão orginal existente no processo de negócio com o grão mais atómico. Como não é praticável a construção de todas as agregações que são possíveis de realizar num Data Warehouse, dada a sua quantidade elevada (Kimball & Ross, 2002), devemos tentar identificar quais as agregações mais úteis. Para isso existem duas técnicas importantes:

- Analisar os padrões de utilização dos "consumidores" da informação presente no Data Warehouse – se conseguirmos esta informação, é mais fácil tomar decisões quanto aos dados a agregar. Uma boa ferramenta de análise é a documentação produzida na fase de definição dos requisitos de negócio, onde conseguimos estudar qual a informação mais relevante para o negócio. Outra forma é o registo estatístico de acesso à informação no Data Warehouse;
- Perceber de que forma a informação se encontra estruturada – se uma hierarquia tiver poucos elementos, então provavelmente não será uma boa candidata a agregação pois o ganho não será elevado. Se por outro lado, o número de elementos em cada nível da hierarquia fôr elevado então poderá ser uma boa candidata à agregação.

No modelo desenvolvido, há uma agregação facilmente identificável: a agregação dos resultados obtidos por semestre lectivo. No protótipo foram efectuadas duas agregações com um grão semestral: "Resultados semestrais – por Aluno" e "Resultados semestrais – por Unidade Curricular".

No que diz respeito à construção da agregação, esta pode ser efectuada de duas formas distintas: ou directamente na ferramenta OLAP, utilizando as capacidades de modelação do motor analítico ou então utilizando a ferramenta de ETL no modelo relacional. Para o protótipo desenvolvido optou-se por esta última técnica devido ao

controlo que permite sobre a agregação desenvolvida – todo o processo pode ser controlado pela ferramenta SQL Server Integration Services.

A Fig. 37 apresenta uma agregação semestral construída tendo por base o protótipo desenvolvido. Face ao modelo dimensional de grão atómico, retiraram-se as dimensões Aluno e Data (de Inscrição e de Avaliação).

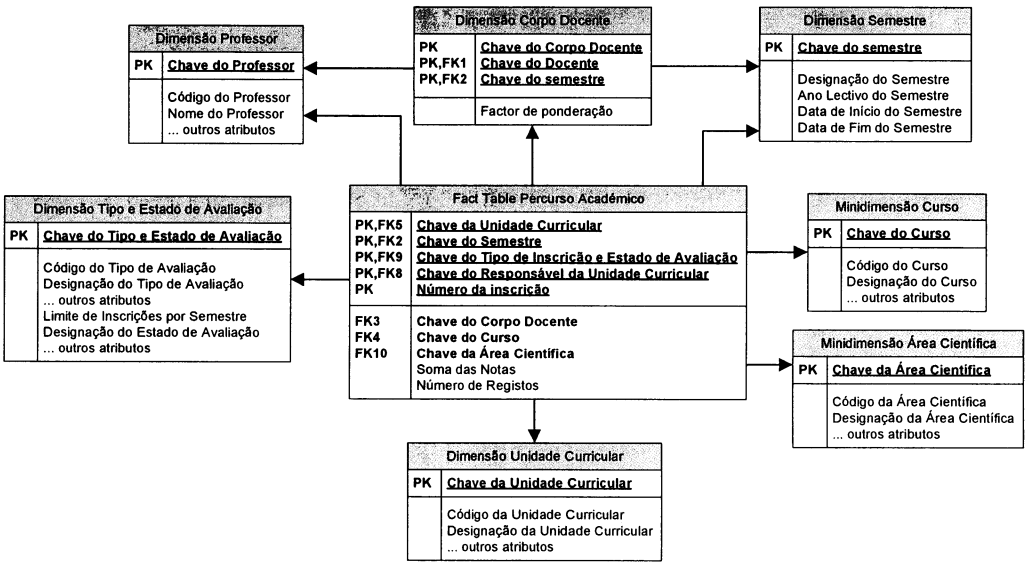


Fig. 37 – Agregação Semestral – por número de inscrição

Esta agregação foi efectuada mantendo a dimensão "Tipo e Estado de Avaliação", para permitir analisar os resultados por tipo de avaliação (Normal, Exame Mensal, etc.) e estado da mesma (Aprovado, Reprovado, etc.).

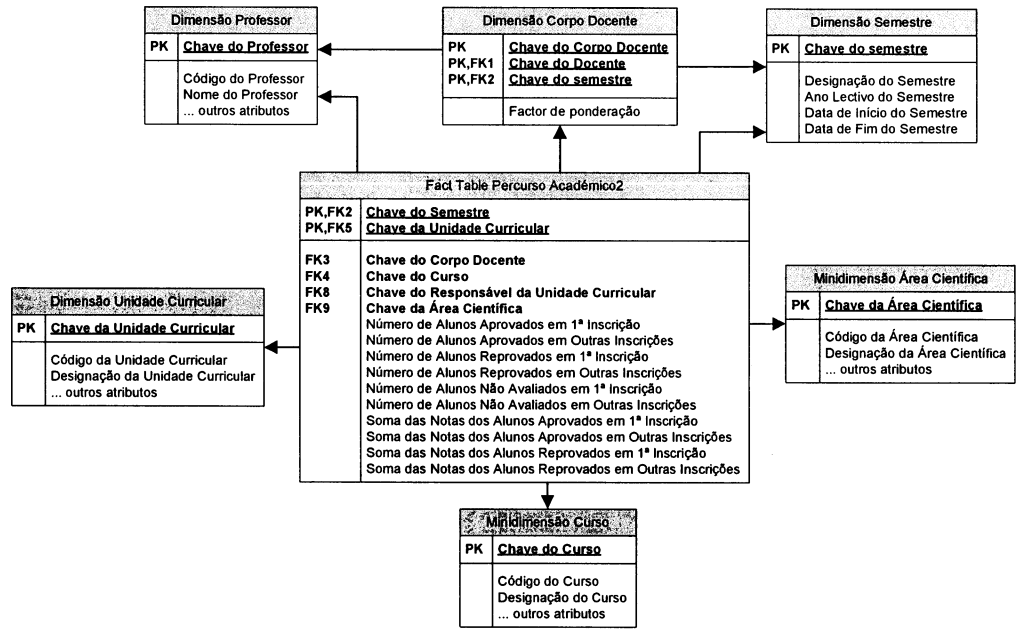


Fig. 38 – Agregação Semestral (detalhe por tipo de Inscrição/avaliação)

Se analisarmos a agregação semestral, já repartida por Tipo de Inscrição, Estado de Avaliação e número de inscrição – convém reparar que na agregação anterior o número de inscrição fazia parte integrante da chave primária, o que mantinha um elevado nível de detalhe no Data Mart e "condicionava" as consultas – constatamos que já nos aproximamos do nível de detalhe pretendido pelos utilizadores visto a informação já se apresentar sumarizada de acordo com as suas necessidades, como iremos verificar no próximo capítulo.

Auditoria e tabela de estágio

Para terminar esta actividade de configuração e desenho físico é necessário abordar um tema muito importante e que permite estabelecer um ponto de partida para a próxima actividade, que incide sobre o processo de ETL: a auditoria ao sistema de armazenamento relacional e a utilização de tabelas de estágio.

A utilização de um sistema de auditoria para monitorizar o carregamento das tabelas físicas que vão armazenar os dados das dimensões e factos permite, por exemplo, controlar a forma como o nosso sistema está ou não bem dimensionado para as necessidades da organização ou detectar e corrigir erros que possam ocorrer durante o processo de ETL.

No protótipo desenvolvido foi adicionado um sistema de auditoria baseado numa tabela de registo da actividade produzida no sistema relacional e através do processo de ETL. Na Tabela 37 podemos ver a estrutura dessa tabela. Cada dimensão, sempre que efectua qualquer processo de alteração na informação que armazena, regista um conjunto de indicadores estatísticos na tabela sobre as operações realizadas.

Tabela 37 – Tabela de Auditoria

Chave_Audit	Chave_AuditPai	NomeTabela	NomePkg	GuidPkg	DataInícioExec	DataFimExec
123			MasterPkg	396eb...	2009-01-06 10:20:36	2009-01-06 10:43:31
124	123	DimAluno	DimAluno	bd394...	2009-01-06 10:20:36	2009-01-06 10:20:36
125	123	DimData	DimData	213ad...	2009-01-06 10:20:36	2009-01-06 10:20:37

Rows_Extracted	Rows_Error	Rows_Inserted	Rows_updated	Rows_AltTipo2	RCntInicio	RCntFim
1208		4	5	2	1204	1210
3673		0	0	0	3673	3673

A tabela de estágio é outra ferramenta importante no processo de gestão física dos dados no Data Warehouse. Estas tabelas permitem efectuar um pré-armazenamento dos dados durante o processo de ETL. Podemos, como foi por exemplo efectuado na construção deste protótipo, construir uma tabela de estágio com a informação actualizada das chaves de todas as tabelas de dimensão que participam num

processo de negócio, de forma a conseguir efectuar a verificação das chaves fictícias para carregar a tabela de factos desde o último carregamento do Data Warehouse.

Área de Estágio e Desenvolvimento

A última actividade do Eixo de Dados consiste na realização do já referido processo "ETL – *Extract, Transform and Load*". Nesta actividade vamos carregar o Data Warehouse com a informação extraída das fontes de dados identificadas nas etapas anteriores. Após extrair a informação, vamos aplicar alguns tratamentos de correcção de erros, validação de informação, normalização de dados, etc. Finalmente, carregamos a informação para as tabelas de factos e de dimensões. A qualidade da informação que se encontra nos sistemas operacionais e outras fontes adicionais vai condicionar directamente esta etapa, pois quanto mais correcta estiver a informação, menor será a necessidade de realizar tarefas como as já anteriormente referidas.

O funcionamento desta actividade dever-se-á reger pela utilização de ferramentas que permitam desenvolver o processo de carregamento da informação de uma forma metódica, ou seja, deve definir-se uma estratégia de execução do processo de carregamento dos dados no Data Warehouse que permita, por um lado, garantir a execução todas as tarefas necessárias de transformação dos dados e, por outro, supervisionar o correcto funcionamento de todo o processo. Para isso deve-se desenvolver um plano de acção que permita a condução dos trabalhos implementados nesta actividade.

No desenvolvimento do protótipo foi definida uma estratégia para o processo ETL, com três níveis de detalhe:

- Controlo de sequência do processo ETL – permite controlar a execução de todo o processo ao nível dos objectos existentes no Data Warehouse que necessitam de carregamento de informação. É também responsável pelo registo de auditoria das tarefas executadas;
- Controlo do processamento das dimensões e tabelas de factos – consiste na realização do processo ETL para cada objecto identificado no nível anterior;
- Implementação de cada tabela de dimensão e de factos – nesta fase é controlada a manipulação da informação, desde os sistemas operacionais até às tabelas de destino.

O nível inicial é que define, não só quando o processo é iniciado mas também quando o mesmo termina, controlando a execução dos outros níveis de detalhe. Se ocorrer algum erro no decorrer deste processo, deverá ser identificado no sistema de auditoria o motivo da falha e, sempre que possível, recolher informação que permita a implementação de um processo de resolução que minimize as consequências da mesma. Este aspecto é especialmente importante no caso do carregamento de grandes volumes de informação, como por exemplo no carregamento inicial do

sistema ou em situações em que a estratégia de actualização dos dados passe pelo carregamento total do Data Warehouse. Se sempre que existir um erro no carregamento de dados, esses registos forem colocados numa área de tratamento de erros e o operador for notificado, a resolução do erro pode passar apenas sobres os registos que produziram a falha, não sendo necessário reiniciar todo o processo de carregamento.

Primeiro nível do processo ETL

Como vimos na actividade anterior, na Tabela 37, foi desenvolvido um sistema de auditoria que vai permitir suportar o processo ETL e, designadamente, os três níveis identificados de desenvolvimento deste.

O primeiro nível regista o início de execução na tabela de auditoria e obtém a data da última execução com sucesso do processo ETL. De seguida, executa o processamento das várias dimensões e tabelas de factos. O processamento das dimensões pode ocorrer em paralelo, para minimizar o tempo de execução. Já com as tabelas de factos optou-se por não efectuar uma execução em paralelo, visto que as necessidades de processamento são maiores.

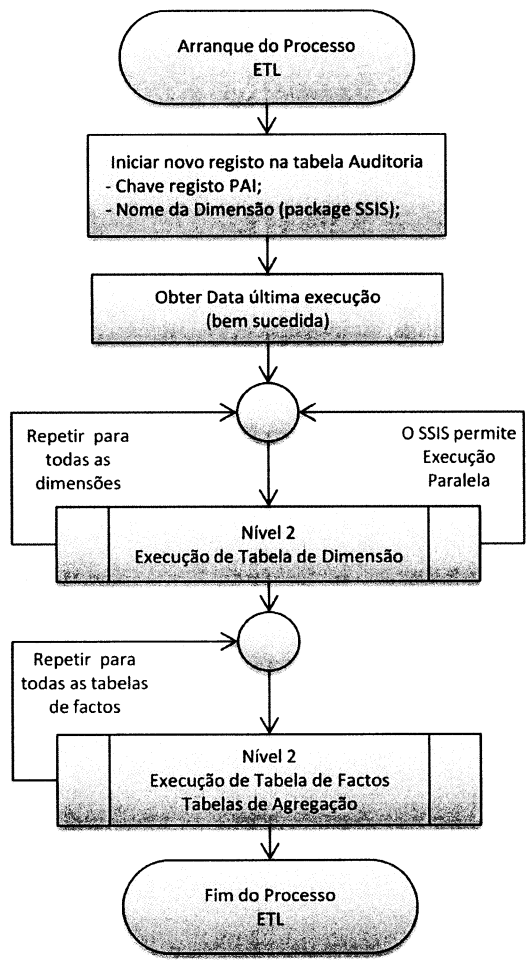


Fig. 39 – Processo ETL – Nível 1

Segundo nível do processo ETL

No segundo nível vamos processar cada uma das várias dimensões do Data Warehouse (ou dos vários Data Mart's, consoante a estratégia de actualização). O processamento das tabelas de factos tem uma sequência algo diferente pelo que tem um diagrama de fluxo próprio.

Os dois primeiros níveis podem ser considerados níveis de controlo: ainda não estamos a trabalhar directamente com os dados mas sim a controlar a execução do processo de ligação entre as várias componentes que serão alvo do ETL.

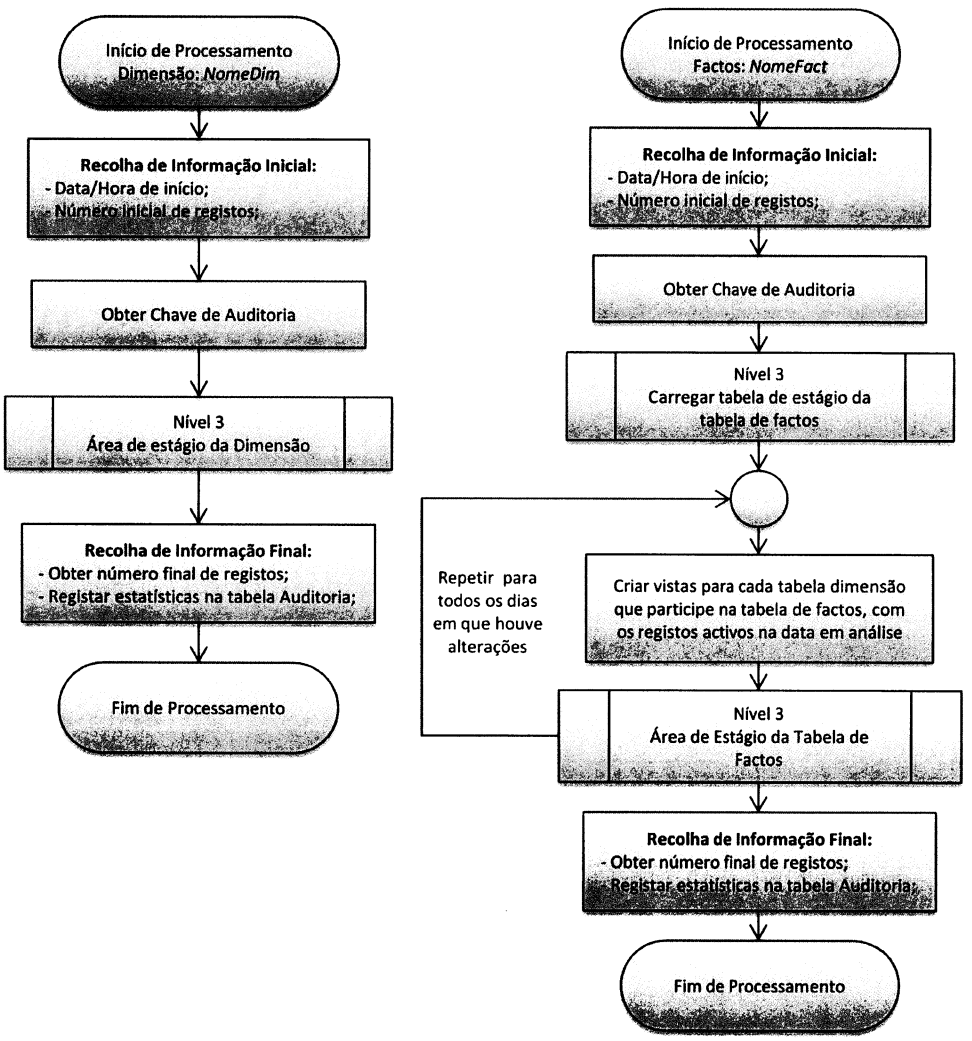


Fig. 40 – Processo ETL – Nível 2

No caso das tabelas de facto, como são utilizadas chaves fictícias para as dimensões, é necessário fazer uma tradução entre estas e as chaves utilizadas pelos sistemas fonte, de forma a carregar a informação na tabela de factos com as ligações referenciais às várias tabelas dimensão. Por outro lado, como as tabelas de dimensão possuem informação activa assim como informação histórica, ou seja, registos "activos" e "inactivos", quando estamos a efectuar o carregamento da informação na

tabela de factos temos de garantir que a ligação à chave fictícia se efectua com o registo da tabela dimensão que se encontra activo na data da ocorrência do facto. Assim, após um carregamento inicial dos registos numa tabela de estágio (onde são efectuadas as transformações necessárias aos dados), devemos identificar todas as datas em que ocorreram os factos a registar no data warehouse e executar um processo repetitivo que cria vistas das tabelas dimensão com a informação activa em cada uma dessas datas. Essas vistas serão depois utilizadas para efectuar a correspondência das chaves fictícias que serão introduzidas na tabela de factos.

Terceiro nível do processo ETL

O terceiro e último nível do processo ETL consiste na extracção dos dados dos sistemas fonte, na transformação dos mesmos, na obtenção de informação estatística a passar aos outros níveis e no preenchimento das várias tabelas de dimensão e de factos.

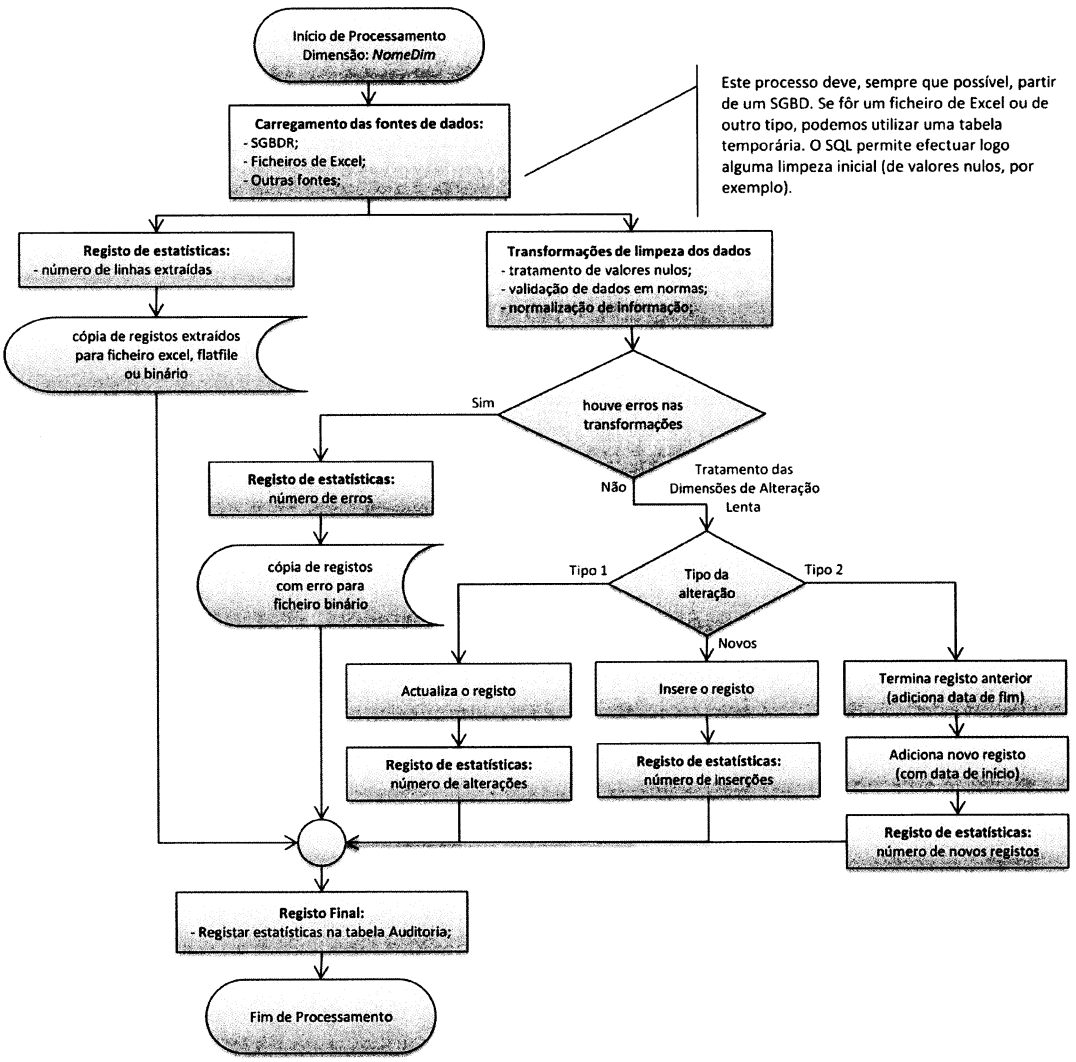


Fig. 41 – Processo ETL – Nível 3 (Dimensões)

Como podemos ver pela Fig. 41, o terceiro nível é o que efectivamente realiza o carregamento dos dados. A existência de erros no processamento produz um ficheiro num formato binário próprio da ferramenta SSIS com os registos que provocaram a falha e que permite a sua posterior análise e reintrodução no sistema de carregamento sem que seja necessário voltar a executar o processo.

No final da execução existem um conjunto de estatísticas que se armazenam na tabela de auditoria e permitem monitorizar e analisar a performance do sistema.

Para o carregamento das tabelas de factos, o processo é semelhante no caso do carregamento da tabela de estágio, sem a utilização do tratamento às dimensões de alteração lenta. No final do carregamento da tabela de estágio, é produzido um *recordset* em memória (um recordset é um objecto utilizado por algumas linguagens de programação e que tem um funcionamento semelhante a uma variável, permitindo armazenar tabelas de dados) com todas as datas em que foram efectuadas alterações no sistema operacional. Essas datas são depois utilizadas no nível 2 por um ciclo que permite construir as views e executar o processo da Fig. 43.

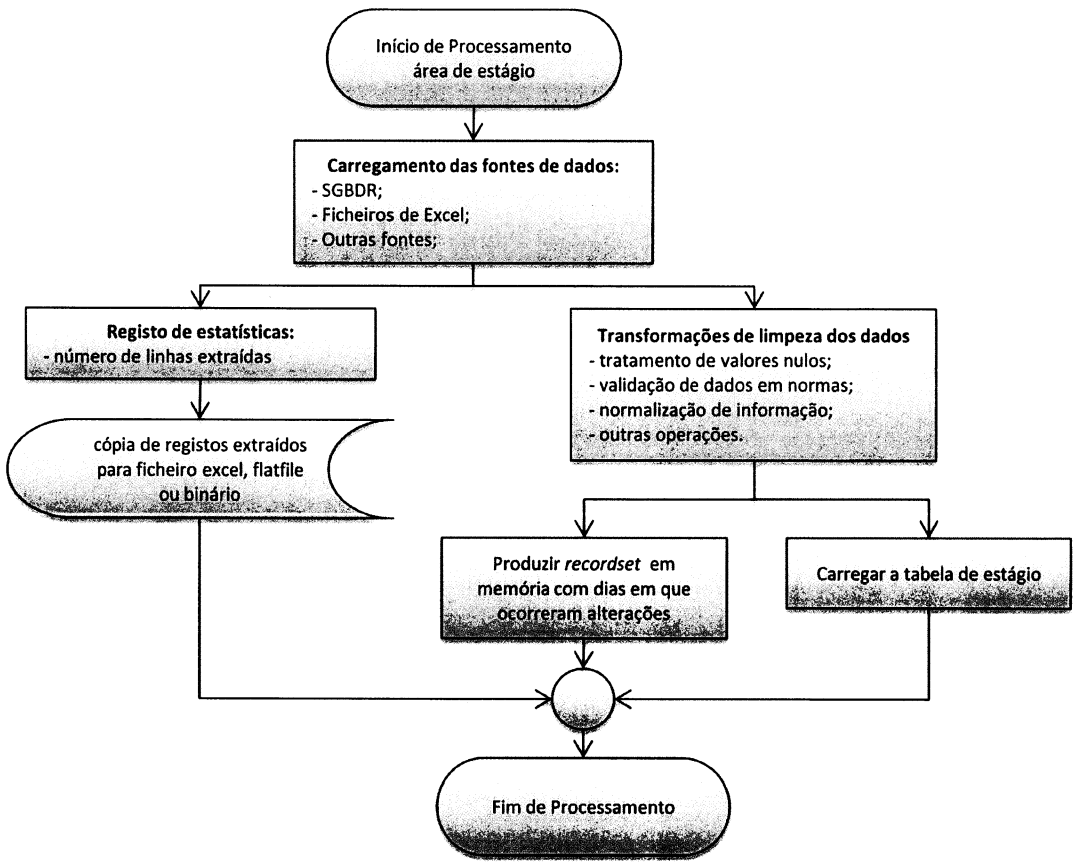


Fig. 42 – Processo ETL - Nível 3 (Área de estágio)

No que diz respeito ao carregamento da tabela de factos, o processo é bastante diferente dos dois anteriores, como podemos verificar na Fig. 42. Esta tarefa tem como objectivo principal obter as chaves fictícias das dimensões que participam no processo

de negócio. Se uma correspondência falhar, ou seja, se não existir na tabela de dimensão a informação sobre o registo que estamos a adicionar à tabela de factos, então temos duas formas de agir: ou inserimos um registo *dummy* ou então falhamos o registo, enviando-o para um ficheiro de erros, que podemos depois analisar e, corrigido o problema, voltar a submeter ao processo para introduzir na tabela de factos.

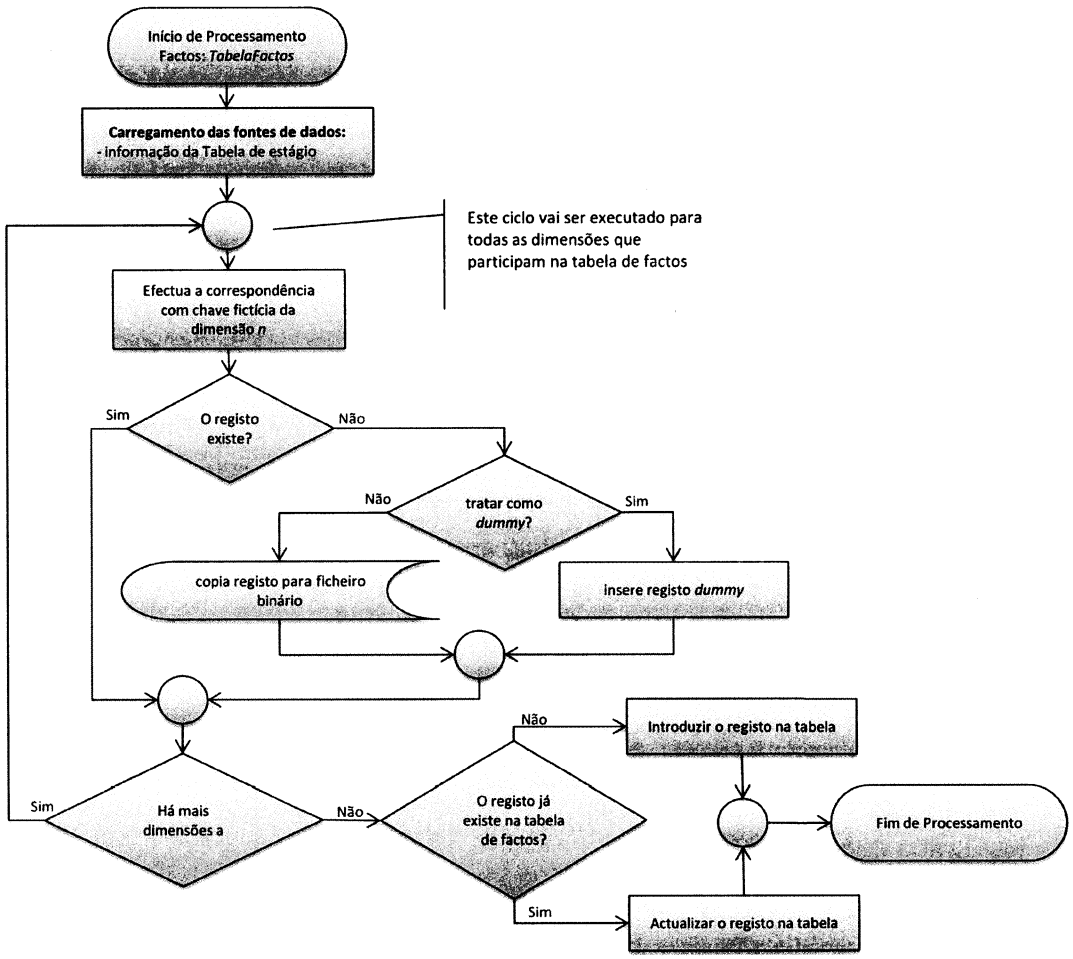


Fig. 43 – Processo ETL - Nível 3 (Tabela de factos)

Frequência de actualização

A frequência de actualização do protótipo difere no tipo de objecto. Assim, as dimensões são actualizadas com uma frequência semanal, enquanto a tabela de factos terá uma frequência de actualização mensal. Isto porque, como o processo actual implementado considera as avaliações efectuadas, à excepção de alguns tipos específicos de avaliação, como por exemplo os exames mensais, a actualização da tabela de factos não tem uma frequência superior ao mês.

Quanto às dimensões, optou-se por uma frequência de actualização semanal porque são mais frequentes as alterações, principalmente nas dimensões Aluno e Professor.

Quando for implementado o próximo processo de negócio, que deverá ser o processo de venda/transacção de produtos e serviços, a frequência de actualização deste processo já passará a ser diária, pois a especificidade do negócio assim o obriga.

Para terminar esta actividade e introduzir a próxima, a ferramenta do pacote de Business Intelligence da Microsoft utilizada para o processo ETL é o SQL Server Integration Services (SSIS) que iremos enquadrar no contexto das ferramentas de BI disponibilizadas pela Microsoft.

No Anexo II podemos encontrar todos os diagramas e scripts efectuados no processo de ETL para o protótipo desenvolvido.

EIXO TECNOLÓGICO

A primeira actividade deste eixo consiste na definição da arquitectura técnica que irá dar suporte ao desenvolvimento do Data Warehouse. A actividade inicia após a definição de requisitos do negócio, precisamente porque só após termos os requisitos definidos conseguimos planear o sistema necessário para garantir os aspectos de performance e fiabilidade necessária para o funcionamento da plataforma.

Em termos de arquitectura técnica do sistema, o modelo dimensional é desenvolvido em primeiro lugar numa base de dados relacional, que representa a solução encontrada no Eixo de Dados, que abordaremos em seguida. Para efectuar o carregamento dos dados nessa base de dados são utilizadas um conjunto de técnicas de recolha e limpeza, que constituem o processo ETL, já abordado no Capítulo 2. Esta base de dados servirá de suporte à base de dados OLAP que irá conter o Data Warehouse e permitirá a sua gestão.

Os sistemas fonte utilizados são, essencialmente, as bases de dados transaccionais do Sistema de Informação da ESTBarreiro/IPS e alguns ficheiros de Excel que contêm informação dispersa pela Organização. Para o acesso à informação no Data Warehouse, existem essencialmente dois grupos de ferramentas:

- As ferramentas analíticas avançadas – são ferramentas que permitem análises avançadas nos dados, como aplicações de data mining para descoberta de conhecimento, ferramentas avançadas de relatório (também conhecidas por *reporting tools*;
- As ferramentas de consulta e relatório – são ferramentas de consulta da informação, como consultas *ad-hoc*, relatórios, análises utilizando folhas de cálculo e tabelas dinâmicas (ou *Pivot Tables*, como também são conhecidas), etc.

No protótipo desenvolvido nesta dissertação, que iremos abordar no sub-capítulo 4.2, as ferramentas utilizadas para a visualização e consulta de informação pertencem essencialmente ao segundo grupo, de ferramentas de consulta também designada

"exploratória" (Kimball, Mundy, & Thornthwaite, 2006). Na Fig. 44, podemos ver um esquema da arquitectura técnica do Data Warehouse desenvolvido. A sigla MDX significa *Multidimensional Expressions* ou expressões multi-dimensionais, em português, e consiste numa linguagem de consulta para bases de dados OLAP, assim como o SQL o é para bases de dados transaccionais.

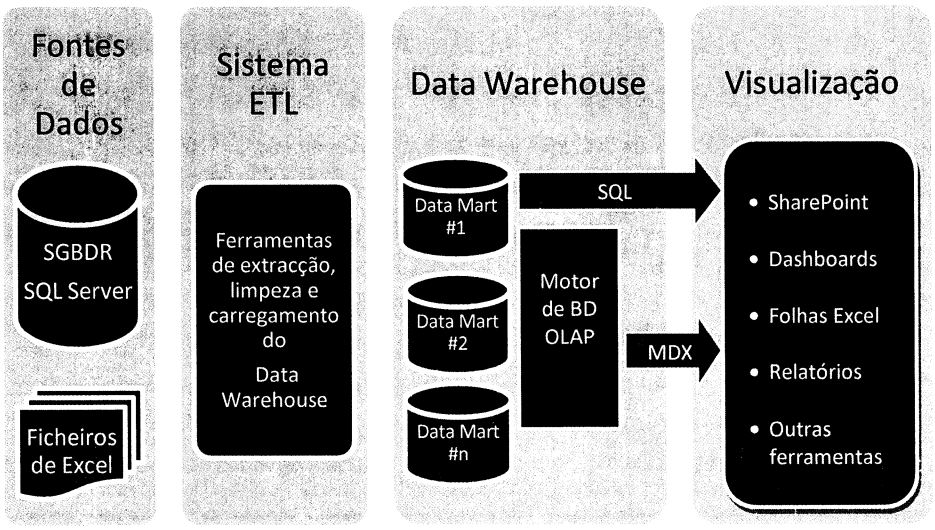


Fig. 44 – Arquitectura técnica do modelo desenvolvido

A segunda actividade, selecção e instalação do produto, consiste em escolher, de entre as soluções existentes no mercado, qual a plataforma de desenvolvimento a utilizar. Como já foi atrás referido, o protótipo foi implementado numa solução Microsoft, o Microsoft SQL Server 2008. A edição utilizada foi a empresarial, pois só esta possui as funcionalidades necessárias ao desenvolvimento do Data Warehouse (a versão *Developer* permite as mesmas funcionalidades que a empresarial, mas é destinada apenas a ambientes de teste).

Esta ferramenta possui duas componentes fundamentais no desenvolvimento e gestão do Data Warehouse: uma ferramenta integradora denominada *Business Intelligence Developer Studio* ou BIDS, que permite construir a solução de Data Warehouse, e a ferramenta *SQL Server Management Studio* ou SSMS, que permite efectuar a gestão da solução desenvolvida. Em termos de ferramentas para desenvolvimento da plataforma BI, onde se integra o Data Warehouse, a ferramenta SQL Server 2008 divide-se em três componentes:

- Integration Services (SSIS) – é a ferramenta que permite integrar a informação proveniente das mais diversas fontes de dados e aplicar as tarefas necessárias de limpeza e transformação dos dados – o processo ETL;
- Analysis Services (SSAS) – consiste na base de dados analítica com suporte ao desenvolvimento dimensional. É a ferramenta que permite construir o Data Warehouse;

- Reporting Services (SSRS) – os serviços de relatório permitem adicionar uma camada de apresentação de informação mediante a construção de relatórios.

A Fig. 45 apresenta a forma como estes componentes interagem no pacote de ferramentas de *Bussiness Intelligence* (BI) da Microsoft.

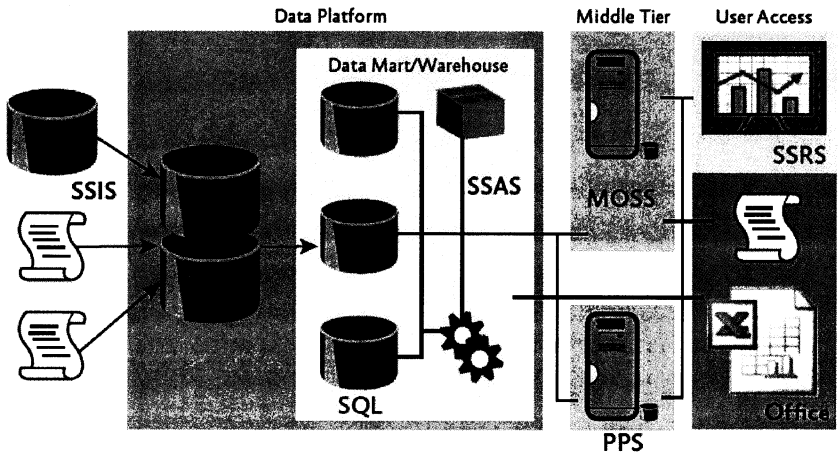


Fig. 45 – Componentes de BI do Microsoft SQL Server 2008

Fonte: (Misner, 2009)

A sigla PPS na Fig. 45 é a abreviatura de *Performance Point Server*, uma ferramenta analítica da Microsoft, que não será objecto de estudo nesta dissertação.

EIXO ANALÍTICO

No que respeita ao eixo analítico, este comporta também duas actividades, a especificação das aplicações analíticas e o seu respectivo desenvolvimento. A primeira actividade consiste na definição de um conjunto de aplicações que permitam satisfazer algumas das necessidades analíticas dos utilizadores, como reports ou dashboards pré-construídos. Estas necessidades são extraídas da actividade anterior, definição de requisitos de negócio. A actividade seguinte consiste no desenvolvimento das aplicações definidas anteriormente. É importante ter em atenção as normas e standards utilizados nos metadados. É aqui que transparece a utilidade do desenvolvimento das aplicações em simultâneo com os outros dois eixos: assim que possamos começar a testar as aplicações nos dados entretanto disponibilizados conseguimos efectuar correcções nos mesmos e perceber as necessidades de afinações nas aplicações.

No que respeita ao Data Warehouse da ESTBarreiro/IPS, o protótipo desenvolvido utiliza para construção das aplicações de análise e apresentação de resultados ao utilizador, duas ferramentas da empresa Microsoft: para construir as aplicações utiliza-se o Microsoft Office Excel 2007 e as suas potencialidades em termos de tabelas dinâmicas (ou *Pivot Tables*, em inglês) para construir relatórios customizados que depois são colocados em Dashboards publicados no portal Microsoft Office SharePoint

Server 2007, disponibilizados a todos aqueles que necessitem de os consultar ou manipular. Esta decisão permite uma série de vantagens para os utilizadores, como iremos abordar no capítulo 5, apresentação de resultados.

IMPLEMENTAÇÃO

Os três eixos de desenvolvimento abordados anteriormente vão convergir nesta actividade, que consiste na implementação do Data Warehouse para a organização, ou seja, é a fase em que os utilizadores vão poder ter contacto com o resultado do desenvolvimento efectuado nas actividades anteriores.

Um aspecto muito importante desta fase é a formação dos utilizadores na exploração dos recursos do Data Warehouse e na forma de organização da informação. No capítulo 5, apresentação de resultados, vamos abordar esta relação dos utilizadores com o Data Warehouse da ESTBarreiro/IPS.

MANUTENÇÃO E CRESCIMENTO

Esta é a última actividade do ciclo de vida dimensional e consiste na manutenção do Data Warehouse e suporte aos seus utilizadores. Este suporte deve acontecer de forma proactiva nas semanas seguintes à formação inicial para permitir o esclarecimento de dúvidas iniciais que possam surgir.

O acompanhamento dos utilizadores é especialmente importante pois permite resolver os problemas de forma proactiva, em vez de esperar os problemas colocados pelos utilizadores, reagindo aos mesmos. É importante não só conquistar mas também manter a confiança do utilizador na informação que consulta, sob risco de o mesmo deixar de utilizar o sistema.

Por outro lado, ao passar para um ambiente de produção devemos ter em consideração os aspectos de performance e estabilidade do sistema, monitorizando o funcionamento do mesmo. Deve existir uma equipa dedicada ao suporte técnico ao sistema, que permita responder às necessidades do mesmo.

Por último, esta actividade também é o ponto de partida para o crescimento do Data Warehouse. No entanto, à semelhança de qualquer outro projecto, para que possa crescer é necessário que exista aceitação do mesmo por parte da organização e que surjam novas necessidades. Devemos por isso ter especial cuidado com a imagem deixada pelo projecto e garantir que o que está implementado serve de montra para o desenvolvimento futuro.

5. APRESENTAÇÃO DE RESULTADOS

5.1. CASO DE ESTUDO

Para permitir testar o modelo desenvolvido e respectivo protótipo, foi implementado um portal para acesso ao Data Warehouse da ESTBarreiro/IPS, utilizando a ferramenta Microsoft Office Sharepoint Server 2007 (MOSS 2007) e que está disponível apenas para os utilizadores que participaram nas reuniões de desenvolvimento, permitindo o acesso a toda a informação disponível no Data Warehouse da ESTBarreiro/IPS. Esta ferramenta permite a integração com as restantes ferramentas de BI da Microsoft, como é possível verificar na Fig. 46.

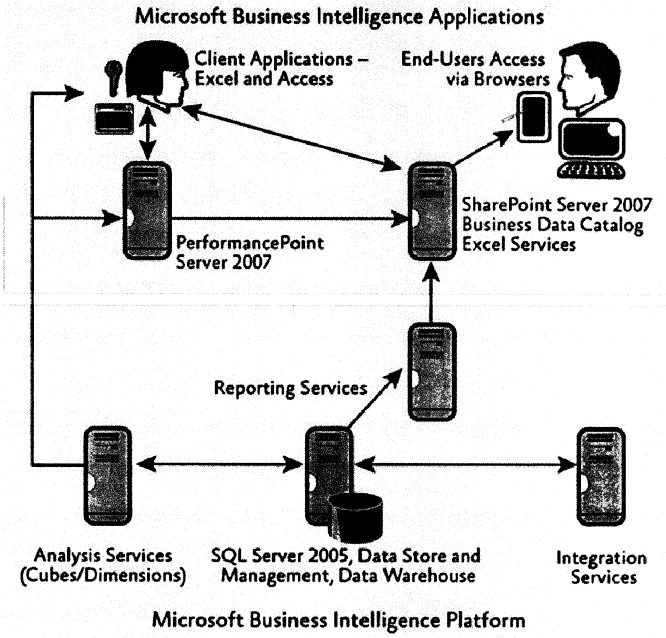


Fig. 46 – Integração do Sharepoint no BI da Microsoft

(Curry & English, 2009)

O acesso à informação é efectuado através da ferramenta Microsoft Office Excel 2007, mas também podem ser utilizadas outras ferramentas como por exemplo a base de dados Microsoft Access ou outras aplicações não Microsoft.

Uma componente bastante importante no Sharepoint 2007 é o motor de serviços de Excel ou *Excel Services*. Esta ferramenta permite amplificar as funcionalidades do Excel 2007 na medida em que disponibiliza a partilha de documentos de Excel utilizando capacidades de controlo de versões e acesso partilhado, tirando partido de um interface de utilizador disponível através de páginas Web (também designada por Web UI). Este interface Web permite ao Sharepoint apresentar a informação das folhas de Excel no portal, partilhando-as com a comunidade. Os serviços de Excel permitem também que a informação a disponibilizar no portal possua um controlo de

segurança aplicado ao objecto, ou seja, podemos definir qual o objecto ou objectos de uma folha de Excel que queremos apresentar no portal, como por exemplo uma tabela de dados formatada, um gráfico ou mesmo um indicador de performance, também designado por KPI (sigla para *Key Performance Indicator*).

A ferramenta Excel 2007 possui integração directa com ferramentas de BI da Microsoft, nomeadamente o Analysis Services, como podemos ver na Fig. 47.

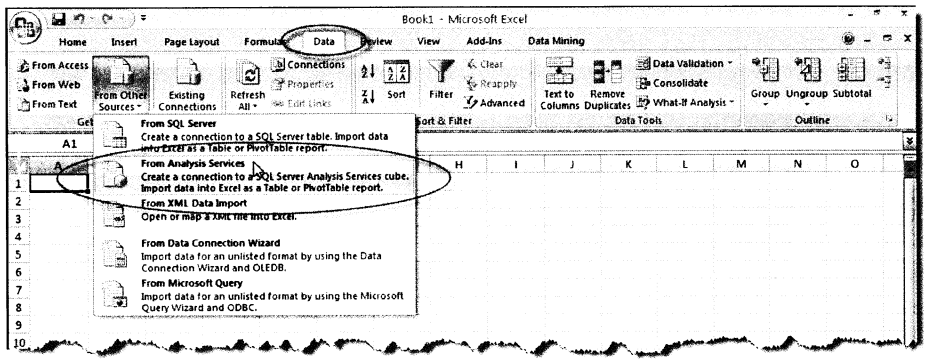


Fig. 47 – Ligação do Excel ao Analysis Services

No caso de estudo, as análises de dados efectuadas utilizam o Excel, não só pelas capacidades que oferece em termos de ferramenta de análise de dados integrada no pacote de BI da Microsoft mas também visto ser uma ferramenta amplamente divulgada na instituição e com a qual os utilizadores possuem experiência de utilização.

A página inicial do portal Sharepoint da ESTBarreiro/IPS é apresentada na Fig. 48.

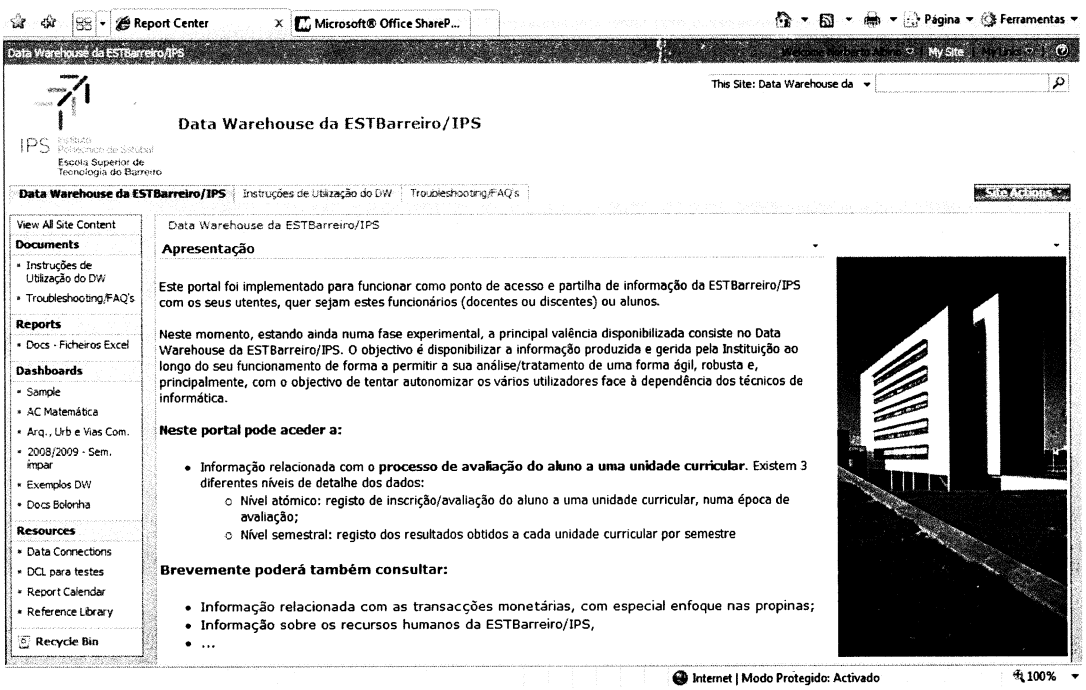
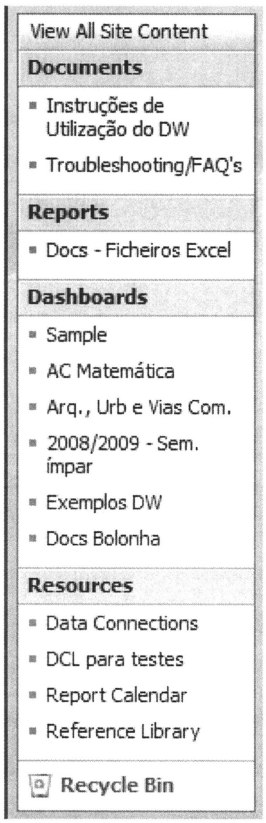


Fig. 48 – Portal Sharepoint da ESTBarreiro/IPS

O portal vai servir como repositório central de partilha de informação, que será "trabalhada" no Excel 2007 e apresentada em dashboards e páginas Web do portal.

O objectivo consiste em permitir aos utilizadores da organização, sem conhecimentos técnicos na área da informática, a possibilidade de realizarem as suas consultas de informação e análises de dados sem recurso ao auxílio do pessoal técnico do Núcleo de informática da ESTBarreiro/IPS.



Assim, para aceder à informação constante no Data Warehouse, os utilizadores possuem uma barra de acesso vertical na esquerda do portal onde podem efectuar as seguintes operações:

- Consultar documentos com as instruções de funcionamento do portal Sharepoint, forma de utilização da informação disponível no Data Warehouse, construção de recursos em Excel e uma área dedicada a perguntas frequentes e erros detectados;
- Aceder a uma área de ficheiros de Excel disponibilizados pelos utilizadores do Data Warehouse;
- Consultar os vários dashboards com informação construída e disponível pelos vários utilizadores;
- Aceder às conexões para os vários níveis de agregação disponíveis no Data Warehouse e diferentes perspectivas criadas no Analysis Services. Estas conexões permitem aceder ao Excel e efectuar a conexão ao Analysis Services de forma automática, sem ter de definir manualmente a conexão como mostrava a Fig. 47.

Fig. 49 – Barra de navegação do Portal

A criação de um ficheiro de Excel com ligação ao Data Warehouse é realizada de forma bastante simples: basta aceder à opção "Data Connections", escolher a ligação desejada e automaticamente surge uma folha de Excel com uma tabela dinâmica pronta a trabalhar com a informação disponibilizada pelo Data Warehouse.

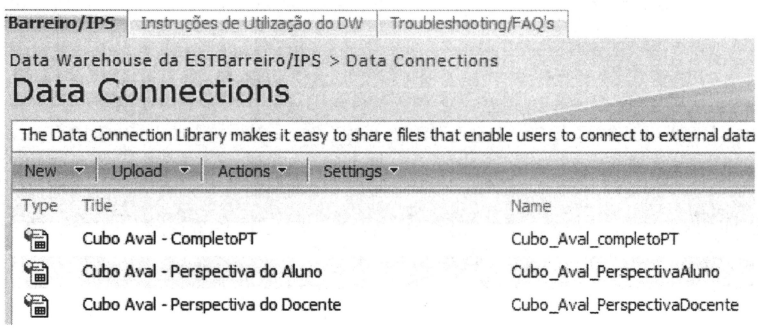


Fig. 50 – Ligações automáticas ao Data Warehouse

Uma vez na folha de Excel, o utilizador pode tratar a informação da forma que mais lhe convier, efectuando as análises de acordo com as suas necessidades e sem depender do Núcleo de Informática da ESTBarreiro/IPS. A Fig. 51 apresenta a folha de Excel pronta a ser utilizada pelo utilizador na análise da informação.

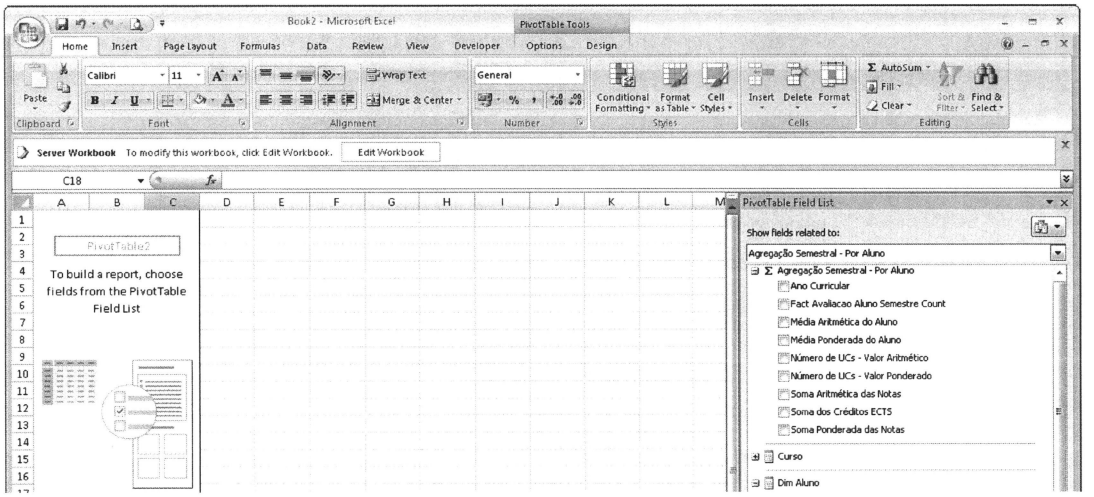


Fig. 51 – Livro de Excel com conexão ao Data Warehouse

A formação inicial na utilização do Data Warehouse inclui, para além da utilização do portal Sharepoint e forma como os dados se encontram estruturados, formação em Excel com dois níveis de detalhe: uma formação genérica na utilização da folha de cálculo e uma formação específica em tabelas dinâmicas e acesso a fontes de dados externas. Utilizando a informação disponível, os utilizadores podem construir, por exemplo, tabelas como a da Fig. 52, que permite apresentar as médias de aprovação por curso e ano lectivo.

	A	B	C	D	E
1	Tipo de inscrição e estado de Avaliação	Aprovado			
2	Estado do Curso	OPEN			
3	Regime do Curso	Diurno			
4					
5		Column Labels			
6	Row Labels	⊞2006/2007	⊞2007/2008	⊞2008/2009	Grand Total
7	CIVD	12,67	12,45	12,39	12,49
8	CRD	13,15	12,19	12,62	12,88
9	GCD		12,39	12,47	12,44
10	QMD			12,29	12,29
11	Grand Total	12,72	12,44	12,40	12,51
12					

Fig. 52 – Exemplo de análise de dados em Excel

Desta tabela podemos criar outros recursos, como o da Fig. 53, que permite apresentar a informação de uma forma gráfica ao utilizador. Depois de criados os recursos desejados na tabela de Excel, podemos disponibilizá-los no portal Sharepoint para consulta e/ou alteração por terceiros.

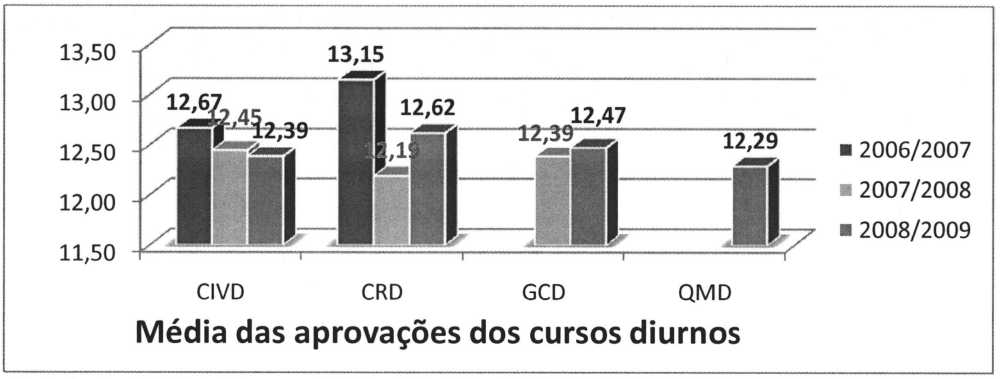


Fig. 53 – Gráfico exemplo de análise de dados em Excel

Se não existisse a ferramenta Data Warehouse na organização para efectuar estas análises, as mesmas teriam de ser efectuadas recorrendo ao sistema operacional, o que levaria à necessidade de produção de código (normalmente SQL) para obter a informação desejada. O código produzido seria, por exemplo, o seguinte:

```
SELECT AnoLectivo, Acronimo, COUNT(*) [Número de Registos], SUM(Nota) [Soma
das notas],
       CAST(SUM(Nota) as decimal(6,0))/CAST(COUNT(*) as decimal(6,0)) [Média]
FROM Curso C, InscricaoDisciplina ID, Disciplina D, AnoLectivo AL
WHERE ID.CodDisciplina = D.CodDisciplina AND
       D.CodCurso = C.CodCurso AND D.Regime = C.Regime AND D.Ciclo = C.Ciclo
       AND ID.CodAnoLectivo = AL.CodAnoLectivo
       AND (C.Estado = 'OPEN') AND ID.Status='Aprovado'
GROUP BY Acronimo,AnoLectivo
ORDER BY AnoLectivo
```

Como podemos verificar, só seria possível ao utilizador efectuar a análise desejada se tivesse conhecimentos avançados da linguagem SQL ou então recorrendo ao auxílio de pessoal técnico que lhe produzisse o código.

Para publicar o ficheiro de Excel no portal podemos gravá-lo do Excel directamente para o portal ou então podemos carregá-lo a partir do portal.

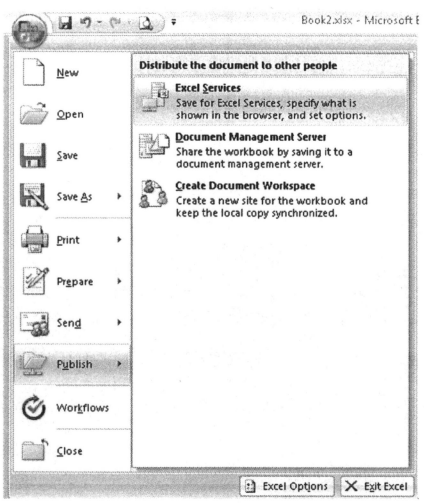


Fig. 54 – Gravação da folha de Excel nos serviços de Excel

No seguimento do exemplo anterior, vamos analisar os passos necessários para publicar o ficheiro no portal a partir do próprio ficheiro de Excel. Em primeiro lugar deverá ser seleccionado o botão Office e escolhida a opção "Publish → Excel Services", como mostra a Fig. 54. Depois, no menu de gravação, deve escolher-se a opção "Excel Services Options" e identificar qual o objecto que se quer publicar no portal Sharepoint. A gravação fica concluída quando identificamos o caminho do servidor (que neste caso seria http://datawarehouse.estbarreiro.ips.pt) e escolhemos o repositório de gravação no mesmo, que deverá corresponder a uma área de ficheiros própria, designada no Sharepoint por DCL – *Data Connection Library.*

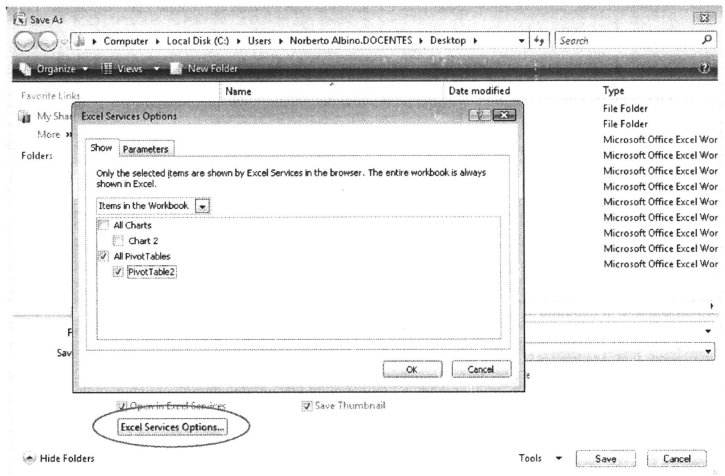


Fig. 55 – Publicação do objecto de Excel no portal

Para publicar o objecto de Excel num dashboard ou página Web do portal, basta seleccionar o dashboard da barra de opções do portal apresentada na Fig. 49 e de seguida inserir o objecto de Excel pretendido num dos vários locais (placeholders) que o dashboard permitir. A Fig. 56 apresenta os exemplos anteriores num dashboard criado para o efeito, "Exemplos dissertação".

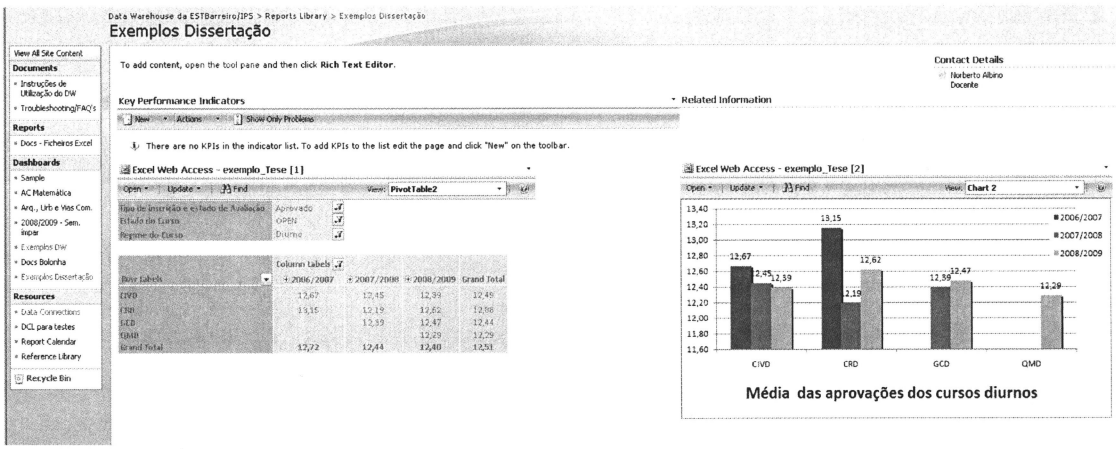


Fig. 56 – Apresentação dos resultados no portal

5.2. ANÁLISES EFECTUADAS

No que diz respeito às análises efectuadas pelos utilizadores do Data Warehouse da ESTBarreiro/IPS, existem duas áreas que claramente mereceram uma atenção especial nos primeiros testes de funcionamento do sistema: a ficha estatística do final de semestre com os resultados de cada Unidade Curricular e os relatórios estatísticos com os "grandes números" da ESTBarreiro/IPS para a elaboração do relatório relativo ao processo de Bolonha.

A ficha estatística de final de semestre (também designada internamente na instituição por modelo 46) é parte integrante do relatório da Unidade Curricular e é uma tarefa que deve ser desempenhada pelo responsável da mesma. Nesta devem constar os resultados obtidos pelos alunos na avaliação final. Deve ser efectuada uma distinção entre os alunos matriculados pela 1ª vez e os alunos com mais do que uma inscrição. A Fig. 57 apresenta um exemplo do modelo disponibilizado no portal para permitir a análise a todas as Unidades Curriculares de uma forma centralizada.

	A	B	C
1			
2	Abreviatura do Curso	Engenharia Civil - Diurno	
3	Calendário Escolar	2008/2009	
4	Designação da Unidade Curricular	Informática e Programação	
5			
6	**Values**		
7	Número de Alunos Aprovados em 1ª Inscrição	10	
8	Número de Alunos Aprovados noutras Inscrições	4	
9	Número de Alunos Reprovados em 1ª Inscrição	5	
10	Número de Alunos Reprovados noutras Inscrições	2	
11	Número de Alunos Não Avaliados em 1ª Inscrição	49	
12	Número de Alunos Não Avaliados noutras Inscrições	9	
13	Média Global das Aprovações	12,29	
14	Média Global das Reprovações	7,86	
15	Média Global dos Avaliados	10,81	
16	Média dos Aprovados em 1ª Inscrição	12,40	
17	Média dos Reprovados em 1ª Inscrição	8,00	
18			

Fig. 57 – Modelo 46: Ficha estatística da Unidade Curricular

Com este modelo torna-se possível disponibilizar a análise dos resultados obtidos nas Unidades Curriculares de uma forma normalizada a todos os membros da comunidade académica que necessitem de aceder à informação.

Por outro lado, é possível também efectuar outras análises relacionadas com estes indicadores, como por exemplo, verificar a evolução dos resultados de uma determinada unidade curricular ao longo do tempo, como podemos verificar na Fig. 58.

	A	B	C	D	E	F
1						
2	Abreviatura do Curso	Engenharia Civil - Regime Diurno (Antigo)				
3	Designação da Unidade Curricular	Geologia				
4						
5		**Column Labels**				
6	**Values**	**⊞2002/2003**	**⊞2003/2004**	**⊞2004/2005**	**⊞2005/2006**	
7	Número de Alunos Aprovados em 1ª Inscrição	25	23	24	4	
8	Número de Alunos Aprovados noutras Inscrições	12	14	11	7	
9	Número de Alunos Reprovados em 1ª Inscrição	16	11	11	4	
10	Número de Alunos Reprovados noutras Inscrições	5	7	10	8	
11	Número de Alunos Não Avaliados em 1ª Inscrição	13	14	18	13	
12	Número de Alunos Não Avaliados noutras Inscrições	9	15	11	17	
13	Média Global das Aprovações	11,76	11,78	11,29	11,36	
14	Média Global das Reprovações	7,38	7,00	6,95	6,50	
15	Média Global dos Avaliados	10,17	10,22	9,66	8,83	
16	Média dos Aprovados em 1ª Inscrição	11,76	12,13	11,25	11,00	
17	Média dos Reprovados em 1ª Inscrição	7,50	7,09	7,36	5,50	
18						

Fig. 58 – Evolução dos resultados por Unidade Curricular

Quanto aos quadros estatísticos para o relatório de Bolonha, estas análises estão mais direccionadas para a visão global da escola, com valores da evolução ao longo do funcionamento da mesma. A Tabela 59 apresenta um exemplo de um quadro com a evolução dos alunos inscritos na ESTBarreiro/IPS ao longo da sua existência.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2	**Row Labels**	**⊞1999/2000**	**⊞2000/2001**	**⊞2001/2002**	**⊞2002/2003**	**⊞2003/2004**	**⊞2004/2005**	**⊞2005/2006**	**⊞2006/2007**	**⊞2007/2008**	**⊞2008/2009**	**⊞2009/2010**	**Grand Total**
3	**1999/2000**	60	48	43	42	34	33	21	15	6	2	3	60
4	**2000/2001**		57	52	45	42	40	35	27	20	10	13	57
5	**2001/2002**			56	49	43	40	34	25	15	10	14	56
6	**2002/2003**				61	54	52	46	39	20	17	15	61
7	**2003/2004**					87	75	63	54	37	28	27	87
8	**2004/2005**						94	72	67	58	49	42	94
9	**2005/2006**							67	58	48	38	31	68
10	**2006/2007**								180	150	115	102	182
11	**2007/2008**									225	181	140	225
12	**2008/2009**										231	170	231
13	**2009/2010**											168	168
14	**Grand Total**	**60**	**105**	**151**	**197**	**260**	**334**	**338**	**465**	**579**	**681**	**725**	**1289**

Fig. 59 – Distribuição de inscrições por ano lectivo

Neste quadro conseguimos observar, por exemplo que a ESTBarreiro/IPS iniciou a sua actividade no ano lectivo 1999/2000 com 60 alunos, que actualmente tem 725 alunos activos e que, desses alunos presentemente activos, três efectuaram a sua primeira matrícula na ESTBarreiro/IPS no ano de início de actividade.

6. CONCLUSÃO E ESTUDOS FUTUROS

O projecto desta dissertação de mestrado permitiu aprofundar os conhecimentos das ferramentas analíticas de bases de dados e mais concretamente da solução Data Warehouse.

Dentro da solução de Data Warehouse foram analisadas as duas arquitecturas mais utilizadas, a arquitectura BUS com base no modelo dimensional e o Data Warehouse empresarial, tendo a decisão de implementação recaído pela primeira, visto ter vários aspectos que lhe conferem vantagem, e de onde se destacam dois que foram amplamente confirmados ao longo do desenvolvimento desta dissertação: a possibilidade de construir o Data Warehouse de uma forma faseada, mediante a utilização de Data Marts conformes, e a facilidade de utilização e simplicidade de compreensão do modelo por parte dos utilizadores não técnicos, confirmada com as funcionalidades descritas no capítulo 5, que permitem uma total autonomia dos utilizadores face ao departamento de informática da organização.

A utilização do desenvolvimento dimensional do ciclo de vida da organização é uma ferramenta que facilita bastante o desenvolvimento teórico e a implementação técnica de todo o projecto. A divisão por actividades ou etapas permite quase que um "guião" para a construção de um Data Warehouse.

Com a evolução rápida das organizações que hoje se verifica e com as alterações em termos das disponibilidades e necessidades de informação, a fase de definição de requisitos de negócio deve ser revisitada sempre que decidimos abordar um novo processo de negócio.

Outra conclusão está relacionada com a qualidade dos dados: a implementação de um projecto de Data Warehouse permite, por si só, aumentar a qualidade dos dados disponíveis. Por um lado porque ao termos de analisar a informação existente nos sistemas operacionais isso permite-nos identificar erros e, na maior parte das situações, tomar medidas para os solucionar (por vezes trata-se de informação que já não conseguimos obter, como no caso de antigos alunos em que a escola lhes tenha perdido o contacto, por exemplo). Por outro lado, ao envolver os actores no projecto de desenvolvimento e ao lhes apresentar vários conjuntos de informação (para efeitos de teste do modelo, da verificação da necessidade de incluir esses mesmos dados ou, por exemplo, de normalizações a efectuar) torna-se possível encontrar informação incorrecta que, pelo conhecimento que têm da organização, por vezes só estes conseguem detectar.

Foi assim confirmada a importância que tem o envolvimento dos utilizadores no desenvolvimento do projecto, fazendo com que contribuam com a sua experiência acumulada para a melhoria da solução proposta e que sintam que também possuem a sua quota-parte de responsabilidade de uma solução bem desenvolvida. Este último

aspecto é crucial para o bom funcionamento do sistema e para a manutenção do mesmo: ao fazer parte da solução, os utilizadores tornam-se proactivos na contribuição para a melhoria do sistema.

A questão da disponibilidade (ou falta desta) e a qualidade dos dados representou mesmo uma das maiores dificuldades na implementação do projecto de Data Warehouse: é possível detectar as deficiências de armazenamento de informação do sistema operacional, sendo este um dos aspectos a abordar em termos de trabalho futuro.

A formação inicial dos utilizadores antes de lhes ser fornecido o acesso ao Data Warehouse foi outro aspecto que se revelou fundamental, pois permitiu "formatar" o conhecimento existente de alguns procedimentos de acesso aos dados, e também os conhecimentos técnicos de algumas ferramentas, principalmente o Excel. Esta formação deverá ter um acompanhamento com formações regulares que permitam manter o nível de actualização com as ferramentas disponíveis e o refrescamento das funcionalidades base a utilizar na exploração do Data Warehouse.

Finalmente, quando o projecto entra em funcionamento e se começa a receber a resposta dos utilizadores, é possível perceber a real necessidade que as organizações possuem em usufruir de um sistema analítico: a facilidade com que o modelo dimensional permite a consulta e manipulação em termos de análise de grandes volumes de informação amplia as limitações do modelo relacional e respectiva linguagem de consulta (SQL) quando utilizados para analisar informação. Assim, podemos concluir que o Data Warehouse é a resposta que permite capacitar as organizações em termos analíticos, permitindo o acesso à informação pelos utilizadores directamente, sem requerer constantemente auxílio dos técnicos de informática. Este aspecto também é importante porque garante aos utilizadores a possibilidade de efectuar as consultas e análises que pretendem, por um lado utilizando o seu conhecimento da organização (sem que precisem de um elemento intermediário para lhe fornecer os dados e assim limitar a exploração dos mesmos), e por outro, beneficiando da disposição dos dados orientada à forma de fazer negócio da organização.

Esta dissertação de Mestrado permitiu introduzir a ferramenta Data Warehouse numa Instituição de Ensino Superior que até então não havia tido qualquer contacto com uma ferramenta do género. Foi possível apresentar aos seus principais quadros de gestão as potencialidades de uma ferramenta analítica, mostrando as vantagens da sua utilização e o que pode fazer em termos da qualidade adicionada às diferentes análises efectuadas na organização.

6.1. Estudos Futuros

Como trabalho futuro, um dos primeiros objectivos passa pelo desenvolvimento de novos Data Marts para modelar os processos de negócio identificados na definição de requisitos da organização e pela consolidação do trabalho já desenvolvido no Data Warehouse da ESTBarreiro/IPS. Por outro lado, a utilização do Excel como ferramenta de análise de informação revelou-se uma solução com bastante potencial, que importa explorar. O crescimento constante da ferramenta, com o lançamento da nova versão previsto para o início de 2010, também obriga a uma necessidade de constante evolução das soluções nela desenvolvidas. Assim, outro dos objectivos para trabalho futuro passa pelo investimento na produção de novos modelos (ou *templates*) de análise baseados nas capacidades disponibilizadas pelo Excel, como por exemplo a utilização dos indicadores de performance (KPI's) ou o desenvolvimento de dashboards customizados. Existem alguns pacotes de funcionalidades adicionais (*add-ins)* que permitem dotar o Excel 2007 com capacidades de mineração de dados, como por exemplo o "*Microsoft SQL Server Data Mining Add-ins for Microsoft Office 2007*", e que importa também analisar como trabalho futuro. A ferramenta Reporting Services abordada no capítulo 5 também é outra das ferramentas importantes para dotar o projecto de Data Warehouse com um interface de utilização que permita a construção de relatórios personalizados pelo utilizador, e que possui integração com o portal Sharepoint.

Como o forte patrocínio existente para o projecto de Data Warehouse se tornou ainda mais vincado com a nova posição de Pró-Presidência do Instituto Politécnico de Setúbal para a área dos Sistemas de Informação, existiu um fortalecimento da possibilidade de estender a aplicação do Data Warehouse a todo o Instituto Politécnico de Setúbal, passando de Data Warehouse da ESTBarreiro/IPS para Data Warehouse do IPS.

É também enquadrado como objectivo futuro o alargamento do acesso ao Data Warehouse, com diferentes níveis de segurança, à restante comunidade escolar ESTBarreiro/IPS.

Em termos de tecnologias, existe também uma série de novas ferramentas que vão chegar ao mercado no início do próximo ano e que interessa explorar. São ferramentas que dão continuidade ao trabalho realizado pela Microsoft na área de Business Intelligence. Assim, um dos objectivos para trabalho futuro neste âmbito é a implementação do Microsoft Office SharePoint Server 2010 (MOSS 2010), a sua integração com o Microsoft Office Excel 2010 e estudar a sua nova funcionalidade PowerPivot que está já disponível em versão beta e promete ser mais um avanço nesta área. Outro aspecto importante em termos de tecnologia é a disponibilização integrada com o Office 2010 da ferramenta Performance Point, que até aqui era um produto vendido em separado. O Microsoft Office Performance Point 2007 é uma

ferramenta de análise e mining de informação que virá acrescentar mais capacidades de análise ao pacote Office da Microsoft.

Um tema que também tem interesse de estudo é o impacto na qualidade dos dados que um Data Warehouse permite trazer a uma organização. Com todas as fases de análise de dados a que o Data Warehouse "obriga", foi possível detectar um conjunto de informação com falhas de consistência ou mesmo erros que de outra forma dificilmente seriam detectados, pelo menos da forma integrada como o foram com as análises efectuadas para a construção do sistema. Por outro lado, a melhoria da disponibilidade da informação aos utilizadores também permite que estes actuem como revisores da mesma, o que é mais um contributo para a melhoria da qualidade dos dados existentes na organização.

7. REFERÊNCIAS

Abraham, D. L., Chen, D., Preston, D., Thomas, D., & Watson, H. J. (2004). Data Warehousing ROI: Justifying and Assessing a Data Warehouse. *Business Intelligence Journal*, Vol. 9, N.º 2, pp. 6-17.

Alimazighi, Z., & Selmoune, N. (2008). A decisional tool for quality inprovement in higher education. *3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA)*, pp. 1-6.

Boehm, B., Egyed, A., Kwan, J., Port, D., Shah, A., & Madachy, R. (1998). Using the winwin spiral model: A case study. *IEEE Computer*, Vol. 31, N.º 7, pp. 33-44.

Breslin, M. (2004). Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. *Business Intelligence Journal*,Vol. 9, N.º 1, pp. 6-20.

Caldeira, C. P. (2008). *Data Warehousing - Conceitos e Modelos.* Edições Sílabo, Lda.

Cardoso, E., Galhardas, H., Trigueiros, M. J., & Silva, R. (2003). A Decision Support System for IST Academic Information. *Informatica an International Journal of Computing and Informatics*, Vol. 27, N.º 3, pp. 313-323.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, Vol. 13, N.º 6, pp. 377-387.

Comingore, D. (2009). *Building a Data Foundation for a BI Solution.* Consultado em Setembro 12, 2009, de Microsoft TechNet: http://technet.microsoft.com/pt-pt/magazine/2009.08.bipartii(en-us).aspx

Cortes, B. (2005). *Sistemas de Suporte à Decisão.* FCA - Editora de Informática, Lda.

Curry, B., & English, B. (2009). *Business Intelligence with Sharepoint, Excel, and Excel Services.* (Microsoft). Consultado em Junho 08, 2009, de Microsoft Technet: http://technet.microsoft.com/en-us/magazine/2009.02.bookbi(printer).aspx

Dijcks, J.-P. (2004). *Business Intelligence Journal: Integrating Data Quality into Your Data Warehouse Architecture.* Consultado em Novembro 27, 2009, de TDWI: http://www.tdwi.org/research/display.aspx?ID=7130

Drucker, P. F. (1988). The Coming of the new Organization. *Harvard Business Review*, Janeiro-Fevereiro, pp. 45-53.

European Commission. (2009). *European Commission Education & Training.* Consultado em Dezembro 02, 2009, de European Commission: http://ec.europa.eu/education/lifelong-learning-policy/doc48_en.htm

Fichter, C. (2009). Thinking Global BI: Data Warehouse Principles For Supporting Enterprise-Enabled Business Intelligence Applications. *The Architecture Journal número 22*, pp. 2-7.

Glassey, O. (2008). A case study on process modelling-three questions and three techniques. *Decision Support Systems*, Vol. 44, N.º 4, pp. 842-853.

GPEARI. (n.d.). *Registo de Alunos Inscritos e diplomados do Ensino Superior (RAIDES).* Consultado em 12 2, 2009, de GPEARI: http://www.gpeari.mctes.pt/archive/doc/documentometodologicoRAIDES.pdf

Hatch, D. (2008). *BI: Is One Version of the Truth Still Out There?* Consultado em Novembro 12, 2009, de TechNewsWorld: http://www.technewsworld.com/story/65359.html

IDC. (2007a). *Microsoft SQL Server 2008 News and Reports.* Consultado em Novembro 23, 2009, de Microsoft: http://www.microsoft.com/sqlserver/2008/en/us/news-reviews.aspx

IDC. (2007b). *Unit4Agresso Press Release.* Consultado em Novembro 21, 2009, de Unit4Agresso: www.unit4agresso.com/files/IDCreport_2006vendor_Shares.pdf

Inmon, W. (2005). *Building the Data Warehouse, 4th Edition.* Wiley Publishing, Inc.

Kimball, R. (2009). *Exploit Your Fact Tables.* Consultado em Novembro 9, 2009, de Information Management: http://www.information-management.com/issues/2007_55/10014852-1.html

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit, Second Edition.* John Wiley & Sons, Inc.

Kimball, R., Mundy, J., & Thornthwaite, W. (2006). *The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset.* Wiley Publishing, Inc.

Londrevie, J., Lindon, D., Dionísio, P., & Rodrigues, V. (1996). *Mercator Teoria e Prática do Marketing (6ª Edição).* Publicações Dom Quixote.

Misner, S. (2009). *Planning Your First Microsoft BI Solution.* Consultado em Setembro 12, 2009, de Microsoft TechNet: http://technet.microsoft.com/pt-pt/magazine/2009.08.introtobi(en-us).aspx

Moody, D. L., & Kortink, M. A. (2000). de Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design. Stockholm, Sweden: Proceedings of the International Workshop on Design and Management of Data Warehouses.

Mundy, J. (2008). *Kimball University: Maintaining Dimension Hierarchies.* Consultado em 12 02, 2009, de intelligent-enterprise.informationweek.com: http://intelligent-enterprise.informationweek.com/showArticle.jhtml;jsessionid=B5WCAHLCE13L5QE1GHPCKHWATMY32JVN?articleID=211600673&pgno=1

Mundy, J., Thornthwaite, W., & Kimball, R. (2006). *The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset.* Wiley Publishing, Inc.

Nunes, M., & O'Neill, H. (2004). *Fundamental de UML.* FCA - Editora de Informática, Lda.

Pearlson, K. E., & Saunders, C. S. (2009). *Managing and Using Information Systems: A Strategic Approach, Fourth Edition.* John Wiley & Sons, Inc.

Pereira, J. L. (1998). *Tecnologias de Bases de Dados.* FCA - Editora de Informática, Lda.

Ramamurthy, K., Sen, A., & Sinha, A. P. (2008). An empirical investigation of the key determinants of data warehouse adoption. *Decision Support Systems*, Vol. 44, N.º 4, pp. 817-841.

Turban, E., & Aronson, J. E. (1997). *Decision Support Systems and Intelligent Systems.* Prentice Hall International Editions.

Turban, E., & Rainer, R. K. (2009). *Introduction to Information Systems, Second Edition.* John Wiley & Sons, Inc.

ANEXOS

I. SCRIPTS DE CRIAÇÃO DO DATA WAREHOUSE

Criação de Tabelas

```
USE [DW_ESTBarreiro]
GO
CREATE TABLE MiniDimCurso
        [Chave do Curso] [int] IDENTITY(1,1) NOT NULL,
        [Código do Curso] [varchar](4) NULL,
        [Designação do Curso] [varchar](100) NULL,
        [Abreviatura do Curso] [varchar](50) NULL,
        [Acrónimo do Curso] [varchar](15) NULL,
        [Regime do Curso] [varchar](15) NULL,
        [Ciclo de Estudos do Curso] [varchar](15) NULL,
        [Abreviatura do Curso para o Cartão] [varchar](16) NULL,
        [Estado do Curso] [varchar](15) NULL,
        [DataInicio] [smalldatetime] NULL,
        [DataFim] [smalldatetime] NULL,
CONSTRAINT [PK_Curso] PRIMARY KEY CLUSTERED
        ([Chave do Curso] ASC) ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE MiniDimAreaCientifica
        [Chave da Área Científica] [int] IDENTITY(1,1) NOT NULL,
        [Código da Área Científica] [varchar](10) NULL,
        [Designação da Área Científica] [varchar](100) NULL,
        [Abreviatura da Área Científica] [varchar](50) NULL,
        [Piso da Área Científica] [varchar](15) NULL,
        [Telefone Geral da Área Científica] [varchar](20) NULL,
        [Responsável pela Área Científica] [varchar](100) NULL,
        [Telefone do Responsável pela Área Científica] [varchar](20) NULL,
        [E-Mail do Responsável pela Área Científica] [varchar](100) NULL,
        [DataInicio] [smalldatetime] NULL,
        [DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_AreaCientifica] PRIMARY KEY CLUSTERED
        ([Chave da Área Científica] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE DimAuditoria
        [Chave_Audit] [int] IDENTITY(1,1) NOT NULL,
        [Chave_Audit_Pai] [int] NOT NULL,
        [NomeTabela] [varchar](50) NULL,
        [Nome_Pkg] [varchar](50) NULL,
        [GUID_Pkg] [uniqueidentifier] NULL,
        [DataInicioExec] [datetime] NULL,
        [DataFimExec] [datetime] NULL,
        [Rows_Extracted] [bigint] NULL,
        [Rows_Error] [bigint] NULL,
        [Rows_Inserted] [bigint] NULL,
        [Rows_Updated] [bigint] NULL,
        [Rows_AltTipo2] [bigint] NULL,
        [RCount_Inicial] [bigint] NULL,
        [RCount_Final] [bigint] NULL,
 CONSTRAINT [PK_DimAuditoria] PRIMARY KEY CLUSTERED
        ([Chave_Audit] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE DimSemestre
        [Chave do Semestre] [int] IDENTITY(1,1) NOT NULL,
        [Descrição do Semestre] [varchar](50) NULL,
        [Ano Lectivo do Semestre] [varchar](9) NULL,
        [Data de Início do Semestre] [varchar](20) NULL,
        [Data de Fim do Semestre] [varchar](20) NULL,
        [DataInicio] [smalldatetime] NULL,
        [DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_DimSemestre] PRIMARY KEY CLUSTERED
        ([Chave do Semestre] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
```

```
CREATE TABLE DimAluno
	[Chave do Aluno] [int] IDENTITY(1,1) NOT NULL,
	[Número de Aluno] [int] NOT NULL,
	[Nome do Aluno] [varchar](150) NULL,
	[Morada - Férias] [varchar](200) NULL,
	[Código Postal - Férias] [varchar](15) NULL,
	[Localidade - Férias] [varchar](50) NULL,
	[Concelho - Férias] [varchar](100) NULL,
	[Distrito - Férias] [varchar](100) NULL,
	[Morada - Aulas] [varchar](200) NULL,
	[Código Postal - Aulas] [varchar](15) NULL,
	[Localidade - Aulas] [varchar](50) NULL,
	[Concelho - Aulas] [varchar](100) NULL,
	[Distrito - Aulas] [varchar](100) NULL,
	[Número do BI] [varchar](15) NULL,
	[Local de Emissão do BI] [varchar](50) NULL,
	[Data de Validade do BI] [varchar](15) NULL,
	[Naturalidade] [varchar](50) NULL,
	[Nome do Pai] [varchar](150) NULL,
	[Nome da Mãe] [varchar](150) NULL,
	[Data de Nascimento] [varchar](15) NULL,
	[Telefone] [varchar](15) NULL,
	[Telemóvel] [varchar](15) NULL,
	[Sexo] [varchar](15) NULL,
	[Nacionalidade] [varchar](50) NULL,
	[Número de Contribuinte] [varchar](15) NULL,
	[E-Mail #1] [varchar](100) NULL,
	[E-Mail #2] [varchar](100) NULL,
	[Data de Entrada] [varchar](15) NULL,
	[Ano Lectivo de Entrada] [varchar](9) NULL,
	[Tipo de Inscrição] [varchar](60) NULL,
	[DataInicio] [smalldatetime] NULL,
	[DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_Aluno] PRIMARY KEY CLUSTERED
	([Chave do Aluno] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE FactAvaliacao_Aluno_Semestre
	[Chave do Aluno] [int] NOT NULL,
	[Chave do Semestre] [int] NOT NULL,
	[Chave do Curso] [int] NOT NULL,
	[Chave do Tipo de Inscrição e Estado de Avaliação] [int] NOT NULL,
	[Número de UCs - Valor Aritmético] [int] NULL,
	[Soma aritmética das Notas] [int] NULL,
	[Número de UCs - Valor Ponderado] [int] NULL,
	[Soma das Notas Ponderada] [int] NULL,
	[Soma dos Créditos ECTS] [decimal](4, 2) NULL,
	[Ano Curricular] [int] NULL,
 CONSTRAINT [PK_FactAvaliacao_Aluno_Semestre_1] PRIMARY KEY CLUSTERED
	([Chave do Aluno] ASC,
	[Chave do Semestre] ASC,
	[Chave do Curso] ASC,
	[Chave do Tipo de Inscrição e Estado de Avaliação] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE DimUnidadeCurricular
	[Chave da Unidade Curricular] [int] IDENTITY(1,1) NOT NULL,
	[Código da Unidade Curricular] [varchar](7) NULL,
	[Designação da Unidade Curricular] [varchar](100) NULL,
	[Abreviatura da Unidade Curricular] [varchar](50) NULL,
	[Tipo da Unidade Curricular] [varchar](30) NULL,
	[Ano Curricular da Unidade Curricular] [varchar](30) NULL,
	[Semestre Curricular da Unidade Curricular] [varchar](30) NULL,
	[Unidades Curriculares Equivalentes] [varchar](50) NULL,
	[Unidade Curricular Opcional] [varchar](40) NULL,
	[Unidade Curricular Funciona nos 2 Semestres] [varchar](40) NULL,
	[Estado] [varchar](10) NULL,
	[DataInicio] [smalldatetime] NULL,
	[DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_UnidadeCurricular] PRIMARY KEY CLUSTERED
	([Chave da Unidade Curricular] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
```

```
CREATE TABLE DimProfessor
	[Chave do Professor] [int] IDENTITY(1,1) NOT NULL,
	[Código do Professor] [int] NULL,
	[Nome do Professor] [varchar](100) NULL,
	[Data de Nascimento do Professor] [varchar](15) NULL,
	[Morada do Professor] [varchar](150) NULL,
	[Localidade do Professor] [varchar](30) NULL,
	[Código Postal do Professor] [varchar](8) NULL,
	[Número de BI do Professor] [varchar](15) NULL,
	[Estado Civil do Professor] [varchar](15) NULL,
	[Data de Validade do BI do Professor] [varchar](15) NULL,
	[Número de Contribuinte do Professor] [varchar](15) NULL,
	[Código da Repartição de Finanças do Professor] [varchar](15) NULL,
	[Bairro Fiscal do Professor] [varchar](50) NULL,
	[Telefone do Professor] [varchar](15) NULL,
	[Telemóvel do Professor] [varchar](15) NULL,
	[E-Mail #1 do Professor] [varchar](100) NULL,
	[E-Mail #2 do Professor] [varchar](100) NULL,
	[Nome do Pai do Professor] [varchar](100) NULL,
	[Nome da Mãe do Professor] [varchar](100) NULL,
	[NIB do Professor] [varchar](21) NULL,
	[Número de ADSE do Professor] [varchar](15) NULL,
	[Dependentes do Professor] [varchar](15) NULL,
	[Sexo do Professor] [varchar](15) NULL,
	[UserName do Professor] [varchar](50) NULL,
	[número do Gabinete do Professor] [varchar](30) NULL,
	[Piso do Gabinete do Professor] [varchar](30) NULL,
	[Extensão telefónica do Professor] [varchar](30) NULL,
	[DataInicio] [smalldatetime] NULL,
	[DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_Professor] PRIMARY KEY CLUSTERED
	([Chave do Professor] ASC ON [PRIMARY]) ON [PRIMARY]
GO
------------------------------------------------------------------------------------------
CREATE TABLE FactAvaliacao_Staging
	[NumAluno] [int] NOT NULL,
	[CodAnoLectivo] [int] NOT NULL,
	[AnoLectivo] [varchar](9) NOT NULL,
	[AnoCurricular] [int] NULL,
	[Semestre] [varchar](5) NOT NULL,
	[CodDisciplina] [varchar](7) NOT NULL,
	[CodCurso] [varchar](4) NOT NULL,
	[Ciclo] [int] NULL,
	[CodAreaCientifica] [varchar](10) NOT NULL,
	[CodResponsavel] [int] NULL,
	[Creditos] [decimal](4, 2) NULL,
	[Peso] [decimal](4, 2) NULL,
	[NumInsc] [int] NULL,
	[DataInscricao] [smalldatetime] NOT NULL,
	[DataAvaliacao] [smalldatetime] NULL,
	[Nota] [int] NULL,
	[TipoInscricao] [int] NOT NULL,
	[TipoAvaliacao] [varchar](20) NULL,
	[Status] [varchar](30) NULL,
	[CodTurma] [int] NULL,
	[Criado Em] [smalldatetime] NULL,
	[Modificado Em] [smalldatetime] NULL,
 CONSTRAINT [PK_FactAvaliacao_Staging] PRIMARY KEY CLUSTERED
	([NumAluno] ASC,
	[CodAnoLectivo] ASC,
	[Semestre] ASC,
	[CodDisciplina] ASC,
	[TipoInscricao] ASC) ON [PRIMARY]) ON [PRIMARY]
GO
------------------------------------------------------------------------------------------
CREATE TABLE DimCorpoDocente
	[Chave do Corpo Docente] [int] NOT NULL,
	[Chave do Professor] [int] NOT NULL,
	[Chave do Semestre] [int] NOT NULL,
	[Chave da Unidade Curricular] [int] NULL,
	[Factor de Ponderação] [varchar](20) NULL,
 CONSTRAINT [PK_CorpoDocente] PRIMARY KEY CLUSTERED
	([Chave do Corpo Docente] ASC,
	[Chave do Professor] ASC,
	[Chave do Semestre] ASC ON [PRIMARY]) ON [PRIMARY]
GO
------------------------------------------------------------------------------------------
```

```
CREATE TABLE DimData
	[Chave da Data] [int] NOT NULL,
	[Data SQL] [smalldatetime] NULL,
	[Descrição da Data] [varchar](100) NULL,
	[Dia da Semana] [varchar](50) NULL,
	[Mês do Ano] [varchar](50) NULL,
	[Número do Mês no Ano] [int] NULL,
	[Dia do Ano] [int] NULL,
	[Dia do Mês] [int] NULL,
	[Número da Semana] [int] NULL,
	[Trimestre] [varchar](20) NULL,
	[Semestre Civil] [varchar](20) NULL,
	[Ano Civil] [varchar](4) NULL,
	[Indicador de Feriado] [varchar](100) NULL,
	[Indicador de Fim de Semana] [varchar](100) NULL,
	[Momento Académico] [varchar](100) NULL,
	[Semestre Lectivo] [varchar](30) NULL,
	[Ano Lectivo] [varchar](9) NULL,
	[DataInicio] [smalldatetime] NULL,
	[DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_DimData] PRIMARY KEY CLUSTERED
	([Chave da Data] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE FactAvaliacao
	[Chave da Data de Inscrição] [int] NOT NULL,
	[Chave da Data de Avaliacao] [int] NOT NULL,
	[Chave do Aluno] [int] NOT NULL,
	[Chave da Unidade Curricular] [int] NOT NULL,
	[Chave do Responsável da Unidade Curricular] [int] NOT NULL,
	[Chave do Corpo Docente] [int] NOT NULL,
	[Chave do Curso] [int] NOT NULL,
	[Chave da Área Científica] [int] NOT NULL,
	[Chave do Semestre] [int] NOT NULL,
	[Chave do Tipo de Inscrição e Estado de Avaliação] [int] NOT NULL,
	[Nota] [int] NULL,
	[Créditos ECTS] [decimal](6, 3) NULL,
	[Peso na média final] [decimal](6, 3) NULL,
	[Número de Inscrição] [int] NULL,
	[Ano Curricular] [int] NULL,
 CONSTRAINT [PK_FactAvaliacao] PRIMARY KEY CLUSTERED
	([Chave do Aluno] ASC,
	[Chave da Unidade Curricular] ASC,
	[Chave do Semestre] ASC,
	[Chave do Tipo de Inscrição e Estado de Avaliação] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE FactAvaliacao_UC_Semestre_PIVOT
	[Chave da Unidade Curricular] [int] NOT NULL,
	[Chave do Semestre] [int] NOT NULL,
	[Chave do Responsável da UC] [int] NULL,
	[Chave do Corpo Docente] [int] NULL,
	[Chave do Curso] [int] NULL,
	[Chave da Área Científica] [int] NULL,
	[Número de Alunos Aprovados] [int] NULL,
	[Número de Alunos Aprovados em 1ª Inscrição] [int] NULL,
	[Número de Alunos Reprovados] [int] NULL,
	[Número de Alunos Reprovados em 1ª Inscrição] [int] NULL,
	[Número de Alunos Não Avaliados] [int] NULL,
	[Número de Alunos Não Avaliados em 1ª Inscrição] [int] NULL,
	[Soma das notas dos Alunos Aprovados] [int] NULL,
	[Soma das notas dos Alunos Aprovados em 1ª Inscrição] [int] NULL,
	[Soma das notas dos Alunos Reprovados] [int] NULL,
	[Soma das notas dos Alunos Reprovados em 1ª Inscrição] [int] NULL,
 CONSTRAINT [PK_FactAvaliacao_UC_Semestre_PIVOT] PRIMARY KEY CLUSTERED
	([Chave da Unidade Curricular] ASC,
	[Chave do Semestre] ASC ON [PRIMARY]) ON [PRIMARY]
GO
-------------------------------------------------------------------------------------------
CREATE TABLE DimTipoInscricaoEstadoAvaliacao
	[Chave do Tipo de Inscrição e Estado de Avaliação] [int] IDENTITY(1,1) NOT NULL,
	[Código do Tipo de Inscrição] [int] NULL,
	[Designação do Tipo de Inscrição] [varchar](50) NULL,
	[Descrição do Tipo de Inscrição] [varchar](200) NULL,
	[Participa na Média Final] [varchar](50) NULL,
	[Altera o número de inscrição] [varchar](50) NULL,
	[Designação do Estado de Avaliação] [varchar](50) NULL,
```

```
        [Descrição do Estado de Avaliação] [varchar](200) NULL,
        [DataInicio] [smalldatetime] NULL,
        [DataFim] [smalldatetime] NULL,
 CONSTRAINT [PK_TipoInscricao] PRIMARY KEY CLUSTERED
        ([Chave do Tipo de Inscrição e Estado de Avaliação] ASC ON [PRIMARY]) ON
[PRIMARY]
GO
-------------------------------------------------------------------------------------
CREATE TABLE FactAvaliacao_UC_Semestre
        [Chave da Unidade Curricular] [int] NOT NULL,
        [Chave do Semestre] [int] NOT NULL,
        [Chave do Responsável da UC] [int] NOT NULL,
        [Chave do Corpo Docente] [int] NOT NULL,
        [Chave do Curso] [int] NOT NULL,
        [Chave da Área Científica] [int] NOT NULL,
        [Chave do Tipo de Inscrição e Estado de Avaliação] [int] NOT NULL,
        [Número de Inscrição] [int] NOT NULL,
        [Número de Registos] [int] NULL,
        [Soma das Notas] [int] NULL,
 CONSTRAINT [PK_FactAvaliacao_UC_Semestre_1] PRIMARY KEY CLUSTERED
        ([Chave da Unidade Curricular] ASC,
        [Chave do Semestre] ASC,
        [Chave do Responsável da UC] ASC,
        [Chave do Corpo Docente] ASC,
        [Chave do Curso] ASC,
        [Chave da Área Científica] ASC,
        [Chave do Tipo de Inscrição e Estado de Avaliação] ASC,
        [Número de Inscrição] ASC ON [PRIMARY]) ON [PRIMARY]
GO
```

Criação de vistas para preenchimento da tabela de factos

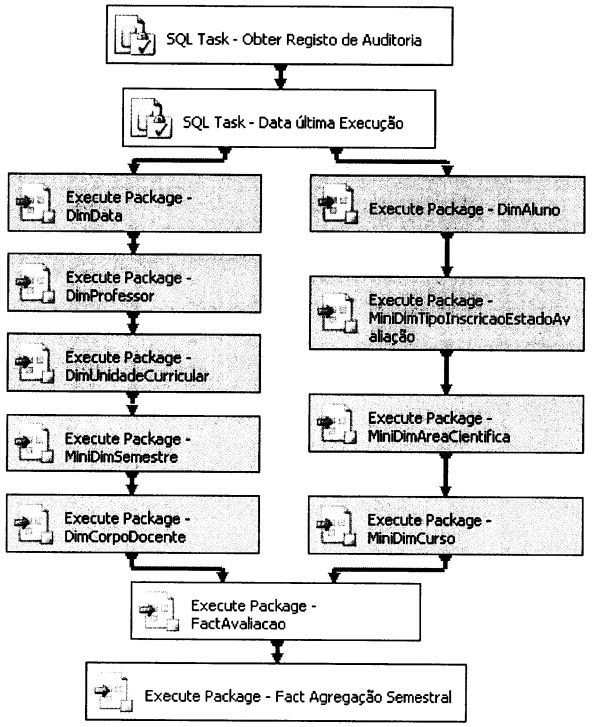
```
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimUC AS
SELECT [Chave da Unidade Curricular], [Código da Unidade Curricular]
FROM DimUnidadeCurricular
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimTipoInscricaoEstadoAvaliacao AS
SELECT [Chave do Tipo de Inscrição e Estado de Avaliação], [Designação do Estado de
Avaliação], [Código do Tipo de Inscrição]
FROM DimTipoInscricaoEstadoAvaliacao
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimSemestre AS
SELECT [Chave do Semestre], [Descrição do Semestre], [Ano Lectivo do Semestre]
FROM DimSemestre
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimCurso AS
SELECT [Chave do Curso], [Código do Curso], [Ciclo de estudos do Curso]
FROM MiniDimCurso
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimAluno AS
SELECT [Chave do Aluno], [Número de Aluno]
FROM DimAluno
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW vTmpDimAC AS
SELECT [Chave da Área Científica], [Código da Área Científica]
FROM MiniDimAreaCientifica
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
CREATE VIEW VTmp_DimV_Responsavel AS
SELECT [Chave do professor], [Código do professor]
FROM DimProfessor
WHERE '{data}' >= DataInicio AND DataFim IS NULL
GO
-------------------------------------------------------------------------------------
```

Procedimento tipo para inserção de registos dummy

```
CREATE PROC USP_InsRegisto_Inexistencias_DimProfessor
AS
BEGIN
   SET IDENTITY_INSERT DimProfessor ON
   IF (SELECT COUNT([Chave do Professor]) FROM DimProfessor WHERE [Chave do Professor]
=-1) = 0
   INSERT INTO DimProfessor ([Chave do Professor], [Código do Professor], [Nome do
Professor], DataInicio)
   VALUES (-1, -1, 'Sem professor Responsável (CC)', getdate())
   SET IDENTITY_INSERT DimProfessor OFF
END
GO
-----------------------------------------------------------------------------------------
CREATE PROC USP_InsRegisto_AlunoNaoAvaliado_DimData
AS
BEGIN
       SET IDENTITY_INSERT DimData ON
              IF (SELECT COUNT([Chave da Data]) FROM DimData WHERE [Chave da Data] =-1)
= 0
              INSERT INTO DimData ([Chave da Data], [Descrição da Data], DataInicio)
              VALUES (-1, 'Sem data de Avaliação', '01-01-1999')--getdate())
       SET IDENTITY_INSERT DimProfessor OFF
END
GO
-----------------------------------------------------------------------------------------
CREATE PROC USP_AddKey_ @CodTurma int
AS
BEGIN
       SET NOCOUNT ON;
       DECLARE @DataInicio smalldatetime
       SET @DataInicio = (Select GETDATE())
       IF NOT EXISTS (SELECT [Código da Turma] FROM DimTurma WHERE [Código da Turma] =
@CodTurma)
              INSERT INTO DimTurma ([Código da Turma], [Designação da Turma],
Datainicio)
              VALUES (@CodTurma, 'Não Insc.', @DataInicio)
END
GO
```

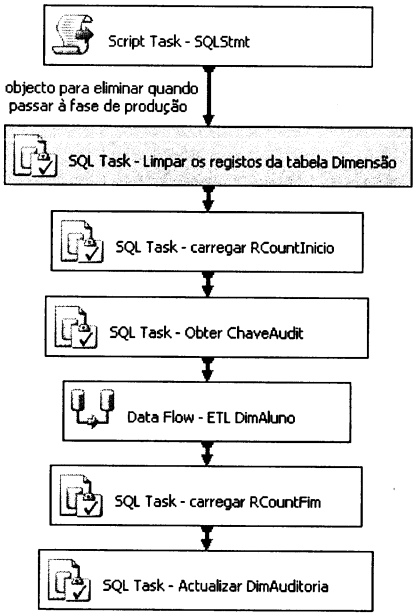
II. ETL – GRÁFICOS DE TRANSFORMAÇÃO

Pacote de Controlo (nível 1)

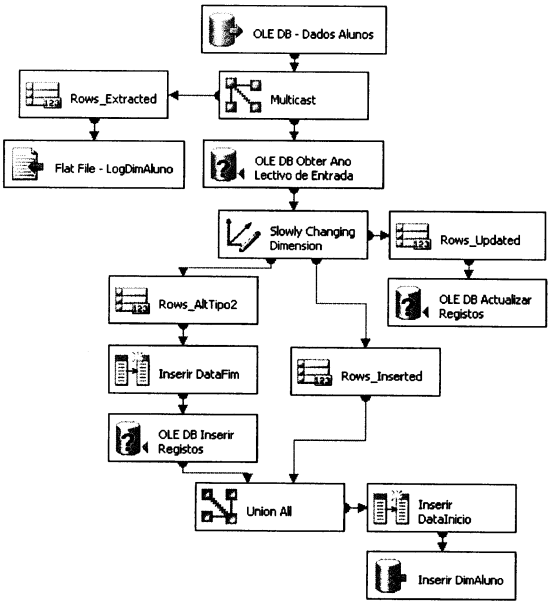


Dimensão Aluno

Nível 2

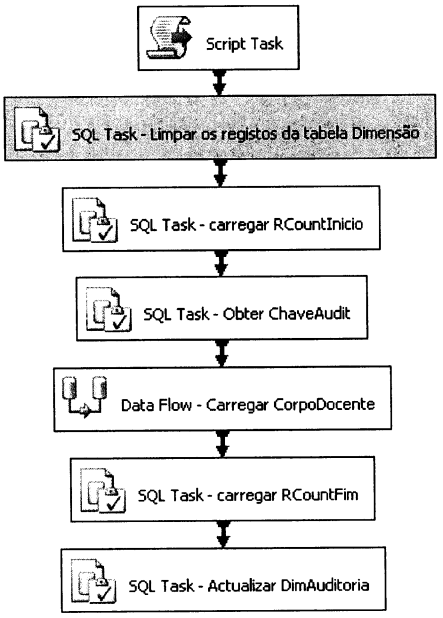


Nível 3

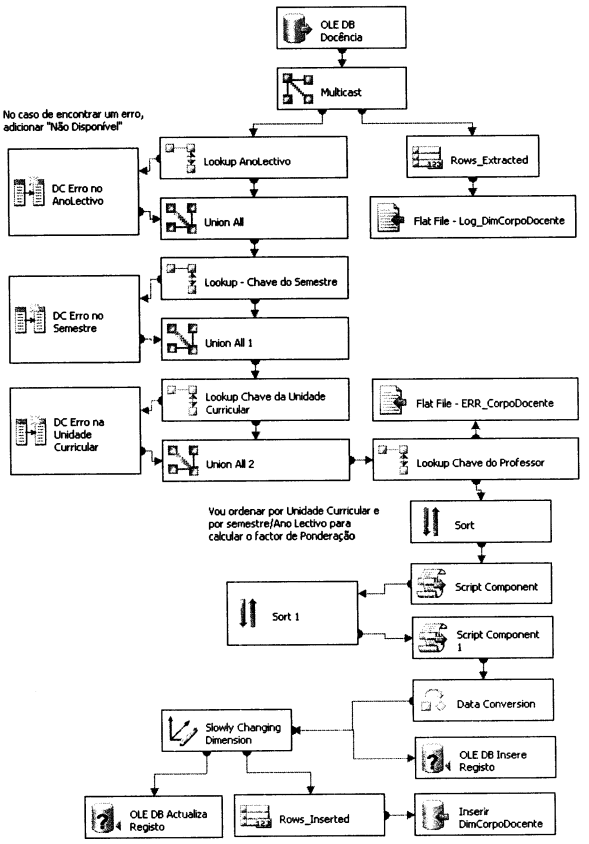


Dimensão Corpo Docente

Nível 2

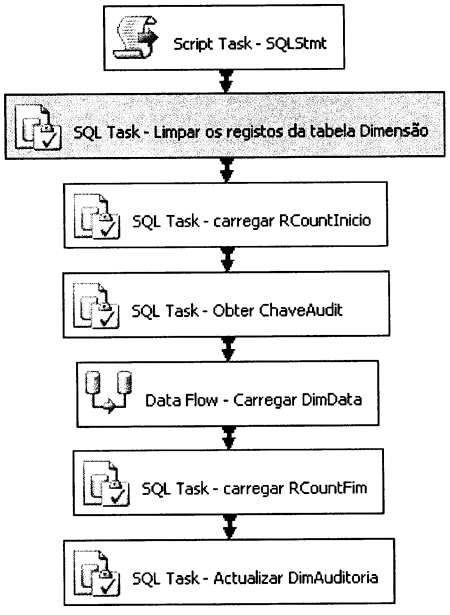


Nível 3

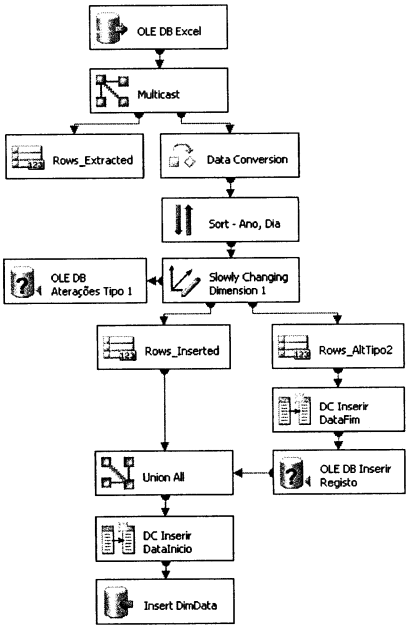


Dimensão Data

Nível 2

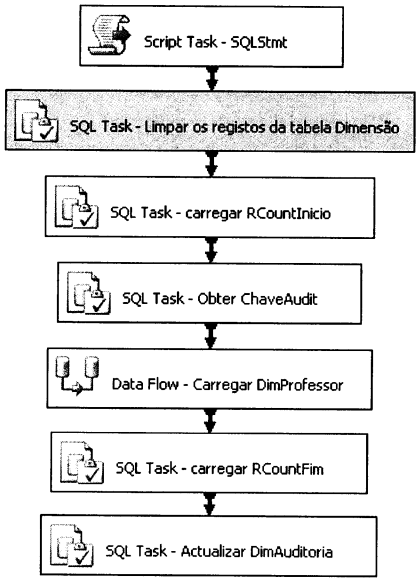


Nível 3

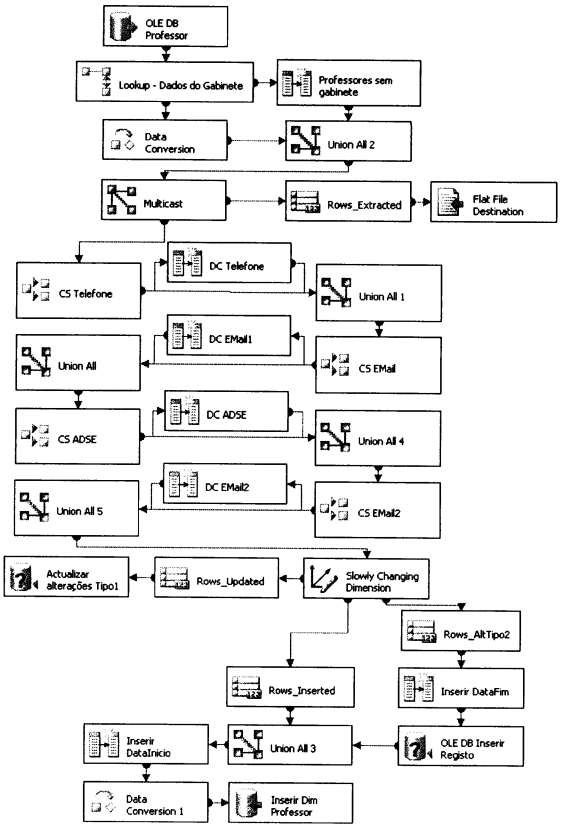


Dimensão Professor

Nível 2

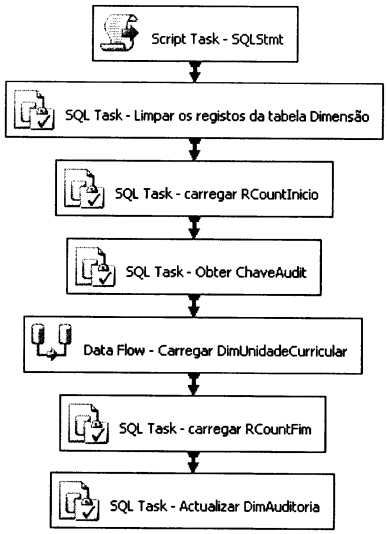


Nível 3

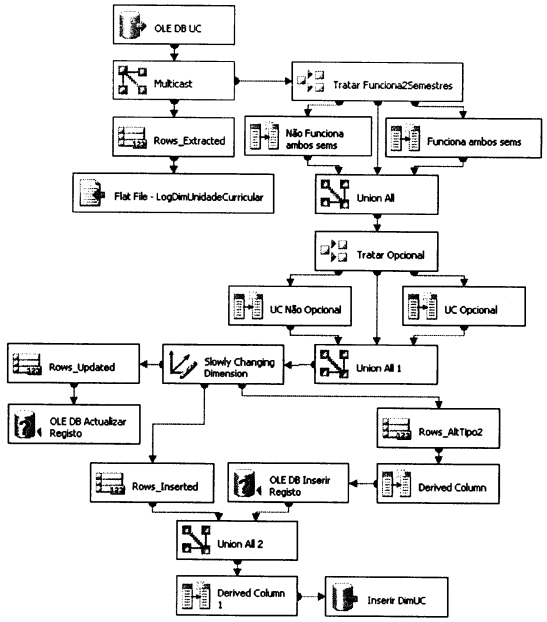


Dimensão Unidade Curricular

Nível 2

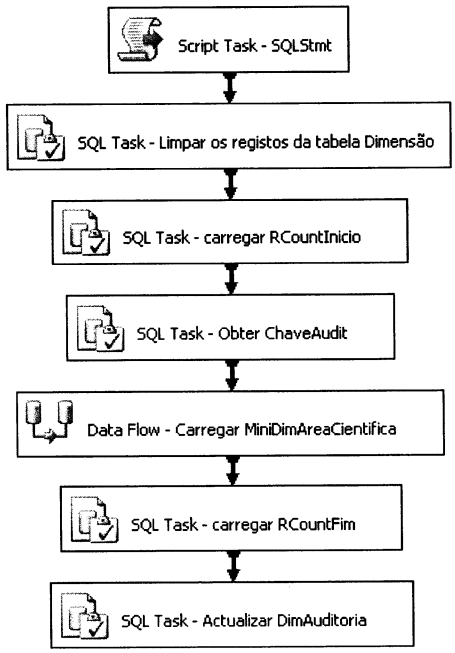


Nível 3

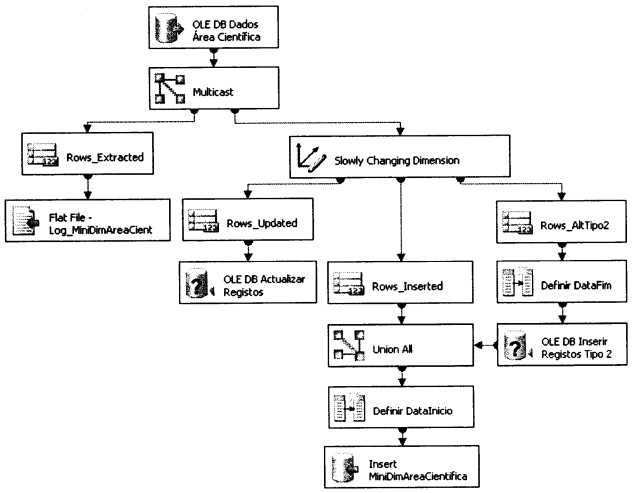


Dimensão Área Científica

Nível 2

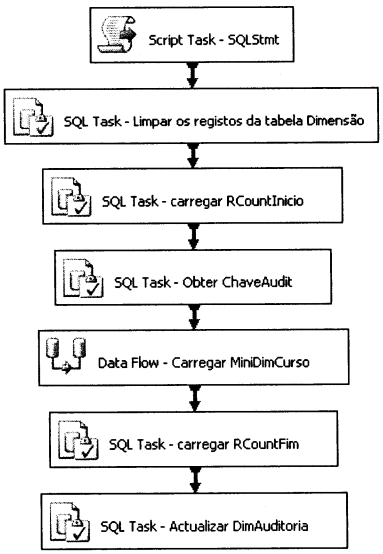


Nível 3

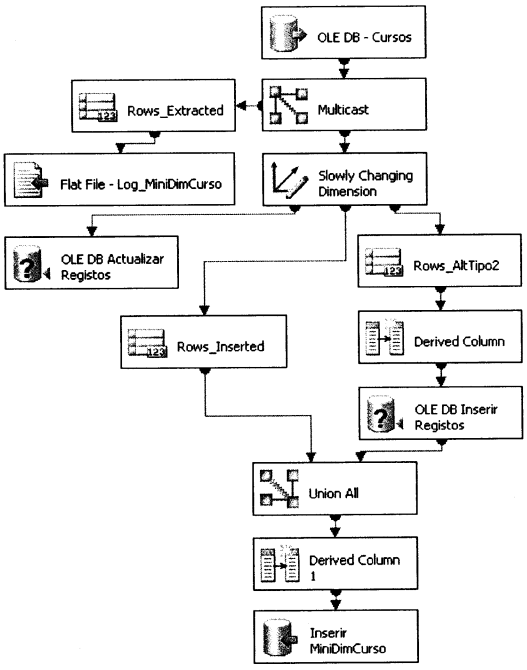


Dimensão Curso

Nível 2

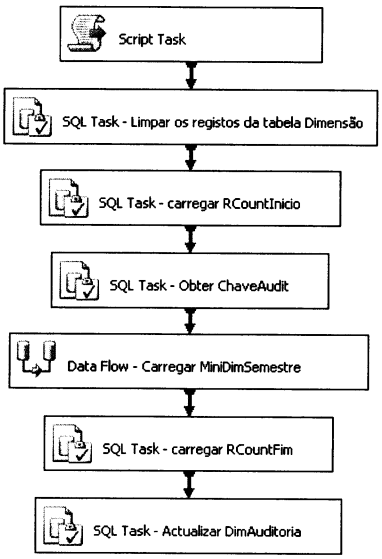


Nível 3

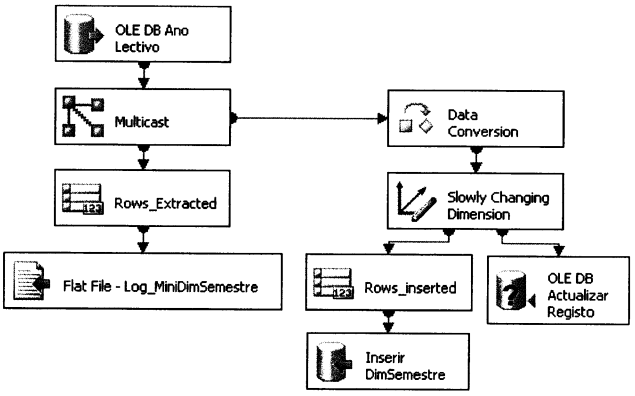


Dimensão Semestre

Nível 2

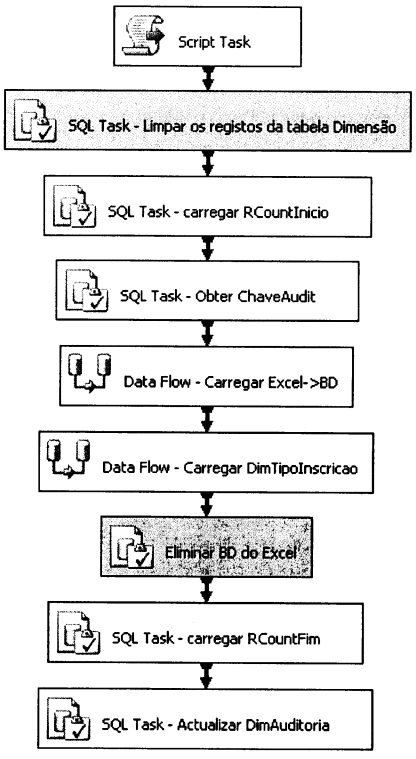


Nível 3

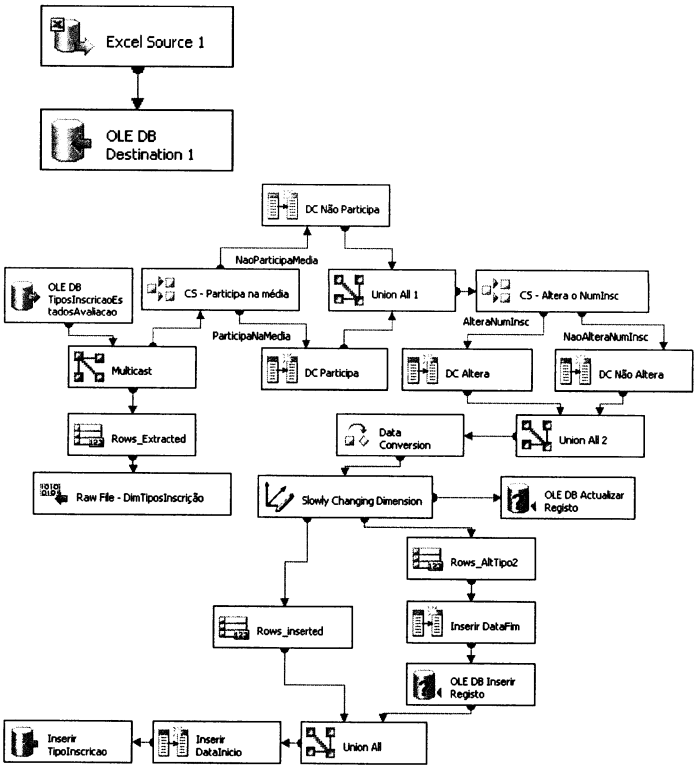


Dimensão Tipo de Inscrição

Nível 2

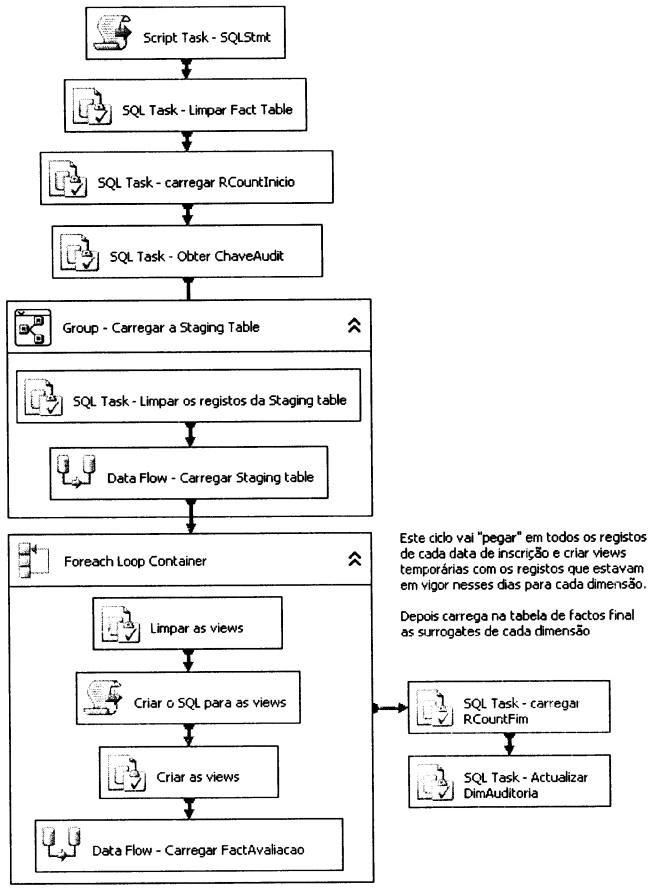


Nível 3

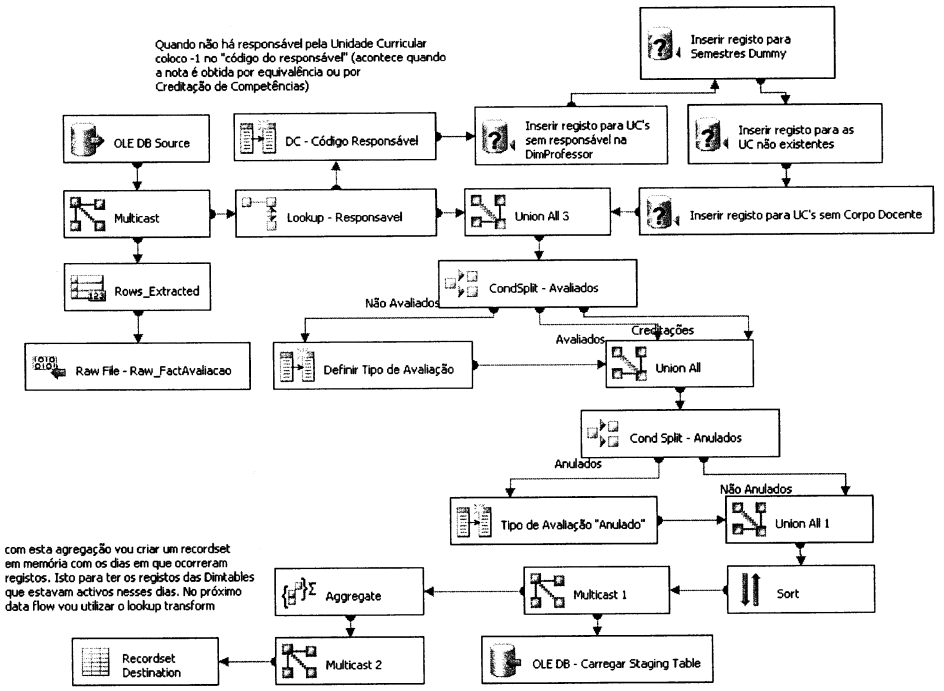


Tabelas de Factos do Data Mart Percurso Académico

Nível 2



Carregamento da tabela de estágio (Nível 3)



Nível 3

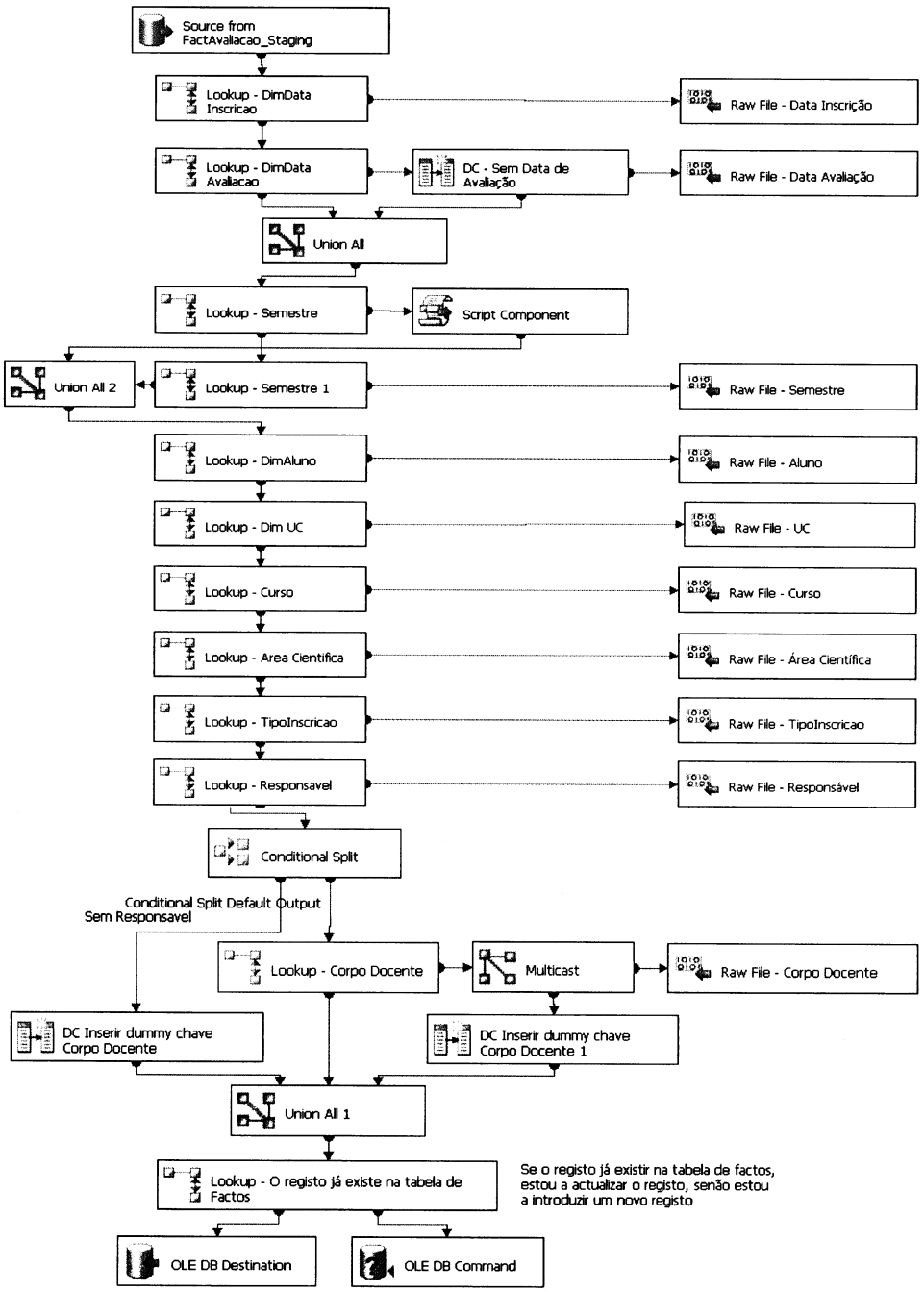
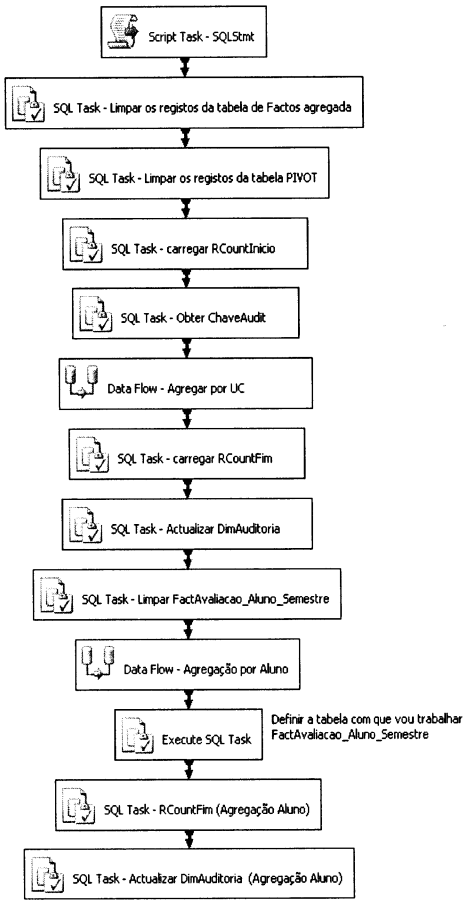
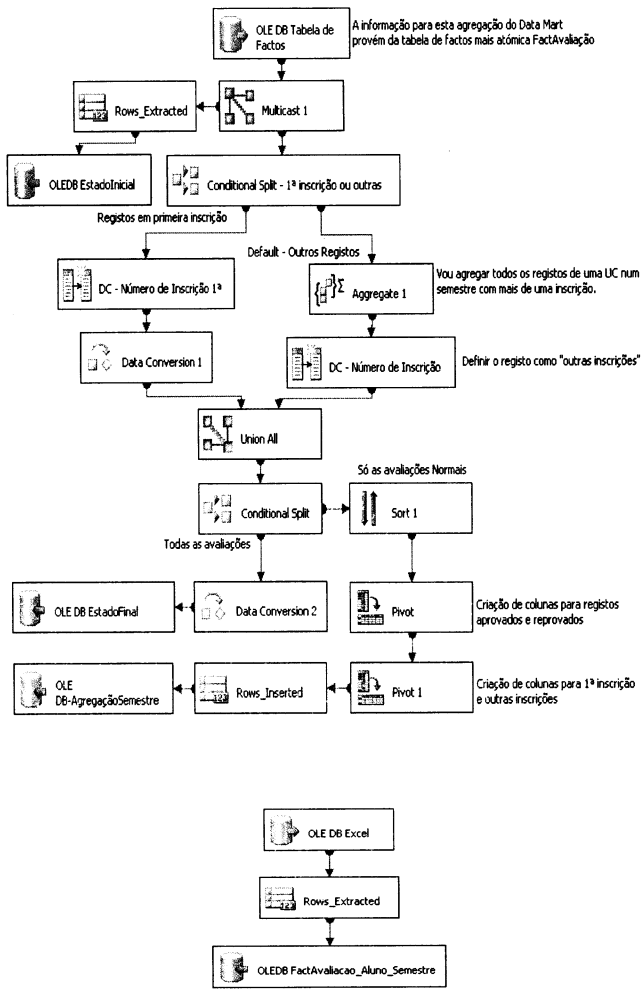

Source from FactAvaliacao_Staging
Lookup - DimData Inscricao
Raw File - Data Inscrição
Lookup - DimData Avaliacao
DC - Sem Data de Avaliação
Raw File - Data Avaliação
Union All
Lookup - Semestre
Script Component
Union All 2
Lookup - Semestre 1
Raw File - Semestre
Lookup - DimAluno
Raw File - Aluno
Lookup - Dim UC
Raw File - UC
Lookup - Curso
Raw File - Curso
Lookup - Area Cientifica
Raw File - Área Científica
Lookup - TipoInscricao
Raw File - TipoInscricao
Lookup - Responsavel
Raw File - Responsável
Conditional Split
Conditional Split Default Output Sem Responsavel
Lookup - Corpo Docente
Multicast
Raw File - Corpo Docente
DC Inserir dummy chave Corpo Docente
DC Inserir dummy chave Corpo Docente 1
Union All 1
Lookup - O registo já existe na tabela de Factos
Se o registo já existir na tabela de factos, estou a actualizar o registo, senão estou a introduzir um novo registo
OLE DB Destination
OLE DB Command

Tabela de Factos de Avaliação Agregados

Nível 2

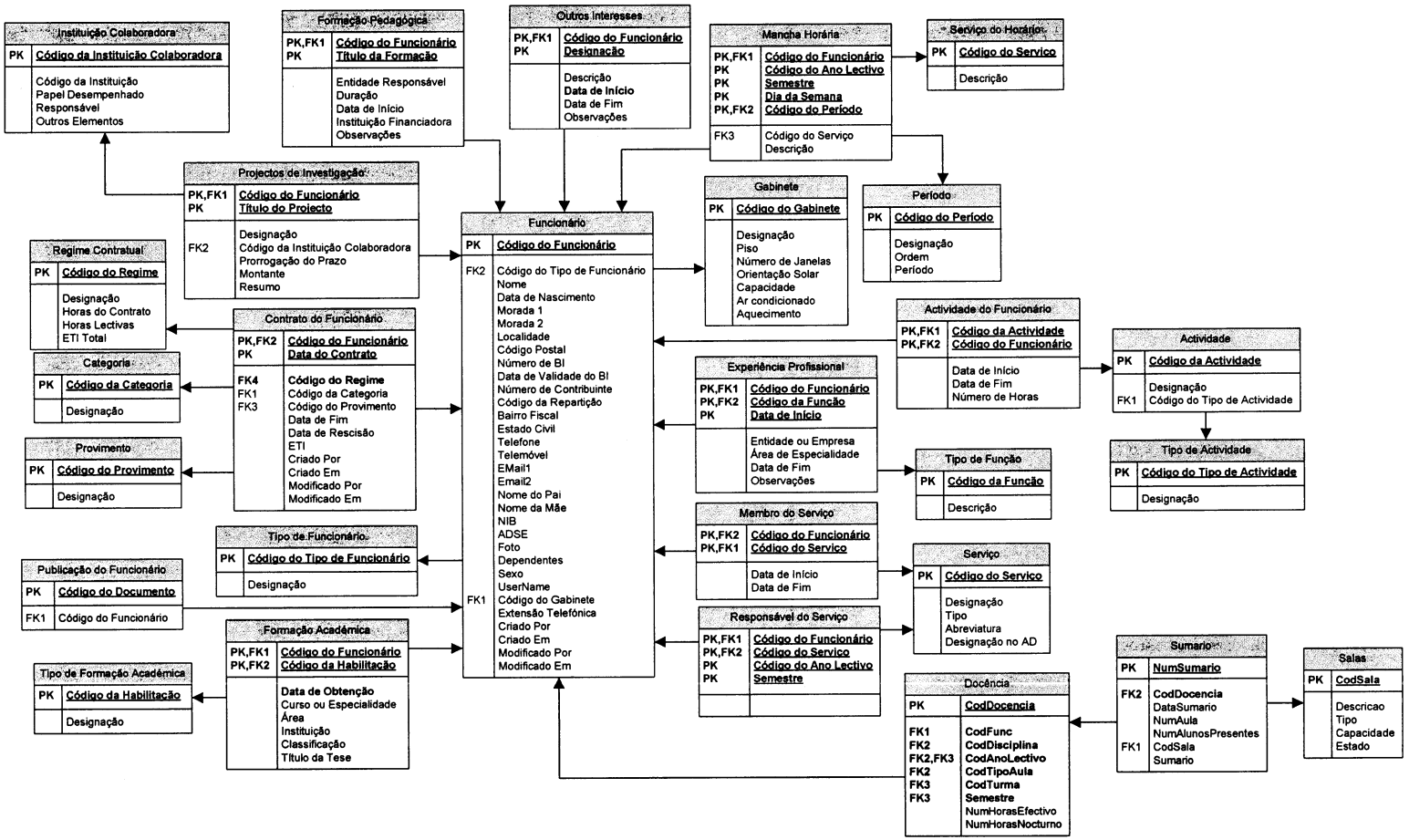


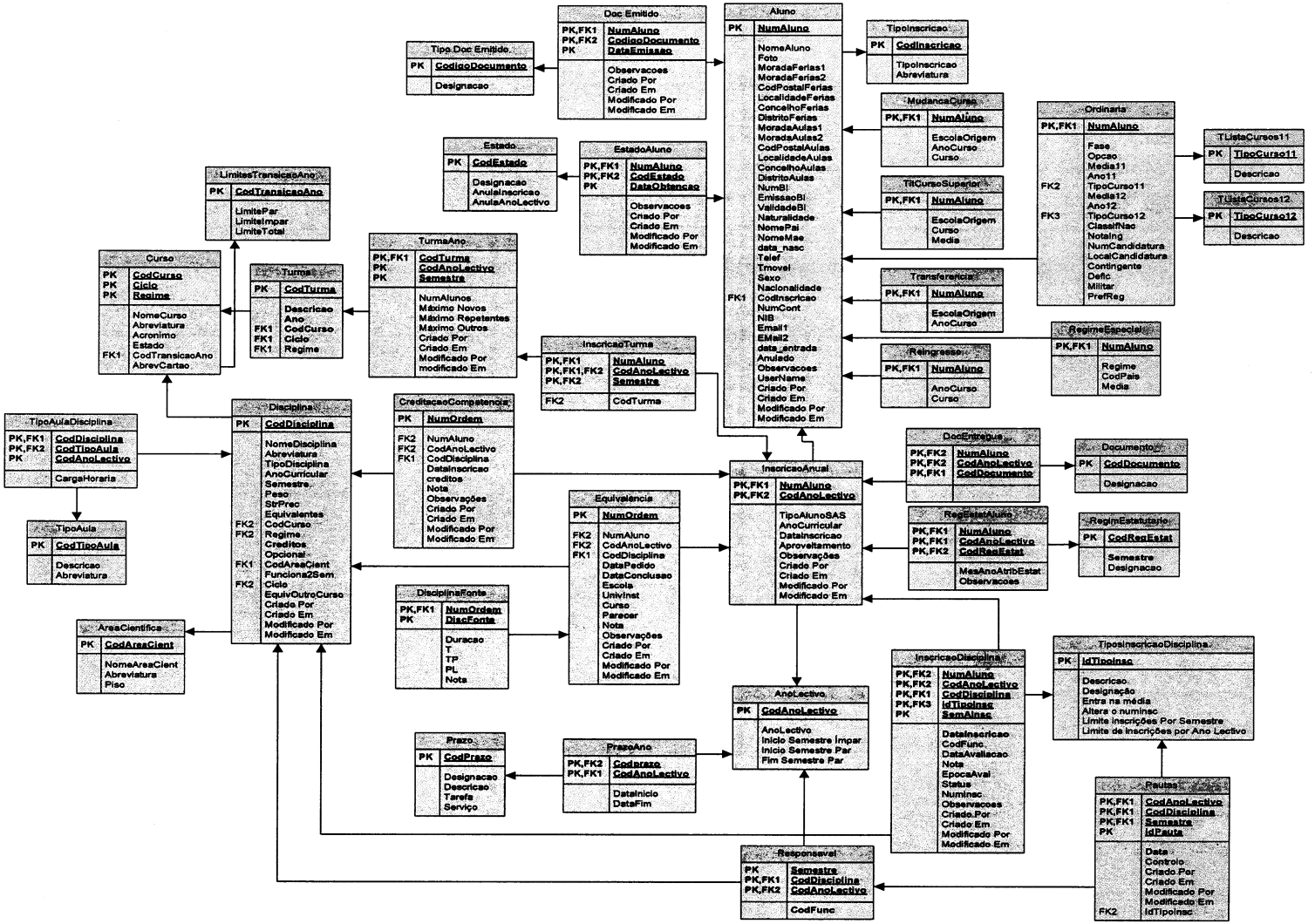
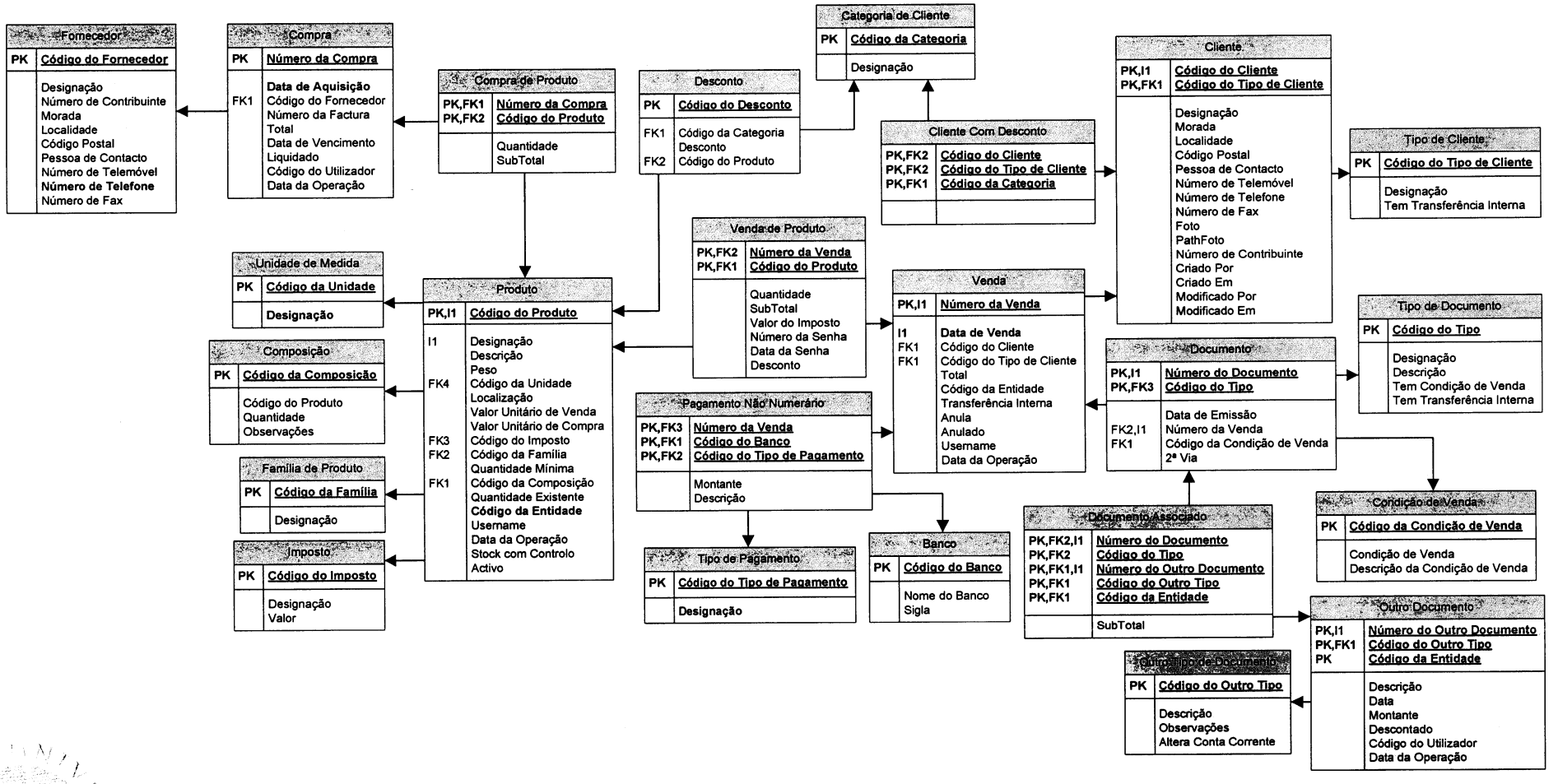
Nível 3



III. BASES DE DADOS RELACIONAIS

Gestão de Recursos Humanos e Serviço Docente




Doc Emitido
PK,FK1 NumAluno
PK,FK2 CodigoDocumento
PK DataEmissao
Observacoes
Criado Por
Criado Em
Modificado Por
Modificado Em
Tipo Doc Emitido
PK CodigoDocumento
Designacao
Aluno
PK NumAluno
NomeAluno
Foto
MoradaFerias1
MoradaFerias2
CodPostalFerias
LocalidadeFerias
ConcelhoFerias
DistritoFerias
MoradaAulas1
MoradaAulas2
CodPostalAulas
LocalidadeAulas
ConcelhoAulas
DistritoAulas
NumBI
EmissaoBI
ValidadeBI
Naturalidade
NomePai
NomeMae
data_nasc
Telef
Tmovel
Sexo
Nacionalidade
FK1 CodInscricao
NumCont
NIB
Email1
EMail2
data_entrada
Anulado
Observacoes
UserName
Criado Por
Criado Em
Modificado Por
Modificado Em
TipoInscricao
PK CodInscricao
TipoInscricao
Abreviatura
MudancaCurso
PK,FK1 NumAluno
EscolaOrigem
AnoCurso
Curso
TitCursoSuperior
PK,FK1 NumAluno
EscolaOrigem
Curso
Media
Transferencia
PK,FK1 NumAluno
EscolaOrigem
AnoCurso
Reingresso
PK,FK1 NumAluno
AnoCurso
Curso
Ordinaria
PK,FK1 NumAluno
Fase
Opcao
Media11
Ano11
FK2 TipoCurso11
Media12
Ano12
FK3 TipoCurso12
ClassifNac
NotaIng
NumCandidatura
LocalCandidatura
Contingente
Defic
Militar
PrefReg
TListaCursos11
PK TipoCurso11
Descricao
TListaCursos12
PK TipoCurso12
Descricao
RegimeEspecial
PK,FK1 NumAluno
Regime
CodPais
Media
Estado
PK CodEstado
Designacao
AnulaInscricao
AnulaAnoLectivo
EstadoAluno
PK,FK1 NumAluno
PK,FK2 CodEstado
PK DataObtencao
Observacoes
Criado Por
Criado Em
Modificado Por
Modificado Em
LimitesTransicaoAno
PK CodTransicaoAno
LimitePar
LimiteImpar
LimiteTotal
Curso
PK CodCurso
PK Ciclo
PK Regime
NomeCurso
Abreviatura
Acronimo
Estado
FK1 CodTransicaoAno
AbrevCartao
Turmas
PK CodTurma
Descricao
Ano
FK1 CodCurso
FK1 Ciclo
FK1 Regime
TurmasAno
PK,FK1 CodTurma
PK CodAnoLectivo
PK Semestre
NumAlunos
Máximo Novos
Máximo Repetentes
Máximo Outros
Criado Por
Criado Em
Modificado Por
modificado Em
InscricaoTurma
PK,FK1 NumAluno
PK,FK1,FK2 CodAnoLectivo
PK,FK2 Semestre
FK2 CodTurma
TipoAulaDisciplina
PK,FK1 CodDisciplina
PK,FK2 CodTipoAula
PK CodAnoLectivo
CargaHoraria
TipoAula
PK CodTipoAula
Descricao
Abreviatura
Disciplina
PK CodDisciplina
NomeDisciplina
Abreviatura
TipoDisciplina
AnoCurricular
Semestre
Peso
StrPrec
Equivalentes
FK2 CodCurso
FK2 Regime
Creditos
Opcional
FK1 CodAreaCient
Funciona2Sem
FK2 Ciclo
EquivOutroCurso
Criado Por
Criado Em
Modificado Por
Modificado Em
AreaCientifica
PK CodAreaCient
NomeAreaCient
Abreviatura
Piso
CreditacaoCompetencias
PK NumOrdem
FK2 NumAluno
FK2 CodAnoLectivo
FK1 CodDisciplina
DataInscricao
creditos
Nota
Observações
Criado Por
Criado Em
Modificado Por
Modificado Em
DisciplinaFonte
PK,FK1 NumOrdem
PK DiscFonte
Duracao
T
TP
PL
Nota
Equivalencia
PK NumOrdem
FK2 NumAluno
FK2 CodAnoLectivo
FK1 CodDisciplina
DataPedido
DataConclusao
Escola
UnivInst
Curso
Parecer
Nota
Observações
Criado Por
Criado Em
Modificado Por
Modificado Em
InscricaoAnual
PK,FK1 NumAluno
PK,FK2 CodAnoLectivo
TipoAlunoSAS
AnoCurricular
DataInscricao
Aproveitamento
Observações
Criado Por
Criado Em
Modificado Por
Modificado Em
DocEntregue
PK,FK2 NumAluno
PK,FK2 CodAnoLectivo
PK,FK1 CodDocumento
Documento
PK CodDocumento
Designacao
RegEstatAluno
PK,FK1 NumAluno
PK,FK1 CodAnoLectivo
PK,FK2 CodRegEstat
MesAnoAtribEstat
Observacoes
RegimEstatutario
PK CodRegEstat
Semestre
Designacao
AnoLectivo
PK CodAnoLectivo
AnoLectivo
Inicio Semestre Impar
Inicio Semestre Par
Fim Semestre Par
Prazo
PK CodPrazo
Designacao
Descricao
Tarefa
Serviço
PrazoAno
PK,FK2 CodPrazo
PK,FK1 CodAnoLectivo
DataInicio
DataFim
InscricaoDisciplina
PK,FK2 NumAluno
PK,FK2 CodAnoLectivo
PK,FK1 CodDisciplina
PK,FK3 IdTipoInsc
PK SemAInsc
DataInscricao
CodFunc
DataAvaliacao
Nota
EpocaAval
Status
NumInsc
Observacoes
Criado Por
Criado Em
Modificado Por
Modificado Em
TiposInscricaoDisciplina
PK IdTipoInsc
Descricao
Designação
Entra na média
Altera o numinsc
Limite inscrições Por Semestre
Limite de inscrições por Ano Lectivo
Pautas
PK,FK1 CodAnoLectivo
PK,FK1 CodDisciplina
PK,FK1 Semestre
PK IdPauta
Data
Controlo
Criado Por
Criado Em
Modificado Por
Modificado Em
FK2 IdTipoInsc
Responsavel
PK Semestre
PK,FK1 CodDisciplina
PK,FK2 CodAnoLectivo
CodFunc

Vendas e transacções de produtos e serviços