Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação

OPEN SOURCE BUSINESS INTELLIGENCE

Filipe Manuel Ferreirinho Nunes

Tese submetida como requisito parcial para obtenção do grau de

Mestre em Sistemas Integrados de Apoio à Decisão

Orientador:
Prof. Doutor Mário Romão, Professor Auxiliar,
ISCTE-IUL

Co-orientador:
Prof. Doutor Paulo Trezentos, Professor Auxiliar,
ISCTE-IUL

Fevereiro, 2012

- Lombada -

OPEN SOURCE BUSINESS INTELLIGENCE

Filipe Manuel Ferreirinho Nunes

fevereiro

2012

Este documento foi criado com recurso ao pacote de produtividade OpenOffice/LibreOffice Writer correndo sobre um sistema Linux.

Resumo

A emergência de pacotes de ferramentas de *open source business intelligence* pode constituir uma alternativa mais barata que as soluções comerciais disponíveis, mantendo a fiabilidade e funcionalidade, principalmente num contexto de crise económica e financeira.

O *open source* surgiu com os primeiros computadores. O termo *open source* não significa *software* gratuito, mas *software* livre, na medida em que o seu código pode ser livremente acedido, modificado e redistribuído, ao abrigo de licenças *open source*.

As ferramentas de *business intelligence* constituem uma classe de aplicações cujo fim é a construção de sistemas de exploração analítica dos dados consolidados da organização, usando tecnologias como OLAP, relatórios e *data mining.*

Existem aplicações *open source* fiáveis, utilizadas generalizadamente em todas as áreas da indústria. O servidor HTTP mais usado do mundo, o Apache, é *open source*, tal como o segundo sistema operativo, o Linux. As mais diversas organizações usam aplicações críticas de negócio suportadas em plataformas *open source*.

No *business intelligence*, a utilização de ferramentas *open source* já se faz há algum tempo, embora esteja ainda no início em Portugal. O Pentaho BI suite é uma dos mais populares a nível mundial.

Para demonstrar a viabilidade deste tipo de soluções em ambientes empresariais em Portugal, foram utilizadas as ferramentas da Pentaho para desenvolvimento de um protótipo, tendo por base o sistema de Anúncios SAPO da PT Comunicações. Este permite concluir pela sua viabilidade, constituindo uma opção com um nível de funcionalidade equivalente ao das soluções tradicionais e com um nível de custos mais baixo.

Palavras chave: *Open Source*, Business Intelligence, Sistemas de Apoio à Decisão, Data Warehousing, OLAP, Data Mining, ETL.

Classificação ACM: H.4.2. INFORMATION SYSTEMS APPLICATIONS - Decision support
K.5.1. LEGAL ASPECTS OF COMPUTING - Licensing

Abstract

The emergence of open source business intelligence tools suites can configure itself as a cheaper alternative, with equivalent reliability and functionality levels, compared to available commercial solutions, specially in a context dominated by economical and financial crisis.

Open source software appeared with the first computers. The “open source” designation doesn't necessarily mean free (as in 'free of charge') software, but software that can be freely accessed, modified and redistributed, under open source licenses.

Business intelligence tools are a class of applications whose purpose is to build systems that allow the organization to analytically explore its consolidated data, using technologies as OLAP, reporting and data mining.

There are reliable, widely used, open source applications in every area of the industry. The most used HTTP server in the world, Apache, is open source, as is the second operating system, Linux. The most diverse organizations use critical business applications built on open source platforms.

In business intelligence, there is significant use of open source tools for some time now, though it's just starting in Portugal. Pentaho BI Suite is one of the most popular suites world wide.

To demonstrate the feasibility of this kind of solution in corporate environments in Portugal, a prototype was built using Pentaho tools, having the PT Comunicações' “Anúncios SAPO” (SAPO Advertisements) system as source. The prototype allows for the conclusion of the solution's feasibility, configuring an option with an equivalent level of functionality and a lower cost level compared to traditional solutions.

Keywords: Open Source, Business Intelligence, Decision Support Systems, Data Warehousing, OLAP, Data Mining, ETL.

ACM Classification: H.4.2. INFORMATION SYSTEMS APPLICATIONS - Decision Support
K.5.1. LEGAL ASPECTS OF COMPUTING - Licensing

Agradecimentos

Este trabalho constitui o final de um trajeto altamente compensador, a nível académico, pessoal e profissional, que não teria sido possível sem a intervenção de algumas pessoas e instituições em particular.

Agradeço em primeiro lugar à saudosa Professora Doutora Maria José Trigueiros, que permanece para mim uma inspiração de força, caráter e dignidade. A sua partida deixou um enorme vazio impossível de preencher.

De forma particular, agradeço aos meus orientadores, Professor Doutor Mário Romão e Professor Doutor Paulo Trezentos, pelas portas que abriram, por todo o apoio que prestaram e pela paciência que comigo tiveram.

Agradeço ainda à PT Comunicações, em particular, a Gonçalo Costa e João Pedro Gonçalves, por terem criado as condições para a implementação do protótipo. Estou também grato à equipa da Caixa Mágica Software na PT Comunicações, em especial ao Miguel Filipe, pela colaboração prestada e disponibilidade evidenciada.

Um obrigado especial também a todos os meus colegas de mestrado, com destaque para o Pedro Perfeito, a Carla Gomes, o João Oliveira, o Henrique Ribeiro e a Magda Vasconcelos, por tudo o que com eles aprendi. O convívio e trabalho com estas pessoas singulares contribuiu em muito para o meu enriquecimento pessoal e o alargar de horizontes profissionais. Obrigado também a todos os docentes, que sempre prestaram todo o apoio necessário, muito para além das simples obrigações profissionais.

Também a todos os amigos e amigas que sempre mostraram interesse por este projeto e que de alguma forma contribuíram para que chegasse a bom porto.

Por último, agradeço aos meus pais, que nos momentos mais difíceis, prestaram um apoio insubstituível; e a Deus, o princípio e o fim de todas as coisas.

Índice

Índice de Figuras

Índice de Tabelas

Abreviaturas

AT&T – American Telephone and Telegraph
BI – Business Intelligence
BIND - Berkeley Internet Name Domain
BIRT – Business Intelligence and Reporting Tools
BSD – Berkeley Software Distribution
CEO – Chief Executive Officer
CERN - Centro Europeu de Investigação Nuclear
CPC – Custo Por Clique
CPM – Cost Per Mili
CRM – Customer Relationship Management
CSV – Comma Separated Values
CTR – Click-Through Rate
DARPA – Defense Advanced Research Projects Agency
DNS – Domain Name System
ECPM - Effective Cost Per thousand iMpressions
ERP – Enterprise Resource Planning
ETL – Extraction, Transformation and Loading
FLOSS – Free/Libre and Open Source Software
FSF – Free Software Foundation
FTP – File Transfer Protocol
GNU – GNU's Not Unix
GPL – General Public Licence
HTML – HyperText Markup Language
HTTP – HyperText Transfer Protocol
IBM – International Business Machines
IIS – Internet Information Services
JDBC – Java DataBase Connectivity
JSP – Java Server Pages
LGPL – Lesser General Public Licence
MDX – MultiDimensional eXpressions
MIT – Massachussets Institute of Technology
NASA – National Aeronautics and Space Administration
ODBC – Open DataBase Connectivity
OLAP – OnLine Analitical Processing
RPM – Revenue Per Mil
SaS – *Software as a Service*
SCD – Slowly Changing Dimension(s)
SGBD – Sistema de Gestão de Bases de Dados
SINFIC - Sistemas de Informação Industriais e Consultoria, SA
SQL – Structured Query Language
TB - TeraByte(s)
TCO – Total Cost of Ownership
UCB – Universidade da Califórnia em Berkeley
UNU-MERIT – Universidade das Nações Unidas
WWW – World Wide Web
XML - EXtensible Markup Language

1 Introdução

"Linux is subversive."

Eric Raymond

1.1 Enquadramento e Motivação do Problema

É geralmente aceite que, num contexto de crise económica, exacerbada por uma crise financeira sem precedentes no mundo globalizado, as organizações retraem e adiam os seus investimentos em projetos de tecnologias de informação na tentativa de reduzir os seus custos, reduzindo-os ao essencial [HEN08][SMI08][IMH08].

A avaliação, seleção, compra, implementação e manutenção de tecnologias de *Business Intelligence*[1] (BI) adequadas às necessidades da organização são atividades tradicionalmente demoradas e caras, o que afasta as organizações deste tipo de projetos no contexto atual, já que estes exigem investimentos avultados [PEN07].

No entanto, permanecem muitas dúvidas e obstáculos que têm levado a resistências na adoção de soluções *open source* BI [VEN08]. Deste modo, a principal motivação para a realização deste trabalho é verificar até que ponto estas dúvidas e questões colocam reais obstáculos à adoção deste tipo de soluções, ou se elas constituem de facto alternativas viáveis para a construção de sistemas BI fiáveis, funcionais e com custos substancialmente mais baixos.

1.2 O Problema

Neste contexto, deverão as organizações adiar os seus projetos de BI até que o ambiente económico recupere? Ou, antes, adotar soluções inovadoras que reduzam drasticamente os custos destes projetos?

O sucesso deste tipo de solução poderá contribuir para o aparecimento de um novo segmento de mercado de ferramentas de BI *open source* em Portugal, que poderia contribuir para satisfazer as necessidades de muitas pequenas e médias organizações que, pelo seu custo, não

1 Business Intelligence – utilização de dados validados, baseados em informação factual, através de ferramentas analíticas de reporting, no sentido de apoiar e melhorar o processo de tomada de decisão, otimizando a gestão da organização [NIC08].

têm hoje acesso a soluções de BI. Coloca-se, no entanto, a questão da viabilidade dos atores neste segmento de mercado, uma vez que não teriam, em princípio, acesso a uma das principais fontes de receita dos vendedores de soluções proprietárias – as licenças de *software*, que, sendo *open source*, é livre e "grátis".

Como anteriormente referido, ainda permanecem dúvidas relativamente ao desenvolvimento deste tipo de soluções, nomeadamente, relativamente ao nível de funcionalidade das ferramentas de BI *open source* e à sua facilidade de implementação, bem como à existência de serviços de apoio técnico e funcional fiáveis e de qualidade.

O problema pode, assim, ser enunciado da seguinte forma:

"Constituem as soluções *open source Business Intelligence* alternativas viáveis, com custos inferiores e qualidade semelhante às soluções de Business Intelligence tradicionais, tanto para grandes organizações como para empresas mais pequenas que não têm normalmente acesso às soluções proprietárias?"

1.3 Objetivos da Dissertação

O principal objetivo deste trabalho é investigar a possibilidade de racionalização dos custos de um projeto de *Business Intelligence* utilizando soluções baseadas em software *open source*, mantendo o nível de funcionalidade e com associação de serviços suporte técnico de qualidade adequados às necessidades das organizações, tal como afirmam as organizações promotoras deste tipo de ferramentas. Este objetivo assume uma importância crucial no contexto económico e financeiro atual, mas não perde importância em períodos de expansão.

Outro objetivo essencial é investigar as vantagens e inconvenientes para as organizações em usar estas soluções, para além dos eventuais benefícios financeiros, ao nível das suas funcionalidades, suporte, fiabilidade e simplicidade.

É também propósito desta tese investigar a viabilidade deste tipo de solução num contexto empresarial, através da construção de um protótipo na área das telecomunicações. Por outro lado, o número de soluções "comerciais" de BI *open source* está em desenvolvimento na Europa e no mundo, de que são exemplos a Pentaho, JasperSoft, BIRT, SpagoBI, Talend, Palo e LogiXML [BIT08]. É, no entanto, de assinalar também o desaparecimento de alguns projetos *open source* BI, como o BEE, Open Intelligence, Open Report e RLIB [BIT08]. É

objetivo deste trabalho investigar o modelo de negócio das empresas produtoras de soluções *open source*.

As questões e objetivos de investigação encontram-se sumarizados na tabela seguinte:

	Questões de Investigação	***Objetivos de Investigação***
1.	Porque motivos poderão as organizações adotar uma solução BI open source em detrimento das soluções proprietárias?	Determinar as vantagens das soluções BI *open source* em relação às soluções tradicionais.
2.	Porque motivos não são estas soluções adaptados com maior frequência?	Determinar os riscos acrescidos e objeções, se existirem, inerentes às soluções de BI *open source*.
3.	Como se organizam, trabalham e obtêm receitas as empresas produtoras de soluções de BI open source?	Determinar o modelo de negócio das empresas produtoras de soluções de BI *open source*.
4.	Como justificar a implementação de uma solução deste tipo numa grande empresa portuguesa?	Implementar um protótipo com base em ferramentas de BI *open source* numa grande empresa portuguesa no sentido de demonstrar ou refutar a sua viabilidade.

Tabela 1: Questões e Objetivos de Investigação

1.4 Metodologia

No desenvolvimento deste trabalho foi utilizada a metodologia de estudo de caso. Robson define estudo de caso como "estratégia de investigação que envolve uma investigação empírica de um fenómeno contemporâneo em particular no âmbito do seu contexto real utilizando múltiplas fontes de informação" [SAU09]. Esta estratégia de investigação tem uma considerável capacidade de responder a questões do tipo "Porquê?", "O quê?" e "Como?" [SAU09]. Tendo em conta as questões e objetivos de investigação atrás enunciados e as recomendações de investigações equivalentes esta abordagem de investigação parece ser adequada.

Para o desenvolvimento do protótipo foi utilizada a metodologia "Business Dimensional Lifecycle" de Kimball, cujos principais passos se encontram ilustrados na Figura 1. Como cada projeto é um caso distinto, a metodologia foi adaptada ao protótipo a desenvolver em concreto. Tratando-se de um protótipo, nem todas as fases foram aplicadas com a mesma profundidade, tendo mesmo algumas sido eliminadas.

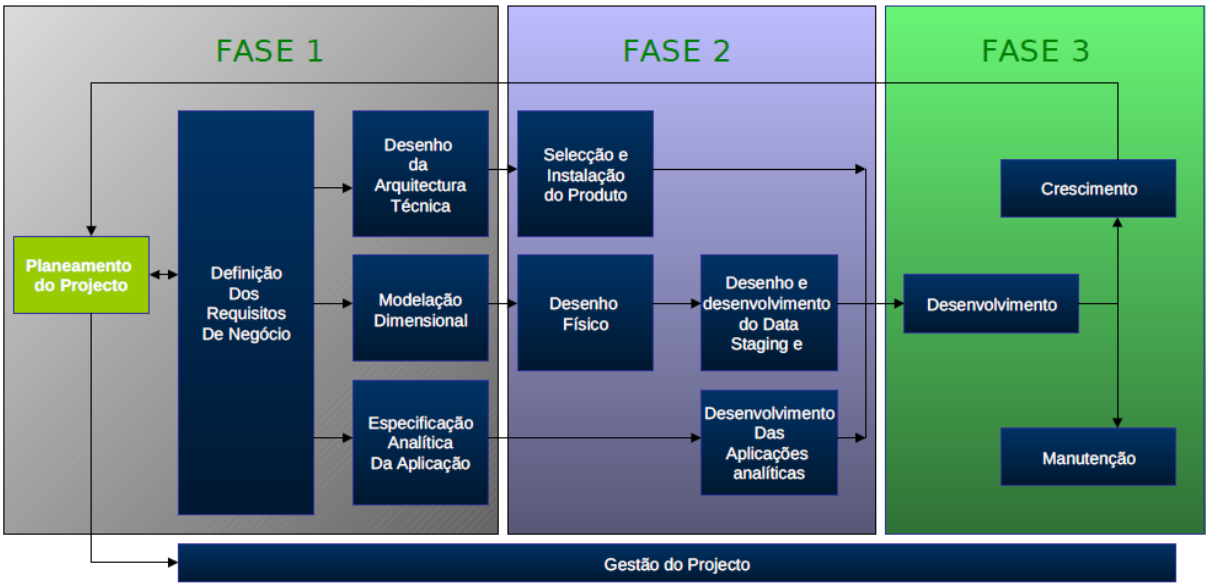


Figura 1: Ciclo de Vida do Projecto [KIM08]

A figura 1 mostra graficamente a sequência de tarefas de alto nível necessárias, segundo [KIM08], para um desenho, desenvolvimento e implementação eficazes do sistema de BI. O diagrama mostra o mapa geral do projeto, em que cada caixa identifica um ponto de controlo.

Em termos genéricos, a Fase 1 corresponde à fase de planeamento e análise e ao conjunto de tarefas que lhes estão associadas, a Fase 2 corresponde à fase de desenho e desenvolvimento a Fase 3 às tarefas de manutenção e crescimento.

Pode-se encontrar uma descrição detalhada desta metodologia em [KIM08].

1.5 Estrutura da Dissertação

Nesta secção apresenta-se sucintamente a estrutura da dissertação.

Assim, no capítulo 1 - Introdução é feito o enquadramento do problema a investigar, são descritos os objetivos da dissertação e é referida a metodologia adaptado para procurar alcançar esses objetivos.

No capítulo 2 - Estado da Arte estabelecem-se os principais conceitos abordados na dissertação, nomeadamente, *Business Intelligence* e *software open source*. É dado um enfoque particular à temática do *open source*, suas origens e evolução, barreiras à adoção e principais aplicações. São também brevemente descritas as principais ferramentas *open source* que podem estar envolvidas na construção de um sistema BI. Por último, são sucintamente

descritos os principais modelos de negócio das empresas que desenvolvem este tipo de *software* e é feita uma investigação de alto nível acerca da sustentabilidade destas empresas.

O capítulo 3 - Prova de Conceito relata o desenvolvimento do protótipo nas suas diversas fases, iniciando-se pela descrição de alto nível da área de negócio Anúncios SAPO, onde este se insere.

Finalmente, o capítulo 4 - Conclusão enuncia as principais conclusões que se retiram do trabalho desenvolvido, tendo como referência os objetivos de investigação inicialmente estipulados.

2 Estado da Arte

Este capítulo descreve o *software* livre atualmente mais relevante, bem como os conceitos que lhe estão inerentes e a sua origem. É também feita uma pequena comparação com as tradicionais soluções proprietárias em termos de custo e funcionalidade. São também abordadas as principais ferramentas de *software* livre existentes no mercado na área de *Business Intelligence*, como não poderia deixar de ser, bem como algumas das suas referências.

2.1 Business Intelligence

Um dos ativos mais importantes de qualquer organização é a sua informação. Esta é mantida quase sempre sob duas formas. Os sistemas operacionais e o *data warehouse*. De uma forma simplista, os sistemas operacionais são aqueles em que a informação é registada e o *data warehouse* é o repositório que suporta a exploração da informação através de ferramentas de BI. É hoje amplamente reconhecido que estas aplicações analíticas têm características muito distintas das dos sistemas operacionais [KIM08].

Assim, Kimball [KIM08] define os seguintes objetivos e características principais do *data warehouse*, a base do BI:

- Tornar a informação da organização rapidamente acessível, compreensível e navegável;
- Tornar a informação da organização consistente e comparável entre diferentes unidades organizacionais;
- Constituir uma fonte de informação adaptável e fiável;
- Garantir segurança no acesso à informação;
- Constituir a fundação dos processos de tomada de decisão.

Os sistemas de BI combinam, assim, a informação do *data warehouse* com ferramentas analíticas no sentido de apresentar informação complexa de forma intuitiva aos responsáveis por processos de tomada de decisão. O BI envolve processos de recolha de dados (a partir dos sistemas operacionais ou outros), a transformação e carregamento desses dados no *data*

warehouse, e a disponibilização de ferramentas analíticas para explorar eficientemente estes dados, com o objetivo de fornecer informação relevante aos decisores no tempo, local e forma adequados [NEG04].

2.2 Definições e História do Movimento open source

As licenças *open source* prometem a todos aquilo que a comunidade designa por liberdade [ROS04]. No que concerne ao *software*, esta liberdade manifesta-se de quatro formas [FSF04]:

1. Liberdade para correr o *software* para qualquer propósito;
2. Liberdade de estudar a forma como o *software* funciona e adaptá-lo às necessidade dos utilizadores;
3. Liberdade de redistribuir cópias do *software;*
4. Liberdade de melhorar o *software* e disponibilizar estas melhorias ao público.

Para um programador, a liberdade para estudar e modificar um *software* implica, necessariamente, aceder ao código fonte (*source code*). Desta forma, o código fonte de uma aplicação tem que estar aberto (*open*) e disponível a todos os interessados, de forma a que a aplicação possa ser estudada, alterada e melhorada [ROS04].

Mas *open source* não significa apenas o acesso ao código fonte. A Open Source Initiative definiu um conjunto de critérios a que uma peça de *software* se deverá conformar para que seja considerada de código aberto [STL04]:

1. Livre Redistribuição – A licença não pode impedir ninguém de redistribuir o *software*, gratuitamente ou não;
2. Código Fonte – O *software* tem de incluir o código fonte, ou, caso contrário, garantir uma forma alternativa de acesso ao mesmo, sem custos, ou apenas com o custo do suporte utilizado, a todos os interessados;
3. Produtos Derivados – O *software* tem que permitir o desenvolvimento de alterações e produtos derivados, bem como a sua distribuição nos mesmos termos que o *software* original;

4. Integridade do Código Fonte de Autor – A licença pode obrigar a que o código original seja distribuído separadamente das alterações efetuadas, no sentido de proteger a integridade do trabalho do autor do *software;*
5. Não Discriminação de Pessoas ou Grupos – A licença não pode discriminar quaisquer pessoas ou grupos de pessoas;
6. Não Discriminação de Áreas de Aplicação – A licença não pode restringir a utilização do *software* a determinado domínio de aplicação, por exemplo, empresas ou áreas de conhecimento;
7. Distribuição da Licença – A licença tem que garantir que as pessoas ou organizações a quem o *software* é redistribuído têm os mesmos direitos inerentes ao *software* original;
8. A Licença Não Pode Ser Específica de um Produto – Os direitos inerentes a um determinado *software* não podem depender de fazer ou não parte de uma distribuição particular. Caso o programa seja extraído da distribuição e redistribuído individualmente, sê-lo-á nos termos da licença da distribuição original;
9. A Licença Não Pode Restringir Outro *Software* – A licença não pode impor restrições ao licenciamento de outras aplicações distribuídas em conjunto com o *software*. Por exemplo, a licença não pode impor que todas as aplicações distribuídas num determinado suporte (CD ou DVD) sejam *open source*;
10. A Licença Tem Que Ser Tecnologicamente Neutra – Nenhum aspeto da licença pode ser prejudicado pela utilização de determinada tecnologia ou interface.

Para visualizar a definição completa, consultar o Anexo A – Open Source Definition.

Rosen (2004) aponta algumas ambiguidades nesta definição, pelo que estabelece um conjunto de princípios com ela consistentes, mas mais facilmente compreensíveis [ROS04]:

1. Os utilizadores são livres de utilizar o *software* para qualquer fim, qualquer que ele seja;
2. Os utilizadores são livres de efetuar cópias do *software* e de as distribuir sem quaisquer encargos;
3. Os utilizadores são livres de criar produtos derivados do *software* e de os distribuir

sem o pagamento de quaisquer encargos;

4. Os utilizadores são livres de aceder e usar o código fonte do *software*;
5. Os utilizadores são livres de combinar o *software open source* com outros tipos de *software*.

Assim, conclui-se que o objetivo deste tipo de *software* é a liberdade, tal como expressa nos pontos mencionados pela Free Software Foundation em [FSF04], o que é conseguido, de um ponto de vista técnico, pelas licenças *open source* [ROS04].

2.2.1 Origens

As origens do *open source* remontam às origens do próprio *software.* Antes dos anos 70, quase todo o *software* era "aberto" ou "livre". Nascera na comunidade científica, onde a informação era, em termos gerais, partilhada, e ninguém tinha pensado numa alternativa. O *software* era desenvolvido pelos fabricantes de hardware, ou pelos seus utilizadores, com o objetivo de satisfazer uma necessidade específica, e era livremente distribuído por grupos de utilizadores ou pelos fabricantes de computadores a outros grupos de utilizadores que pudessem ter a mesma necessidade. Uma vez que o *software* apenas corria nos sistemas de cada fabricante, um bom produto de *software* podia ajudar a vender o sistema [KAV04].

Nos anos 70 nasceu o primeiro sistema operativo moderno, o UNIX, pela mão da AT&T. Dado tratar-se de uma empresa de telecomunicações que operava em regime de monopólio, estava proibida por decreto de desenvolver atividades comerciais noutros setores. Deste modo, a AT&T estava inibida de explorar comercialmente o UNIX, licenciando-o, por um dólar, a várias universidades e outras entidades [MEE08].

Quando o decreto que proibia a AT&T de explorar comercialmente o UNIX cessou a sua vigência, a comunidade científica tratava-o como um projeto de investigação, não como um produto comercial, e partilhava livremente as melhorias e inovações sobre ele desenvolvidas entre si. Assim, o UNIX começou a ser explorado comercialmente, quer pela AT&T, quer por outras empresas, como a IBM e a Sun, o que levou ao aparecimento de várias versões comerciais do UNIX, e impediu que as inovações e melhorias fossem partilhadas [MEE08]. No final dos anos 80, quase todos os fabricantes de hardware tinham a sua versão de UNIX [KAV04]. O movimento do *software* livre surgiu como resposta a esta "privatização" do

UNIX [MEE08].

A comunidade científica sentia assim uma forte necessidade de ter sistemas operativos disponíveis sob a forma de código fonte. A palavra “livre” refere-se, assim, à disponibilidade do código fonte, não ao preço [MEE08].

Entretanto, na Universidade da Califórnia em Berkeley (UCB), em meados dos anos 70, estava em utilização o UNIX Versão 4. Em 1977, foi compilada a primeira Berkeley Software Distribution, constituída por um conjunto de melhorias e inovações desenvolvidas por engenheiros da Universidade, seguida de distribuições mais ou menos anuais. De certa forma, a responsabilidade pela evolução do sistema UNIX passou para a UCB de 1980 a 1995, com o financiamento da Defense Advanced Research Projects Agency (DARPA) . No entanto, cada utilizador final tinha de obter a sua licença da AT&T [KAV04].

Em 1989, o código relativo a redes de comunicação foi disponibilizado em separado ao abrigo da licença Berkeley, que impunha muito poucas restrições. Assim, parte do código original da AT&T podia ser substituído pelo código BSD e “liberto” da necessidade de uma licença da AT&T. Em 1991, foi disponibilizado o 386/BSD ao abrigo da licença Berkeley, substituindo a totalidade do código original e eliminando a necessidade de uma licença da AT&T. O 386/BSD deu origem a três versões, o NetBSD, o OpenBSD e o FreeBSD, este último ainda em utilização, direcionado para PC's e para uma base de utilizadores semelhantes à do Linux [KAV04].

Na Europa, várias pessoas tentaram desenvolver sistemas operativos abertos compatíveis com o UNIX. Um deles foi o MINIX, da autoria de Andrew Tanenbaum, na Universidade Vrije de Amesterdão, preenchendo assim o vazio deixado na comunidade académica. Outro, que viria a ser muito mais famoso, foi o Linux, desenvolvido por um jovem programador de Helsínquia, Linus Torvalds, em 1991 [MEE08].

Entretanto, o projeto GNU (“GNU's not UNIX”) nasceu com o objetivo de construir uma alternativa livre ao UNIX. A sua missão era o desenvolvimento de um sistema operativo completo, enquanto que o Linux era apenas um *kernel.* O GNU defrontava-se com sérios problemas no desenvolvimento do seu *kernel,* e o Linux veio a ser a peça final no puzzle, tendo a FSF convencido Torvalds a disponibilizá-lo nos termos hoje conhecidos por General Public Licence, uma das licenças *open source* mais comuns nos dias de hoje. Assim, o sistema

que hoje conhecemos como Linux resulta desta combinação GNU/Linux [MEE08].

A recessão económica do início da década de 90 levou a que a indústria de tecnologias de informação necessitasse de reduzir drasticamente os seus custos, e muitos programadores desempregados precisavam de se ocupar, o que levou a que o *open source* florescesse e que a popularidade do Linux crescesse exponencialmente, especialmente na Europa [MEE08].

Outro facto que levou ao grande crescimento do movimento *open source* foi o advento da Internet.

A Internet é surpreendentemente antiga, remontando ao final dos anos 60. Em 1969 existiam nós da ARPANET, a precursora da Internet, em Stanford e em Los Angeles. Em 1972 nascia o protocolo FTP e foi inventado o correio eletrónico em Paris. O Sendmail, ainda hoje, o programa mais usado para encaminhamento de correio eletrónico, é *open source* e foi escrito em 1975 por Eric Allman em Berkeley. A Internet passou a usar o protocolo TCP/IP em 1981, altura em que foi escrito o programa *open source* BIND (Berkeley Internet Name Domain) que implementa o protocolo DNS (Domain Name System) e que corre em 95% dos servidores web, sendo o programa mais usado na Internet ainda hoje. Desde 1981 até aos anos 90, o número de utilizadores da Internet duplicou todos os anos [KAV04].

Em 1991, nasceram o HTML e a World Wide Web (WWW), desenvolvidos no CERN (Centro Europeu de Investigação Nuclear). Em 1993 deu-se o aparecimento do primeiro *browser web*, o Mosaic, e em 1995 nasceu o primeiro servidor *web*, o Apache, ainda hoje o mais utilizado, resolvendo assim problemas de dificuldade de utilização que impediam a expansão da Internet [KAV04].

A Internet é a *"killer application"* do *open source*, tendo sido responsável pelo uso generalizado deste tipo de aplicações. Os seus componentes mais importantes são *open source* [KAV04].

Conclui-se, assim, que o *open source* não é propriamente uma novidade, confundindo-se as suas origens com as do próprio *software* em geral [MEE08].

2.2.2 Implicações

De acordo com Kavenaugh [KAV04], o movimento *open source* está a ter profundas implicações ao nível da indústria do *software*, nomeadamente, grades reduções nos custos de:

- Sistemas operativos;
- Servidores *Web*;
- Sistemas de gestão de bases de dados;
- Aplicações de escritório;
- *Browsers web*;
- Variado *software* pessoal e profissional.

É também possível que os baixos preços do software '*core*' leve ao aumento da procura de aplicações mais especializadas do topo da pilha e serviços associados. Este tem sido o comportamento de diversas indústrias ao longo dos anos; por exemplo, o valor da carroçaria dos automóveis tem caído face ao ao valor dos restantes componentes [KAV04].

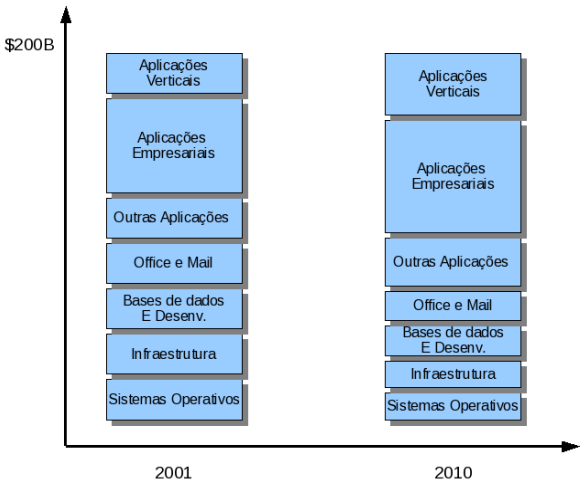


Figura 2: Subida do Valor na Pilha de Software [KAV04]

A Figura 2 mostra a evolução prevista da estrutura da indústria de *software* segundo Ray Lane, ex-CEO da Oracle, de 2001 a 2010. Em 2001, a indústria valia cerca de 200 mil milhões de dólares, prevendo-se que pudesse chegar a 400 mil milhões em 2010. No entanto, dado o estoirar da bolha das empresas “.com” e a ascensão do software *open source*, entre outros fatores, previa-se então que o valor da indústria permanecesse em níveis equivalentes aos do início da década. O que se alterou foi a proporção relativa de cada componente da pilha, ocorrendo uma redistribuição do valor dos componentes mais abaixo para os componentes mais acima. Isto deve-se essencialmente à adoção de *software open source* nos componentes mais abaixo. À medida que os componentes mais básicos vão embaratecendo, existem mais recursos para investir nos componentes mais acima, refletindo-se esse facto na

estrutura da indústria [KAV04].

2.3 Vantagens Open Source

Kavanaugh [KAV04] define as seguintes possibilidades inerentes à utilização de soluções *open source*:

- Acesso ao código fonte;
- Alteração e redistribuição do código fonte;
- Opção de compra de diferentes fornecedores e adoção de diferentes plataformas;
- Utilização de formatos abertos, evitando formatos proprietários;
- Integração entre diferentes ferramentas;
- Redução do custo e esforço de licenciamento;
- Desenvolvimento internacional eficaz;
- Escolha entre um grande grupo de profissionais qualificados.

Porque são estas vantagens importantes? Em primeiro lugar, relativamente ao acesso ao código, a sua não acessibilidade pode tornar difícil e burocrática a correção de bugs. Mesmo os serviços de suporte da empresa que produz ou implementa o *software* podem não ter acesso total ao código, o que não é tão raro como se pensa. Ainda que apenas seja necessária a consulta de manuais, é provável que estes estejam mais atualizados e sejam mais precisos que os de *softwares* fechados, uma vez que o código está acessível à comunidade [KAV04].

Para além de “ver” o código fonte, existem também várias razões para que se possa também alterá-lo e até redistribuí-lo. Desde logo porque é necessário que o código esteja atualizado para que valha a pena vê-lo. Em segundo lugar, o facto de os utilizadores poderem contribuir tem comprovadamente sido um ponto forte dos grandes projetos *open source*.

Por último, com a possibilidade de alterar e redistribuir o código, é também possível a criação de um produto concorrente do original com base neste. Esta é, aliás, a última linha de defesa do mundo *open source*, na medida em que, caso os *owners* de um determinado projeto deixem, de alguma forma, de se interessar pelo mesmo, um grupo concorrente pode continuar

onde o primeiro parou, e continuar a dar suporte aos utilizadores atuais [KAV04]. Foi o que se passou recentemente com o OpenOffice. Após a compra da Sun pela Oracle, a Document Foundation deu origem a uma divisão daquele projeto, nascendo o LibreOffice.

É necessário efetuar um planeamento cuidadoso para salvaguardar a possibilidade de adquirir *software* de diferentes vendedores, o que é mais seguro que confiar numa única plataforma proprietária. A liberdade de escolha reduz os custos, e o passado tem demonstrado que é boa ideia estar-se preparado para migrar os sistemas da organização para plataformas diferentes. Os sistemas *open source* são comprovadamente portáveis entre plataformas. Em primeiro lugar, porque o código fonte está acessível, bem como as ferramentas para o compilar/interpretar em diferentes plataformas. Em segundo lugar, porque existem versões de vários sistemas para diferentes plataformas, quando isso faz sentido [KAV04]. Por exemplo, várias aplicações disponíveis para sistemas Linux têm também versões equivalentes para sistemas Windows, como sejam o OpenOffice/LibreOffice ou o Gimp.

O *software* fechado implementa normalmente formatos proprietários, não os publica e altera-os frequentemente. Os exemplos mais comuns são os antigos formatos de ficheiros do Microsoft Office, mas este facto também se verifica em sistemas como SAP e PeopleSoft, entre outros. Muitos dos documentos que criamos podem permanecer válidos durante anos, podendo mesmo ser necessário convertê-los várias vezes para formatos diferentes, por exemplo, para os publicar na Web ou visualizá-los num PDA [KAV04].

A disponibilidade do código-fonte e o carácter público dos formatos de dados faz com que seja sempre possível, com maior ou menor dificuldade, fazer com que diferentes aplicações possam funcionar de forma integrada. Muitas vezes, existem problemas de integração mesmo com aplicações do mesmo fornecedor. Por exemplo, a Microsoft utiliza diferentes estruturas de armazenamento de dados incompatíveis entre si em diversos dos seus produtos (base de dados Jet, SQL Server, “Metabase” do IIS, *registry* do Windows, ficheiros XML numa grande variedade de formatos, entre outros). No mundo *open source*, com formatos abertos, estes problemas são menos frequentes, e, mesmo que ocorram, a organização pode, ela própria, realizar as alterações necessárias, ao invés de estar dependente de terceiros para o fazer [KAV04].

O *software open source* tem, em quase todos os casos, menos custos que o *software*

proprietário. É possível encontrar situações em que tal não ocorre, mas elas são raras e baseiam-se, normalmente, em pressupostos não realistas. Por exemplo, numa comparação entre Linux e um outro sistema operativo proprietário, o sistema Linux pode ser baseado em *hardware* mais caro, usar serviços e ferramentas de desenvolvimento mais caras e inflacionar os custos com formação. Tudo isto já foi feito em estudos publicados. É certo que existem algumas categorias de custos que favorecem os incumbentes, como os custos com formação e contratação. Mas, à medida que o *open source* se desenvolve, estes custos tendem a baixar [KAV04].

É necessário notar, no entanto, que qualquer análise de retorno do investimento baseada simplesmente em custos é sempre falível. Como já foi referido, basta que os pressupostos se alterem para que surjam diferenças assinaláveis, além de que os benefícios intangíveis são muitas vezes subvalorizados ou não considerados [GUL05].

Também a gestão das licenças do software proprietário pode representar um acréscimo de custos administrativos. Muitas organizações compram mais licenças que as necessárias devido às dificuldades de gestão do uso concorrente (o chamado *shelfware*). Outra vantagem é o controlo da organização sobre as atualizações do software, que não fica dependente do calendário dos fornecedores [KAV04]. Por definição, estas questões não se colocam quando se utiliza software *open source*.

Para organizações multinacionais ou de mercados mais pequenos, a utilização de *software* localizado pode ser uma questão importante. A decisão de desenvolver uma versão localizada de aplicações proprietárias pode não ser viável se o mercado não o justificar. Já com o software *open source*, este pode ser traduzido para uma determinada língua por recursos locais mesmo que isso não seja economicamente viável de um ponto de vista central [KAV04].

Também a grande disponibilidade de profissionais qualificados na área de Linux não constituirá um obstáculo à sua adoção, uma vez que, em termos funcionais, o Linux é idêntico ao UNIX e foi nele inspirado [KAV04]. No universo Business Intelligence português, no entanto, quase não existem referências *open source*, o que pode constituir um obstáculo inicial à adoção destas soluções.

Têm, no entanto, sido publicados diversos anúncios de recrutamento pela SINFIC, que atua

nos mercados português, angolano e moçambicano, contando com cerca de 500 colaboradores e apresentando um volume de negócios de cerca de 14,5 milhões de euros em 2009 [SIN10], solicitando profissionais com competências em Pentaho [SIF10]. Também a XpandIT apresenta no seu sítio *web (www.xpand-it.com)* uma referência de BI para a Brisa usando Pentaho.

Inmon [INM07], pioneiro nas soluções de *Business Intelligence*, define ainda um conjunto de vantagens específicas de ferramentas ETL[2] *open source*, mas que pode ser generalizado para o universo das ferramentas Business Intelligence *open source*:

- Não existem barreiras à adoção: qualquer pessoa pode descarregar a solução da Internet, experimentá-la e começar a implementá-la nos seus projetos – tudo sem o controlo dos vendedores;
- Os custos iniciais e totais são baixos;
- O pagamento, quando existe, é baseado na utilização real, sem *shelfware;*
- Não existem custos de runtime;
- Grande oferta de especialistas no mercado, sem necessidade de recorrer aos serviços de consultoria dos vendedores.

O próprio processo de avaliação e seleção das ferramentas é mais facilitado levando em consideração soluções *open source*. A Pentaho, uma das empresas líderes em soluções *open source* BI, identifica as seguintes [PEN07]:

- Investigação inicial: tipicamente, antes de serem contactados vendedores, a organização tenta recolher informação através de artigos, brochuras e conferências, entre outros. A avaliação de ferramentas *open source* distingue-se por um muito maior nível de acesso a informação relevante, de transparência e de abertura, bem como o acesso a um produto totalmente funcional, para além dos recursos tradicionalmente disponíveis;
- Filtragem de informação: em fases mais avançadas no processo de avaliação, o acesso

2 ETL – Extraction, Transformation and Loading. Ferramentas de extração, transformação (limpeza e integração) e carregamento de dados, utilizadas sobretudo em sistemas Business Intelligence e de integração de dados.

à informação, a versões de demonstração com funcionalidades limitadas, a experiências de clientes, entre outros fatores importantes, é mediado e filtrado pelos vendedores. Isto acontece, não por eventual desonestidade dos vendedores, mas porque os seus interesses não estão perfeitamente alinhados com os da organização, na medida em que, apesar do seu interesse no sucesso do seu potencial cliente, o principal objetivo nesta fase é, legitimamente, "fechar uma venda". Nas soluções *open source* o controlo que os vendedores podem exercer sobre o acesso à informação é limitado, uma vez que a organização tem, no limite, acesso ao próprio código fonte;

- Seleção: Para além das funcionalidades das ferramentas, que a não ser em situações específicas, não são frequentemente determinantes no sucesso de um projeto (há alguns projetos que falham, mas muito poucos por falta de funcionalidades específicas), existem muitos fatores a ter em conta na fase da seleção. A Pentaho apresenta como determinante *o total cost of ownership* (TCO), que é substancialmente mais baixo nas soluções *open source*. Outro facto a ter em conta é a disponibilidade de serviços de suporte, que a maioria das soluções disponibiliza, diretamente ou através de parceiros, através de subscrições, tal como os vendedores tradicionais.

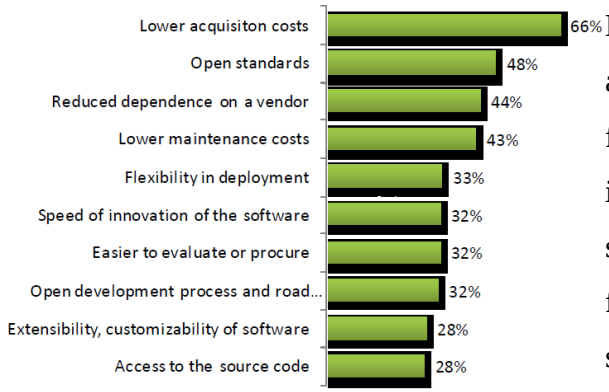


Figura 3: Razões para a utilização de ferramentas open source *[MAD09]*

Deste modo, mesmo que a organização venha a optar por uma solução proprietária, pode fazê-lo de uma forma mais consciente e informada, pela simples consideração de soluções *open source* [PEN07], para além do facto de que, pela sua simples existência, as soluções *open source* BI, contribuem para a descida dos preços das soluções tradicionais [VES07].

Em jeito de resumo, um estudo efetuado pela BeyerNETWORK [MAD09] indica que as principais razões que levam as organizações a adotar soluções *open source* são as que constam da figura 3.

2.3.1 Impacto Económico

As vantagens do código aberto identificadas para as empresas em particular refletem-se na economia dos diferentes países em termos mais gerais. No estudo "Economic Impact of FLOSS on Innovation and Competitiveness of the EU ICT Sector", encomendado pela Comissão Europeia à Universidade das Nações Unidas (UNU-MERIT), chegam-se a algumas conclusões importantes [GHO06]. Desde logo, uma larga maioria das empresas europeias usa ou pensa usar software FLOSS[3].

O desenvolvimento, controlo de qualidade e distribuição das aplicações livres e *open source* (FLOSS) de qualidade em uso na Europa custariam doze mil milhões de euros às empresas para reproduzir internamente. Esta base de código tem duplicado em cada dezoito a vinte e quatro meses, e este crescimento deverá continuar por vários anos [GHO06].

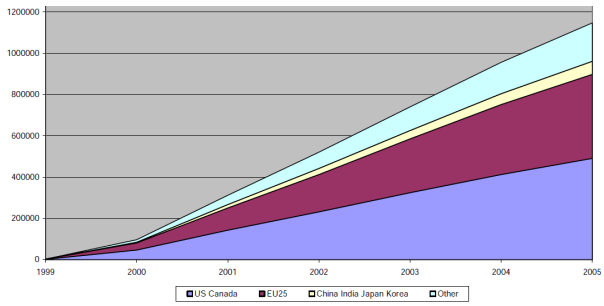


Figura 4: Participantes em Projetos FLOSS por Região [GHO06]

Esta base de software FLOSS representa cerca de 131 000 horas-homem de esforço dedicado exclusivamente por programadores. Uma vez que este esforço de desenvolvimento não é diretamente pago, representa cerca de 800 milhões de Euros em trabalho voluntário por ano, dos quais cerca de metade ficam na Europa [GHO06].

As empresas investiram cerca de 1,2 biliões de Euros no desenvolvimento de aplicações FLOSS, que foram depois disponibilizados livremente. Estas empresas representam pelo menos 565 000 postos de trabalho e 263 mil milhões de Euros em volume de negócios [GHO06]. Segundo previsões feitas em 2006, no mercado de tecnologias de informação, os serviços relacionados com *software* FLOSS deverão atingir 32% em 2010, e 4% do PIB

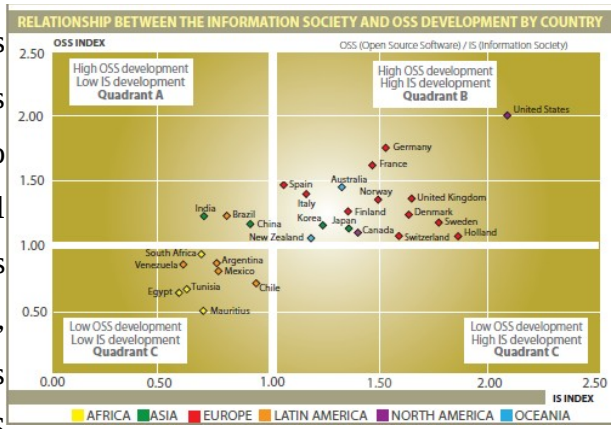


Figura 5: Relação entre a Sociedade de Informação e o Desenvolvimento de Software Open Source *por País* [CEN10]

3 FLOSS – Free/Libre and *open source* software – software Livre e *open source*

europeu. O *software* FLOSS suporta diretamente uma quota de 29% do *software* desenvolvido na União Europeia (43% nos Estados Unidos) [GHO06].

As empresas produtoras de *software* proprietário empregam menos de 10% dos programadores americanos, o que sugere um grande potencial de crescimento do emprego relacionado com desenvolvimento de *software* FLOSS [GHO06].

Estima-se que o desenvolvimento de *software open source* possa representar uma poupança de cerca de 36% em despesas de investigação e desenvolvimento na produção de *software*, que se pode traduzir em maiores lucros ou investimento noutras linhas de investigação [CEN10].

É interessante também verificar a relação entre o grau de desenvolvimento económico de um país ou território e o seu grau de adoção de *software* livre. Num estudo publicado em 2010 pelo Observatório Nacional de *Software Open Source* espanhol, conclui-se que são os países com as economias mais fortes e com maior grau de desenvolvimento na área das tecnologias de informação que lideram o desenvolvimento e adoção de *Software Open Source*, nomeadamente, os Estados Unidos a Austrália e a Europa Ocidental. Verifica-se ainda que países com economias emergentes, como o Brasil, a China e a Índia, têm índices de adoção de *Software* livre superiores ao esperado [CEN10].

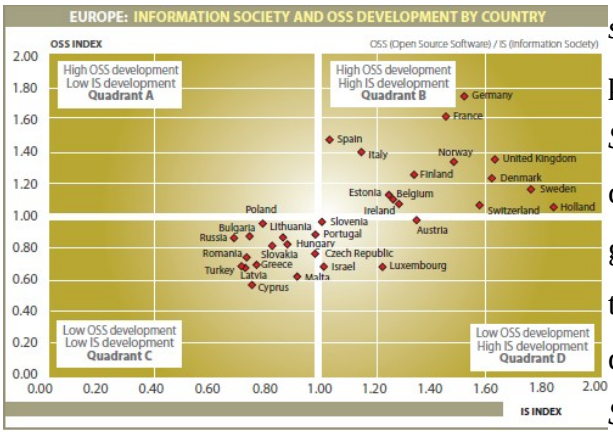


Figura 6: Europa: Sociedade de Informação e Desenvolvimento OSS por País [CEN10]

Na figura 6 podemos observar a mesma relação apenas para a Europa, que mostra Portugal numa posição média em ambos os índices, o que leva à conclusão que, também na área das Tecnologias de Informação, temos ainda um caminho a percorrer para chegar ao nível dos nossos principais parceiros europeus.

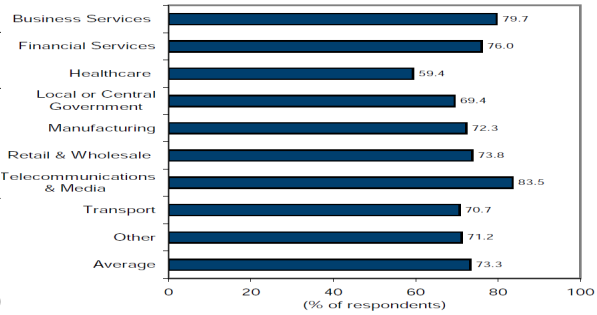


Figura 7: Utilização de FLOSS na Europa por Sector de Atividade [GHO06]

Através da observação da figura 7, verifica-se que todos os setores de atividade estudados em [GHO06] apresentam níveis de adoção superiores a 50%, com destaque para os sectores financeiro, empresarial e das telecomunicações.

Desde 2006, data em que estes dados foram publicados, o *software open source* tem aumentado a sua penetração no sector público na Europa, liderada neste esforço pela Alemanha, pela França e pela Espanha. A própria União Europeia tem contribuído neste sentido, através do financiamento de projetos tecnológicos que resultam em desenvolvimentos *open source*, envolvendo empresas e universidades de diversos países [CEN10].

2.4 Mitos

Apesar de todas as vantagens já mencionadas relativamente às soluções abertas, existem ainda alguns preconceitos relativamente a estas que constituem verdadeiras barreiras à sua adoção O Gartner Group [DRI01] identifica os seguintes como sendo os principais:

- O *software open source* é totalmente grátis – O *software* em si é, de facto, muitas vezes grátis, embora nem sempre seja esse o caso. No entanto, ao avaliar uma solução aberta, há outros custos a ter em conta, nomeadamente, os custos de suporte, documentação e formação;
- É uma moda passageira. É contra a natureza humana trabalhar gratuitamente – Como já vimos, o *open source* existe praticamente desde que existem computadores, embora não fosse reconhecido como tal. Com esta designação, existe há mais de 20 anos e é utilizado de forma generalizada;
- Ninguém coordena o desenvolvimento – Na verdade, a maioria do *software open source* é rigorosamente controlado por um reduzido número de colaboradores chave;
- Não existe a quem recorrer para suporte – Muitos vendedores de versões comerciais licenciadas (o chamado *open source* comercial) suportam diretamente as suas soluções. Além disso, existe um sem número de empresas que prestam suporte às mais variadas soluções, e é sempre possível recorrer a uma vasta comunidade de utilizadores e participantes no desenvolvimento, embora nem sempre este tipo de suporte seja aceitável pelas organizações, que procuram, não só suporte, mas

responsabilidade [LEY00];

- Qualquer pessoa pode alterar o *software* , que acaba por se tornar instável e pouco seguro – Este mito está relacionado com a coordenação dos projetos *open source* que, como já referimos, é restrita e rigorosa;
- Quando o principal impulsionador sai do projeto, este morre – A maioria dos projetos *open source* são extraordinariamente resistentes a este tipo de fenómenos, continuando a ser suportados até muito depois dos líderes originais terem saído;
- Os projetos *open source* acabam por se dividir, tal como o UNIX – Os sistemas *open source* são a prova de que as leis de seleção natural se aplicam ao *software*.

A Ingres Corporation, que distribui o SGBD *open source* 'Ingres' identifica, adicionalmente, os seguintes mitos, que é necessário desfazer [MCG08]:

- O *open source* é um nicho do mercado de tecnologias de informação – Desde o aparecimento do Linux e do Apache, o *open source* tem vindo a tomar uma forte posição no mercado;
- As ferramentas *open source* não podem suportar aplicações críticas – Muitos opositores do movimento *open source* têm alimentado esta ideia, bem como a de que os produtos *open source* não têm qualidade. No entanto, os produtos *open source* estão sujeitos aos mesmos testes de desempenho, stress, funcionais, de segurança e de regressão que os produtos fechados. Para além disso, o modelo de desenvolvimento *open source* permite o envolvimento dos utilizadores na fase inicial do ciclo de desenvolvimento, o que leva a que erros e falhas de desenho sejam identificados mais cedo;
- As empresas *open source* não controlam a sua propriedade intelectual - A verdade é que o *software open source* está sujeito às mesmas leis de direitos de autor que o *software* fechado, mas os fornecedores de *software open source*, por opção, partilham esses direitos com outros, o que resulta num desenvolvimento mais rápido e evita a monopolização dos produtos. Nenhuma organização pode controlar o preço dos serviços de suporte, e a concorrência para fornecer serviços de suporte a um preço atrativo reduz os custos para os clientes;

- A tecnologia *open source* não é suportada a nível profissional – Hoje em dia, as empresas produtoras de soluções *open source* oferecem serviços profissionais de suporte e consultoria, possibilitando que o *software* corra aplicações críticas em grandes empresas. Na verdade, o modelo de negócio da maioria das empresas produtoras de software *open source* depende da aquisição de serviços de suporte e consultoria;
- O *software open source* não é seguro – Trata-se de um mal-entendido comum considerar-se que o *software open source* é mais vulnerável a ataques que o *software* fechado, simplesmente porque é mais provável que o código seja adulterado. Na verdade, o *software open source* é seguro e utiliza normas em termos de metodologias de desenvolvimento e técnicas de segurança. Todas as alterações ao código são sujeitas a uma avaliação rigorosa pela comunidade antes de serem aceites. Deste modo, a deteção de vulnerabilidades é mais pro-ativa, expondo quaisquer adulterações do código, e disponibilizando mais segurança desde o início do desenvolvimento.

Conclui-se, assim, que o *software open source* constitui hoje uma alternativa viável para grandes e pequenas organizações [MCG08].

2.5 Barreiras à Adoção

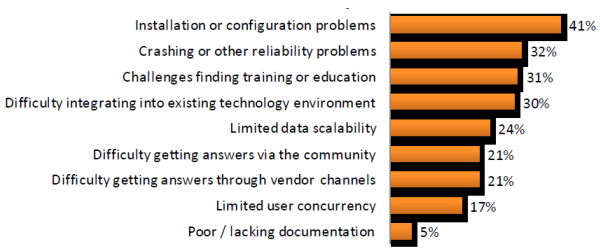


Figura 8: Problemas comuns com software open source [MAD09]

Para além de alguns mitos que persistem em relação ao *software open source*, que podem constituir barreiras à adoção, é necessário referir também alguns problemas reais que podem impedir a sua adoção pelas organizações.

Num estudo efetuado pela BeyerNETWORK [MAD09], 47% das organizações respondentes que efetivamente avaliaram aplicações *open source* nos seus processos de seleção de ferramentas, referiram ter tido algum tipo de problema.

As dificuldades de instalação e configuração foram a maior dificuldade reportada neste estudo, o que, na maior parte dos casos, está relacionado com a natureza modular do *software*

open source. O segundo maior problema reportado está relacionado com a fiabilidade das ferramentas [MAD09]. Estas conclusões estão sumarizadas na figura 8.

Problemas de instalação, configuração, fiabilidade e escalabilidade revelam que algumas soluções ainda não atingiram a maturidade necessária à sua utilização em grandes projetos.

É de realçar, no entanto, que muitas soluções proprietárias têm também alguns destes problemas.

2.6 Licenças Open Source

Todas as licenças *open source*, por definição, permitem ao utilizador o exercício de todos os direitos de *copyright* no que respeita ao *software* licenciado. Existem, no entanto, dois tipos básicos de licenças; as hereditárias e as permissivas. A principal diferença é o facto das licenças hereditárias imporem restrições significativas ao exercício de certos direitos – normalmente, o de redistribuição, na medida em que toda a redistribuição pelo utilizador terá de ser feita ao abrigo da licença original [MEE08].

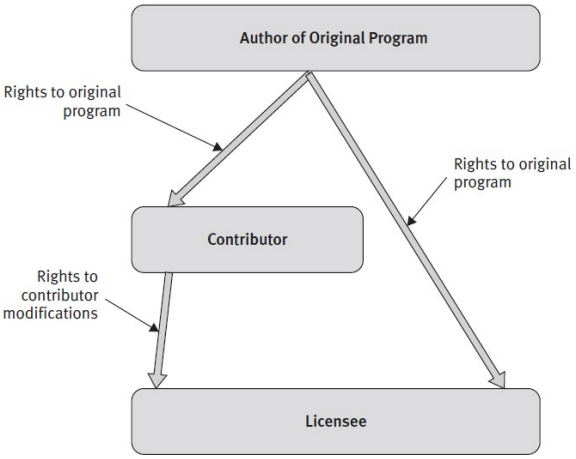


Figura 9: Modelo de Licenciamento Direto [MEE08]

Um dos conceitos mais difíceis de apreender em matéria de licenciamento *open source* é o facto de constituir um modelo de licenciamento direto. A maioria das licenças *open source* são permissões concedidas diretamente pelo autor a quem quer que deseje utilizar o *software* ao abrigo da licença. Se o utilizador contribuir com o seu próprio código para o *software*, pode, ele próprio, licenciá-lo diretamente nos mesmos termos. Deste modo, em dados módulos, os direitos referentes a diferentes porções de código podem provir de diferentes autores [MEE08].

Existem ainda algumas licenças proprietárias que, não sendo *open source*, incorporam alguns dos princípios deste movimento, merecendo, por isso, referência [STL04]:

- Sun Community Source License – desenvolvida pela Sun para incorporar alguns dos

benefícios *open source* em dois dos seus produtos proprietários – Jini e Java;

- Microsoft Shared Source Initiative – resposta da Microsoft ao movimento *open source*. Através deste esquema de licenciamento, a Microsoft desenvolveu um conjunto complexo de licenças que permitem o acesso ao código fonte de forma mais ou menos restrita, baseando-se numa escala com cinco atributos chave:
  - A possibilidade de ver e referenciar o código sem o alterar;
  - A possibilidade de melhorar o debugging com acesso ao código fonte;
  - A possibilidade de alterar o código apenas para uso local;
  - A possibilidade de distribuir produtos baseados em código alterado para fins não comerciais;
  - A possibilidade de comercializar produtos baseados em código fonte modificado.

Para uma lista de licenças *open source* mais completa, consultar o Anexo B – Licenças open source Mais Comuns.

2.7 Open Source Business Intelligence

No sentido de estudar as ferramentas *open source* atualmente existentes para BI, este trabalho adota a arquitetura de referência da figura 10, resultante da adaptação e simplificação da arquitetura proposta por Kimball em [KIM08].

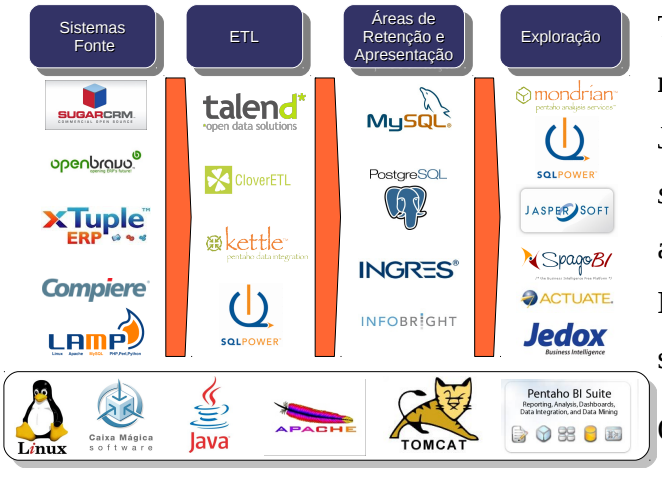


Figura 10: Arquitetura de referência para open source Business Intelligence

Todos os componentes desta arquitetura de referência são *open source*, à exceção do Java, que, não sendo totalmente *open source*, é de utilização livre, distribuído ao abrigo da licença Sun Community Source Licence, em termos relativamente semelhantes ao *software open source*.

Os componentes mencionados na figura não pretendem ser uma listagem exaustiva de todas as ferramentas *open source* que

podem ser incorporadas numa solução deste tipo. Poderão mesmo existir arquiteturas híbridas, em que alguns componentes são abertos e outros não, sendo esta uma situação comum.

Em todo o caso, é possível desde já concluir que, de um ponto de vista teórico, é possível construir um sistema de *Business Intelligence* usando exclusivamente componentes *open source*, incluindo os sistemas fonte, como mostra a figura e veremos com mais detalhe nas secções seguintes.

2.7.1 Infraestrutura

Dependendo do sistema a desenvolver, os componentes necessários para a infraestrutura, em termos de sistema operativo, servidores *web* e servidores aplicacionais, entre outros, podem variar. Neste estudo, consideram-se os seguintes (*open source* ou similar):

- Linux (genérico);
- Caixa Mágica (distribuição portuguesa de Linux);
- FreeBSD;
- Servidor *web* Apache;
- Servidor Aplicacional Tomcat.

Consideram-se ainda como parte da Infraestrutura os sistemas de bases de dados relacionais:

- MySQL;
- PostgresSQL;
- Ingres;
- InfoBright.

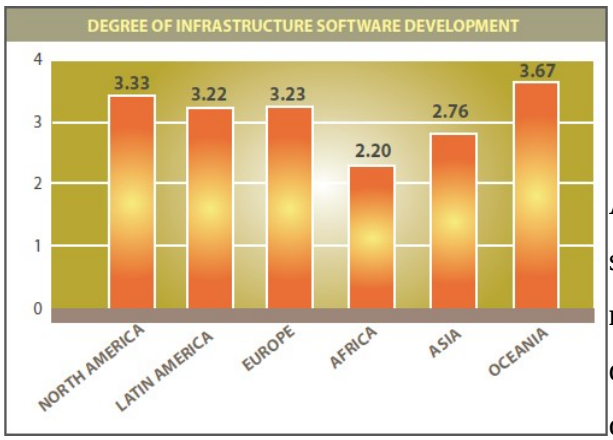


Figura 11: Grau de Desenvolvimento do Software *de Infraestrutura [CEN10]*

A figura 11 mostra o grau de penetração do *software open source* (numa escala de 1 a 5) nesta categoria, mais elevado que em qualquer outra [CEN10], o que valida a conclusão enunciada em 2.2.2 - Implicações, em que se sugere que, devido à redução de

custos do *software* de infraestrutura resultante da adoção crescente de soluções *open source* a este nível, as organizações orientam as poupanças que daí resultam para *software e serviços* mais especializados.

2.7.1.1 Linux

O Linux é o segundo sistema operativo mais usado em servidores em todo o mundo, a seguir ao Windows [KAV04]. Pode ser encontrado em data centers por todo o mundo [NIC08]. Segundo a IDC, em 2002, a Microsoft vendeu 2.533.671 servidores Windows, o que representa um crescimento de 14% face a 2001. No mesmo ano, também segundo a IDC, foram instalados 485.679 servidores Linux, cerca de 20% da quota do Windows, mas com um crescimento de 35% em relação ao ano anterior [KAV04].

Num estudo da Forrester Research na Europa e América do Norte, 26% das empresas inquiridas declararam utilizar de alguma forma o Linux como sistema operativo desktop, enquanto 76% referem que o utilizam como servidor, o que representa um impacto significativo no mercado de sistemas operativos servidores [FOR07].

Apesar da utilização bastante menos expressiva como sistema operativo desktop, é importante referir que o Linux é o segundo sistema operativo cliente mais utilizado a nível mundial, tendo destronado o Mac OS em 2003 [KAV04].

2.7.1.2 Linux Caixa Mágica

O Linux Caixa Mágica é uma distribuição portuguesa do sistema operativo Linux gerida pela Caixa Mágica Software e baseada na distribuição franco-brasileira Mandriva Linux. Dirige-se às empresas, particulares, educação e administração pública. Foi criado em Outubro de 2000 com a atribuição do Prémio Milénio Expresso 2000 a três investigadores do ISCTE que colaboravam na Associação para o Desenvolvimento das Telecomunicações e Técnicas de Informática (ADETTI): Daniel Neves, José Guimarães e Paulo Trezentos [WIK09].

Organismos e empresas que utilizam totalmente ou parcialmente o Linux Caixa Mágica [WIK09]:

- Câmara Municipal de Barcelos;
- J. P. Sá Couto – incluído no portátil “Magalhães”, ao abrigo da iniciativa 'e-escolinhas'

do Ministério da Educação;

- Ministério da Educação, em 13 mil computadores nas escolas secundárias, em dual boot com o Windows;
- Linius – Versão do Caixa Mágica personalizada para o Ministério da Justiça;
- Inforlândia – portáteis InSys distribuídos ao abrigo do programa e-escolas, do Ministério da Educação.

2.7.1.3 FreeBSD

No mundo *open source*, a palavra "Linux" é quase sinónima de "sistema operativo", mas este não é o único sistema operativo aberto baseado no UNIX. De acordo com o Internet Operating System Counter, em Abril de 1999 14,6% dos servidores *web* corriam FreeBSD. Algumas das empresas mais ativas na Internet, como a Yahoo!, usam FreeBSD [LEH09].

O FreeBSD, cujas origens já analisámos, não é de uso tão generalizado como o Linux, havendo menos produtos disponíveis para esta plataforma, apesar do FreeBSD possuir um pacote de compatibilidade com o Linux [LEH09]. No entanto, as aplicações servidoras mais importantes, como o Apache e vários motores de bases de dados estão disponíveis, constituindo assim uma boa opção para um servidor dedicado [KAV04].

Embora o FreeBSD seja o mais conhecido, existem outros projetos BSD, com diferentes objetivos [LEH09]:

- FreeBSD – alto desempenho e facilidade de utilização:
- NetBSD – portabilidade;
- OpenBSD – segurança e fiabilidade;
- DragonFlyBSD – desempenho e escalabilidade *(clustering).*

Como sistema cliente, é sobretudo usado como componente do Mac OS X, que é o terceiro sistema operativo desktop mais utilizado, a seguir ao Windows e ao Linux. Ao longo dos últimos anos, o sistema operativo da Macintosh tem sido reescrito, e o seu *kernel* é baseado no FreeBSD. No entanto, o Mac OS X não é *open source*, à exceção do seu *kernel.* O interface gráfico e API's são proprietários, assim como a maioria das aplicações para ele

desenvolvidas [KAV04].

2.7.1.4 Apache

O Apache é um dos projetos *open source* mais bem sucedidos até hoje. É o servidor *web* mais usado em todo o mundo, tendo conquistado uma quota de cerca de dois terços (67%, contra 21% da Microsoft) de todos os servidores web instalados. Corre em múltiplos sistemas operativos, incluindo o Linux, a maioria das versões de UNIX, Windows e Novell NetWare [KAV04].

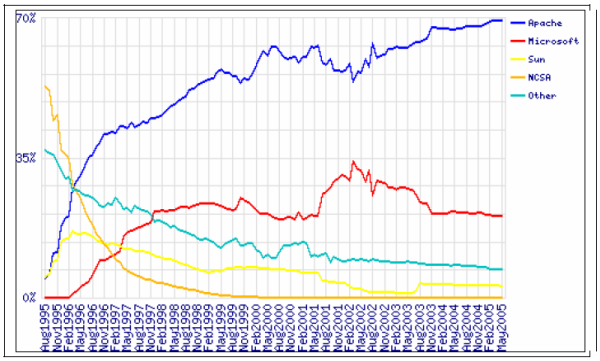


Figura 12: Quota de mercado dos servidores web mais importantes [GHO06]

O Apache é distribuído ao abrigo da licença BSD, tornando-o atrativo tanto para aplicações comercias como não comerciais. A Amazon e o Yahoo! são exemplos de grandes sítios *web* que usam esta aplicação servidora [LEE02].

2.7.1.5 Java, Tomcat e Jboss

É possível construir aplicações simples com Java Server Pages (JSP), a forma mais comum de desenvolvimento de aplicações Java. Aplicações mais complexas necessitam de um servidor aplicacional Java [KAV04].

Para aplicações JSP, a escolha *open source* recai sobre o Jakarta Tomcat. A nível de servidores aplicacionais, o JBoss é a solução *open source* mais usada, com uma quota de cerca de 25% dos sistemas em produção [KAV04].

O servidor aplicacional JBoss e o motor de *servlets* Tomcat são usados em muitas empresas para aplicações operacionais críticas, e existem diversas aplicações comerciais baseadas nestas tecnologias [OPT07].

As aplicações JSP e Java são multi-plataforma, correndo em Linux, Windows, UNIX e outros sistemas operativos, se desenvolvidas e testadas corretamente, sendo esta a sua principal vantagem [KAV04].

2.7.1.6 MySQL

O sistema de gestão de bases de dados (SGBD) MySQL nasceu em 1996 pela mão de três escandinavos; Axmark, Larsson e Widenius. Em 1994-95, com a emergência da WWW, sentiram a necessidade de criar um SGBD especialmente adequado à web [BUR04].

De forma a penetrar rapidamente no mercado, mas, ainda assim, obter receitas com licenças, optaram por um esquema de licenciamento dual como base para o seu modelo de negócio. Assim, o MySQL é distribuído gratuitamente através de uma licença *open source*, ou através de uma licença proprietária de baixo custo. Ao fim de quatro meses após o lançamento, já tinham sido feitos cerca de 1000 downloads a partir do sítio web da empresa. As primeiras licenças comerciais foram vendidas em 1997 [BUR04].

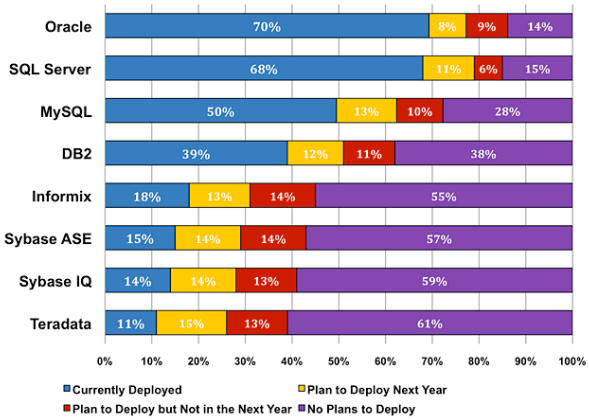


Figura 13: Instalações de bases de dados e planos de desenvolvimento (fonte: Gartner Group 2008, citado em www.mysql.com)

O MySQL, entretanto adquirido pela Sun, por sua vez adquirida pela Oracle, reclama para si o título de base de dados *open source* mais popular do mundo.

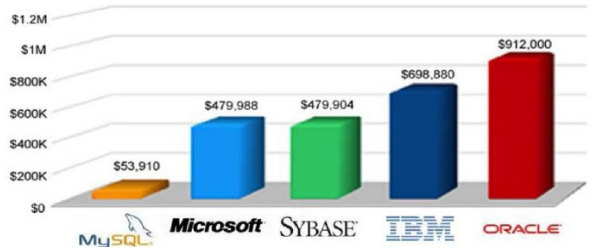


Figura 14: Comparação a 3 anos dos custos de licenciamento dos SGBD mais importantes [MYS09]

O MySQL apresenta como proposta de valor essencial a redução do TCO em mais de 90% em relação aos SGBD tradicionais. O gráfico seguinte compara os custos de três subscrições do MySQL Enterprise Gold, que inclui suporte 24 horas por dia, 7 dias por semana com os custos de produtos equivalentes da Microsoft, Sybase, IBM e Oracle, e é ilustrativo desta redução [MYS09].

Há, no entanto, que ter em atenção que esta informação é fornecida pelo vendedor, e que os custos reais das soluções alternativas variam de conforme os acordos comerciais das respetivas empresas e dos seus clientes. No entanto, a adoção do MySQL em vez de um SGBD proprietário pode representar uma poupança significativa.

O MySQL apresenta ainda algumas características e diferentes “motores de armazenamento” que o tornam adequado para uma solução de Data Warehousing [MYS07]:

- Particionamento de dados e índices;
- Sem limites práticos de armazenamento (1 tablespace tem um limite de 110 TB);
- Tabelas em memória com o motor Memory, ideal para tabelas de dimensão;
- SELECTS e INSERTS rápidos, com o motor MyISAM, não transacional;
- Compressão de dados até 80%, com o motor Archive;
- Dados armazenados por colunas, com o motor BrightHouse, o que aumenta o desempenho e facilita a compressão de dados;
- Suporte multi-plataforma.

Tem como principais referências o Google, Sabre Holdings, Associated Press, Suzuki e a NASA [MYS08].

2.7.1.7 PostgreSQL

Este popular SGBD foi desenvolvido pela Universidade da Califórnia em Berkeley, no Departamento de Ciências Informáticas. Foi pioneiro em muitos conceitos que apenas foram disponibilizadas muito mais tarde em alguns SGBDs comerciais. É distribuído nos termos da licença BSD, o que permite a sua alteração e redistribuição por qualquer entidade, com qualquer objetivo, seja ele privado, comercial ou académico [POS08].

2.7.1.8 INGRES

A Ingres Corporation, que resulta de um *spin-off* da Computer Associates reclama para este SGBD o título de “Leading open source corporate database” [ING08]. O SGBD Ingres foi, até há pouco, um produto comercial muito bem sucedido da Computer Associates, que, como resposta ao fenómeno MySQL, abriu o seu código, tornando-o um produto *open source*. Assim, passou a existir um produto

Figura 15: INGRES - referências mais importantes

open source reconhecido no mercado pelas maiores empresas mundiais [GAR05].

2.7.1.9 Infobright

Figura 16: Infobright - principais referências

A solução BrightHouse Data Warehousing, da Infobright, foi desenhada para lidar com os desafios dos grandes *data warehouses* dos operadores de telecomunicações com um custo baixo sem comprometer o desempenho. A Infobright apresenta as seguintes vantagens para a sua solução [MYS08]:

- Poupança de espaço em disco de 90%;
- 10% do custo total associado a produtos similares;
- Baixos tempos de resposta para *queries ad-hoc* analíticas em comparação com outros produtos;
- Não requer a utilização de esquemas específicos (Ex. Esquema em Estrela);
- Não requer vistas materializadas nem estratégias complexas de particionamento e indexação de dados;
- Reduz os custos iniciais e operacionais do *data warehouse*, reduzindo o número de servidores e o espaço de armazenamento necessários, bem como os custos de manutenção associados;
- Corre em *hardware* standard, de baixo custo;
- Compatível com as mais reconhecidas ferramentas de *Business Intelligence*, como Cognos, Business Objects e Pentaho, entre outros.

O Infobright integra-se com o MySQL, aproveitando as funcionalidades disponibilizadas pelos conectores MySQL (C, JDBC, ODBC, .NET, Perl, etc.). O MySQL disponibiliza também funções de catalogação, tais como definições de tabelas, vistas, utilizadores, permissões, etc, armazenados numa base de dados MyISAM [INF08].

2.7.2 Sistemas Fonte

Nesta secção, abordam-se alguns sistemas baseados em ferramentas *open source* que, tipicamente, constituem fontes de dados para as soluções de Business Intelligence.

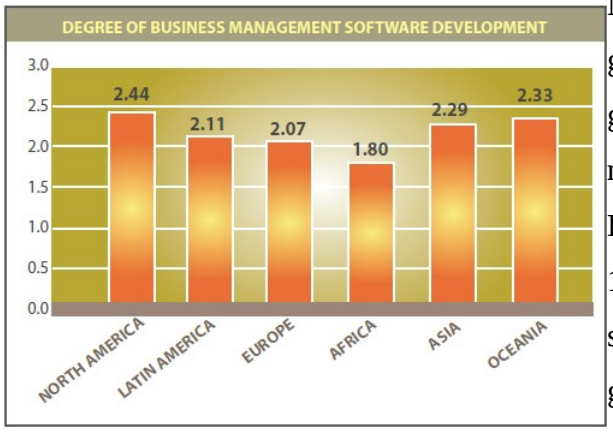


Figura 17: Grau de Desenvolvimento no Software de Gestão [CEN10]

Nesta categoria incluem-se muitas soluções de gestão, como sejam sistemas ERP ou CRM. O grau de penetração do *software open source* nestas soluções é mais baixo, sendo que na Europa é de 2,07 numa escala de 1 a 5 (figura 17). Como se verá nesta secção, existem soluções de gestão *open source* usadas em grandes projetos e suportadas a nível profissional, inclusivamente premiadas, o que sugere a existência de um nicho de mercado a explorar pelas empresas que atuam nesta área.

2.7.2.1 Openbravo

O ERP Openbravo é um produto *open source* funcional e integrado, baseado na *web*, oferecendo às empresas uma solução de baixo custo para otimizar os seus processos de negócio. Inclui também funcionalidades básicas de CRM e Business Intelligence [OPE09].

Figura 18: Prémios angariados pelo Openbravo (www.openbravo.com)

O Openbravo já conquistou diversos prémios, atribuídos por publicações especializadas, de onde se destacam os da figura 18.

2.7.2.2 xTuple

O ERP xTuple, também *open source*, assenta numa solução cliente-servidor, com um cliente com interface gráfico e uma base de dados PostgreSQL. A edição PostBooks é a versão totalmente gratuita e adequada para organizações mais pequenas. As edições Standard e a Manufacturing são licenciadas comercialmente, embora todas as versões sejam desenvolvidas

sobre o mesmo código, diferindo apenas na base de dados. Em todas elas, o cliente tem acesso ao código fonte, quer das aplicações cliente, quer das servidoras [XTU09].

2.7.2.3 Compiere

O Compiere é outro dos ERPs *open source*, que apresenta também funcionalidades de CRM, constituindo uma solução de gestão integrada e adaptável [COM08]. Este produto apresenta como principal inovação o facto de armazenar a lógica de negócio num Dicionário de Aplicações na base de dados, em vez de estar *hard*-coded no *software* [COM09].

2.7.2.4 SugarCRM

O SugarCRM é o gigante na categoria *open source* das soluções CRM. Nascido em 2004 pela mão de John Roberts, Clint Oram e Jacob Taylor, o seu código fonte já foi descarregado mais de três milhões de vezes. Mais de 12.000 empresas usam o SugarCRM, incluindo a Honeywell International, a Starbucks Corporation, o First Federal Bank e a BDO Seidman LLP [HAK08]. Em Portugal, destaca-se a solução implementada pela DRI para o serviço UZO da TMN.

2.7.3 Business Intelligence

Uma das principais razões pela qual as soluções de BI comerciais têm tido dificuldade em captar clientes fora do âmbito das grandes organizações está relacionada com a perceção de exagero no valor dos preços e licenças. Com o aparecimento de soluções *open source* suportadas profissionalmente, com modelos de licenciamento por subscrição, deixam de existir estas barreiras à entrada, embora não se possa afirmar a gratuitidade de qualquer solução, como já se referiu. À medida que que o BI *open source* progride, substituindo mesmo algumas implementações tradicionais, aumentará o número de soluções disponíveis [WIS10].

As soluções *open source* BI ainda têm uma penetração bastante limitada. Em inquéritos realizados pelo Gartner Group realizados em 2008, menos de 2% dos inquiridos mencionaram utilizar algum tipo de produto *open sour*ce BI [BIT08], o que tanto pode ser visto tanto como uma barreira à adoção, tanto como um nicho de mercado com potencial de crescimento.

Nas secções seguintes encontram-se sucintamente descritas as principais soluções *open source Business Intelligence* atualmente disponíveis, que estão referenciadas no relatório da

Gartner “Who’s Who in Open-Source Business Intelligence” [BIT08].

2.7.3.1 Talend Open Studio

Philip Howard, numa análise da Blood Research, afirma, acerca do Talend Open Studio, que esta ferramenta merece ser seriamente considerada mais ou menos em todas as situações [HOW09].

O Open Studio é uma ferramenta de geração de código Java, Perl ou SQL, que pode depois correr-se onde e quando necessário [HOW09].

A Talend é a única empresa *open source* neste mercado com escritórios fora do seu mercado local, isto é, na China, Estados Unidos, França e Alemanha, permitindo a oferta de suporte 24x7 em todos os continentes [HOW09].

2.7.3.2 Pentaho BI Suite

A Pentaho foi fundada em 2004, e disponibiliza uma das mais conhecidas soluções *open source Business Intelligence*, que apresenta um grande leque de funcionalidades, rivalizando com algumas ofertas comerciais como o Business Objects, Cognos ou Oracle [BIT08].

A oferta da Pentaho inclui a Pentaho BI suite, que, por sua vez, inclui funcionalidades analíticas, *dashboard*, *data mining*, OLAP e ETL. Os clientes podem optar pela versão *open source*, sem custos, ou por uma subscrição, disponível em diferentes níveis de serviço. A subscrição possibilita também acesso a funcionalidades não disponíveis na versão *open source* [BIT08].

2.7.3.3 Jaspersoft BI

A JasperSoft é outra marca muito bem estabelecida no mercado *open source* BI. Fundada em 2001, refere que foram feitos três milhões de downloads do seu produto e intitula-se “O líder de mercado do *open source* BI”, com mais de 8000 clientes comerciais. A família de produtos JasperSoft inclui o JasperServer, JasperReports, Jasper Report Designer, JasperAnalysis (servidor OLAP) e o JasperETL, baseado em tecnologia Talend [BIT08].

2.7.3.4 Actuate

A Actuate, fundada em 1993, foi a primeira empresa cotada em bolsa a entrar no mercado

open source BI, em 2004. A 'Business Intelligence and Reporting Tool' (BIRT) é baseada no Eclipse e constitui a base da linha de produtos comerciais Actuate. Através do projeto BIRT, a Actuate disponibiliza uma pequena parte da sua solução de reporting através de uma licença *open source* [BIT08].

2.7.3.5 Outros

Há medida que a comunidade *open source* BI vai crescendo continuamente, mais empresas começam a atuar neste mercado. A Jedox, sedeada na Alemanha disponibiliza o servidor OLAP Palo, e tem tido um crescimento sustentado. A SQL Power Group, de Toronto decidiu recentemente abrir o código das suas soluções BI proprietárias [BIT08].

2.8 Modelos de Negócio

Nesta secção serão abordados alguns dos modelos de negócio e desenvolvimento *open source* mais comuns, com especial enfoque no chamado *open source* comercial, em utilização pela maioria das empresas que produzem e suportam soluções de BI *open source.*

Desde o início do projeto GNU/Linux em 1983, o modelo de desenvolvimento *open source* evoluiu para novas formas de cooperação à volta do conceito mais ou menos vago de uma comunidade de contribuidores. Existem comunidades de pessoas individuais; pequenas, médias e grandes empresas; entidades públicas, universidades e centros tecnológicos e de investigação. Todas partilham o princípio de que o *software open source* é uma estratégia eficaz na melhoria dos processos de investigação, desenvolvimento e inovação, tornando possível o estabelecimento de modelos de negócio viáveis promovendo a colaboração [CEN10].

Esta evolução deu origem ao surgimento de três tipos de comunidades [CEN10]:

- Comunidades *ad-hoc* que se formam para levar a cabo projetos específicos;
- Fundações para projetos maiores que requerem a formalização das regras de funcionamento da comunidade;
- Comunidades suportadas por empresas, individualmente ou em consórcio.

Partindo destas últimas, o Observatório Nacional de *Software Open Source* espanhol

estabelece cinco modelos de negócio [CEN10]:

1. Modelos de subscrição – Distribuição gratuita de *software open source* com a subscrição de serviços associados, pagos, de manutenção e consultoria;
2. Serviços de valor acrescentado - prestados por membros da comunidade com base no conhecimento gerado;
3. *Software* como serviço (SaS) – O cliente usa o *software* remotamente, sem necessidade de instalação nos seus sistemas, pagando apenas pela utilização;
4. Modelo híbrido – O cliente tem acesso a *software* ao abrigo de licenças *open source*, e pode adquirir funcionalidades adicionais ao abrigo de outro tipo de licenças;
5. *Cross-selling* – Distribuição de *software open source* em conjunto com outros produtos.

Muito poucos vendedores usam, no entanto, exatamente as mesmas estratégias de desenvolvimento, licenciamento, vendas e obtenção de receitas [45108].

Assim, o 451 Group, num estudo publicado em 2008, chega às seguintes principais conclusões [45108]:

1. A maioria das empresas utiliza algum tipo de licenciamento comercial de forma a gerar receitas a partir de *software open source*;
2. Uma grande parte dos vendedores combina código *open source* com código desenvolvido fora do âmbito da comunidade nas suas soluções;
3. Os vendedores que utilizam modelos híbridos de desenvolvimento e licenciamento suportam maiores custos de desenvolvimento e *marketing*, mas conseguem equilibrar estes custos com a geração de receitas a partir de *software* licenciado comercialmente;
4. A linha divisória entre *software* livre e proprietário tem vindo a desvanecer-se, à medida que componentes de *software* livre são incorporados em soluções proprietárias, e o *software* livre vai tendo também uma vertente comercial.

As estratégias mais comuns têm como elementos principais um modelo de desenvolvimento híbrido, controlado pelos vendedores; e um modelo de licenciamento *open-core*[4] com o

4 Licenciamento *open-core* - Modelo de licenciamento em que as funcionalidades básicas de um produto são

licenciamento comercial e serviços de subscrição como principais fontes de receitas [45108]. É este o modelo seguido pela maioria das empresas envolvidas na disponibilização de soluções BI *open source*.

2.8.1 Modelos de Desenvolvimento

De acordo com o 451 Group [45108], a maioria dos vendedores de soluções *open source* adota um dos seguintes modelos de desenvolvimento:

- *open source* corporativo - O *software* é distribuído ao abrigo de uma licença *open source*, e todas as contribuições são públicas, mas o desenvolvimento é predominantemente efetuado pelos colaboradores do vendedor;
- *open source* comunitário – O *software* é desenvolvido publicamente por uma comunidade de programadores individuais, da qual faz parte também o vendedor;
- *open source* misto – A solução oferecida pelo vendedor resulta de uma combinação de diferentes projetos *open source* desenvolvidos por várias empresas e comunidades;
- *open source* híbrido – A base da solução é distribuída ao abrigo de uma licença *open source*, mas algumas funcionalidades adicionais são proprietárias.

A maioria dos vendedores investigados usam este modelo, em combinação com o *open source* corporativo. De assinalar que a maioria dos vendedores de soluções de BI *open source* abordados neste trabalho usam também esta última categoria, embora, por exemplo, o MySql seja um puro *open source* corporativo.

2.8.2 Estratégias de Licenciamento

O estudo do 451 Group [45108] identifica as seguintes estratégias de licenciamento utilizadas na distribuição das soluções comerciais *open source*:

- Licenciamento dual – O mesmo produto é distribuído ao abrigo de uma licença *open source* a alguns utilizadores e de uma licença comercial a outros, que podem ter alguma espécie de resistência às licenças *open source*, justificadamente ou não;

fornecidas ao abrigo de licenças *open source* e as funcionalidades adicionais são fornecidas ao abrigo de licenças comerciais [45108].

- Licenciamento *open-core* – O *core* da solução é disponibilizado através de uma licença *open source*, mas existem também versões empresariais ou profissionais em que estão incluídas funcionalidades desenvolvidas em código fechado. Por vezes designado por 'modelo *up-selling*';
- Aberto-e-fechado – Os produtos *open source* são complementados por outros, proprietários, desenvolvidos e comercializados em separado;
- *open source* puro – Toda a solução é disponibilizada ao abrigo de uma licença *open source*;
- *Assembled open source* – A solução utiliza código de diferentes projetos *open source* que utilizam várias licenças diferentes;
- Fechado – O produto é baseado em código *open source*, mas é disponibilizado apenas ao abrigo de licenças comerciais.

Menos de metade dos vendedores utilizam esquemas de licenciamento *open source* puros (41%). As estratégias mais comuns são o '*open-core'* (24%), onde se incluem o Talend, xTuple, Pentaho e Jasper; o 'aberto-e-fechado' (15%), onde encontramos a Actuate; e o licenciamento dual (15%), de que é exemplo o MySQL.

2.8.3 Fontes de Receitas

O estudo do 451 Group [45108] refere as seguintes fontes de receitas para os vendedores de soluções *open source* comerciais:

- Licenças comerciais – licenças não aprovadas pela FSF;
- Subscrições – Contratação periódica de serviços de suporte, que podem também dar acesso a atualizações de produtos e a versões profissionais e empresariais;
- Suporte e serviços – Pedidos pontuais de suporte e outros serviços, como consultoria e formação;
- *Hardware embebido* – O *software* é distribuído em conjunto com um '*appliance*' ou outro tipo de *hardware*;
- *Software as a Service (SaaS)* – Os utilizadores pagam pelo acesso e utilização do

software através da Internet;

- *Software embebido* – O *software open source* é fornecido como parte de um pacote de *software* comercial;
- Publicidade – A utilização do *software* é gratuita, o vendedor obtém receitas através de publicidade associada ao produto;
- Desenvolvimento à medida – Os utilizadores pagam a personalização do *software* às suas necessidades particulares;
- Outros produtos e serviços – O *software open source* não é usado como fonte de receitas. Para isso, são disponibilizados outros produtos e serviços complementares.

Verificou-se no estudo que a maioria das receitas provenientes destas soluções têm origem em licenças comerciais (25%) e em serviços de subscrição (34%). É importante referir que muito poucas empresas deste segmento de mercado dependem de uma única fonte de receitas entre as identificadas, mas a uma combinação de várias fontes.

2.8.4 Sustentabilidade

No sentido de procurar avaliar a sustentabilidade financeira das empresas produtoras de soluções de BI *open source*, chegou-se à conclusão que elas são, na sua maioria, empresas privadas segundo o direito americano, ou seja, são empresas não cotadas que não divulgam as suas demonstrações financeiras. Assim, só indiretamente, através de outros indícios, é possível fazê-lo.

Em relação à JasperSoft e à Pentaho, as empresas que distribuem as soluções de BI *open source* mais populares, existem dados relativos a 2007 e 2008 publicados num estudo da Nine Lives Media [PAN09], que construiu um *ranking* das maiores empresas do universo *open source* com base nas suas relações com os seus parceiros. A JasperSoft e a Pentaho ocupam o quinto e o décimo terceiro lugares, respetivamente A tabela abaixo sumariza os dados disponíveis para estas duas empresas.

	JasperSoft		Pentaho	
	2008	**Δ**	**2008**	**Δ**
Receitas Totais	N/A	75,00%	N/A	N/A
Receitas Provenientes de Parceiros	60,00%	20,00%	30,00%	0,00%
Número de Parceiros	510	45,71%	110	120,00%
Número de Colaboradores	75	25,00%	70	40,00%

Tabela 2: JasperSoft e Pentaho - Principais Indicadores [PAN09]

Estes dados, apesar de escassos, permitem verificar que ambas as empresas se encontram em crescimento.

A estrutura acionista da JasperSoft, segundo consta do seu sítio *web* (*www.jaspersoft.com*) conta com financiadores de capital de risco e algumas empresas de referência, nas quais se incluem:

- Adams Street Partners;
- Doll Capital Management;
- Morganthaler Ventures;
- Partech International;
- Red Hat;
- SAP Ventures;
- Scale Venture Partners.

A JasperSoft apresenta ainda um conjunto de referências importantes, nomeadamente:

Open Source Business Intelligence

Alten Italia	Austin Energy	British Telecom	CardSmith	Cast Consult	Cincom
Disytel	Easypay	Ebuilder	EnergySys	Entrust	Entuity
Genologics	GGB	Handysoft	Hospira MedNet	Monolith Software	Open Mind Networks
Qmetry	Remote Learner	Revol Wireless	SADAD	Saudi Telecom	Spiral Universe
Tata	Telvent	US Naval Safety Center	Tomax	University of East London	Universidade de Valência
Vinteh Business Solutions	Virgin Money	Webteam	WJ Bradley	Yarris	QED Financial

Tabela 3: Principais Referências JasperSoft

A Pentaho apresenta na sua estrutura acionista entidades como as listadas abaixo, obtidas a partir do seu sítio *web* (*www.pentaho.com*):

- Benchmark Capital;
- Index Ventures;
- New Enterprise Associates.

Apresenta como principais referências as seguintes:

Accession Media	Aspiro	BizIntel	BNSF Logistics	Boyne Resorts	Brussels Airport
Cardiac Science	Centro	Close Premium Finance	Cognitive Match	ControlPay	Coopservice
Desjardins	Estalea	Fidelitone	Harris	Lifetime Networks	Loma Linda University
Burough of Camdem	London Oncology Clinic	Mainzer Volksbank	Marketo	Mozilla	MySQL AB
National Health Service Islington	Online Buddies	Otto International	Power Costs	Quartet Financial Systems	Scholarship America
Sheetz	Sircon	Specsavers	Rednano	Spidex Software	Stonegate Senior Caer

Sun Microsystems	Swiss Colony	Swissport	Transporte Medicatel	TV 2 (Noruega)	US Naval Air Systems Command
	Uniloc	Unionfidi	ZipRealty		

Tabela 4: Principais Referências Pentaho

Em resumo, com os dados disponíveis, estas empresas parecem ser sólidas fiáveis, embora tal não se possa afirmar com toda a certeza sem uma análise de dados financeiros que não estão disponíveis.

Tratando-se de soluções *open source*, o papel dos integradores de sistemas assume uma importância crucial, uma vez que são eles quem implementa as soluções para os seus clientes. Entre integradores de sistemas, revendedores e parceiros tecnológicos, a Pentaho apresenta uma extensa lista de parceiros que pode ser consultada em *www.pentaho.com/partners*, que inclui nomes como a Accenture, a HP, a Red Hat e a Sun. Conta mesmo com um parceiro português, a Xpand IT Solutions, que apresenta no seu sítio *web* (*www.xpand-it.com*) um caso de sucesso de implementação de uma solução de BI baseada em Pentaho para a Brisa[5].

Também a JasperSoft apresenta uma rede de parceiros abrangente, que pode ser consultada em *www.jaspersoft.com/partners*, que inclui nomes como a HP, a IBM, a Oracle, a Novell, a Red Hat, a Sun, e a Unisys.

5 A descrição completa deste projeto pode ser encontrada em *www.xpand-it.com/pt/montra-produtos-e-portfolio/16-business-intelligence/122-brisa-maintenance-datawarehouse*

3 Prova de Conceito

No sentido de justificar e demonstrar a viabilidade da adoção de uma solução de BI *open source* numa grande empresa portuguesa, foi realizada uma prova de conceito usando apenas ferramentas *open source* na PT Comunicações, no âmbito do serviço de anúncios online do SAPO.

Todo este serviço é suportado por um sistema operacional construído sobre a plataforma LAMP (Linux, Apache, MySQL, PHP/Perl/Python), constituída apenas por componentes *open source*, o que constitui um bom exemplo de uma aplicação crítica de negócio numa grande empresa construída sobre soluções *open source*.

3.1 Negócio

O serviço Anúncios SAPO é bastante semelhante ao Adwords do Google, mas especialmente adequado ao mercado português. Os potenciais anunciantes registam-se em http://anuncios.sapo.pt, abrindo a sua conta.

Depois de provisionada a conta do anunciante, este poderá criar campanhas publicitárias, que poderão incluir vários grupos de anúncios, que, por sua vez, conterão os anúncios individuais. Cada campanha pode ser segmentada por palavras, em que o anunciante selecionará as palavras que deseja que despoletem a impressão dos seus anúncios com a ajuda de um assistente presente na página, ou por *site*, em que são escolhidos os *sites* em que os anúncios serão impressos, de entre *sites* do SAPO e/ou dos seus parceiros. Pode, adicionalmente, dirigir a campanha aos internautas de um determinado distrito do país.

Posteriormente, o anunciante tem a possibilidade de criar o seu primeiro anúncio para a campanha assim definida, após o que poderá definir um limite diário para o custo da sua campanha, bem como um custo por clique (CPC) máximo.

Para além do CPC, estão ainda definidas, entre outras, as seguintes métricas para medir o desempenho de anúncios, grupos de anúncios e campanhas, bem como a sua rentabilidade:

- Click-Through Rate (CTR) – Número médio de cliques por cada 100 impressões;
- Conversões – Utilizadores que realizam a ação desejada (comprar um produto,

registar-se numa página, etc);

- Cost Per Mili (CPM) – Custo médio por cada 1000 impressões.

As formas mais comuns de compra de anúncios por parte dos anunciantes são a compra por CPC, CPM e CPA (*Cost Per Action*, ou custo por ação, em que o anunciante paga por ações específicas, como uma compra, ou a submissão de um formulário). É de realçar que, no caso dos Anúncios SAPO, apenas é disponibilizada a compra por CPM.

3.2 Planeamento

Nesta fase, é necessário, em primeiro lugar, fazer uma definição cuidada do projeto Para isso, começamos por avaliar o grau de prontidão da organização para a implementação de uma solução de BI [KIM08]. No âmbito desta prova de conceito, considerar-se-á apenas a unidade organizacional Anúncios SAPO na PT Comunicações.

3.2.1 Avaliação do Grau de Prontidão Organizacional

Começou-se pela análise da informação disponibilizada atualmente pela plataforma Anúncios SAPO, bem como a realização de reuniões preparatórias com alguns utilizadores chave da referida informação, técnicos e comerciais. Estas iniciativas tiveram como objetivo, não só avaliar a prontidão da unidade organizacional, mas também efetuar o levantamento de alguns requisitos de negócio a implementar na prova de conceito.

Kimball [KIM08] identifica cinco critérios chave para a avaliação do grau de prontidão organizacional para uma solução de BI:

1. Forte patrocínio executivo – Trata-se do fator mais crítico para o sucesso de qualquer projeto, não constituindo exceção a implementação de uma solução de BI. A presente prova de conceito é patrocinada pelos principais responsáveis técnico e comercial, profissionais respeitados na organização, tendo sempre mostrado a disponibilidade necessária a esclarecer questões técnicas e de negócio, com uma visão realista do âmbito e objetivos da iniciativa. Conclui-se que existe um patrocínio adequado da iniciativa por parte da organização;
2. Motivação de negócio – A gestão dos Anúncios SAPO depende fortemente da tomada de decisões informadas e atempadas. Através das reuniões realizadas, foram

identificadas algumas insuficiências nos processos atuais Nomeadamente, existem relatórios suportados em folhas de cálculo atualizadas manualmente todos os dias. Não existe informação histórica anterior a 2006 fora destes relatórios. Não é possível atualmente identificar os *sites* em que os anúncios são impressos, a não ser que as campanhas sejam segmentadas por *site*. Os utilizadores da área comercial mostraram uma clara necessidade de mais, melhor, e mais atempada informação. Adicionalmente, o custo da iniciativa é virtualmente nulo para a organização, uma vez que se utilizaram exclusivamente ferramentas *open source* e o esforço de implementação não é remunerado, por se tratar de um projeto com uma vertente académica;

3. Aliança entre a equipa técnica e equipa de negócio – O mestrando, que assumiu o papel de equipa técnica nesta iniciativa, teve sempre apoio adequado por parte dos utilizadores de negócio;
4. Cultura analítica – Como já foi referido, os responsáveis do serviço Anúncios SAPO possuem uma forte cultura analítica, tomando decisões o mais informadas possível, confiando em factos e números trabalhados diariamente em folhas de cálculo e evidenciando a necessidade de mais e melhor informação para fundamentar as suas decisões;
5. Viabilidade técnica – Em termos técnicos, foram assumidos alguns riscos, uma vez que toda a prova de conceito foi implementada com recurso a um único computador de secretária, com recursos limitados. A experiência da "equipa técnica" era, também, limitada, resultante, sobretudo, das aulas e projetos de mestrado, e sem um conhecimento profundo da plataforma utilizada. No entanto, dado tratar-se de um projeto com uma vertente académica, é natural que assim seja, sendo esta também uma condição necessária para a concretização dos objetivos da iniciativa.

Pode-se, assim, concluir que o grau de prontidão organizacional do serviço Anúncios SAPO é elevado, havendo, deste ponto de vista, todas as condições para o sucesso da iniciativa.

3.2.2 Levantamento Preliminar de Requisitos de Negócio

Nas reuniões com utilizadores chave, foram identificadas várias questões de negócio a que um sistema de BI deveria ser capaz de responder, relacionadas com as pesquisas efetuadas no

portal SAPO e com os Anúncios SAPO propriamente ditos.

Os utilizadores entrevistados forneceram mesmo uma folha de cálculo à "equipa técnica" com um conjunto de requisitos de negócio para uma eventual solução de BI que viesse a ser implementada na área dos Anúncios SAPO, antes ainda do início da presente iniciativa, o que vem demonstrar o seu forte patrocínio e a sua visão e expectativas claras em relação a uma solução de BI.

Esta folha de cálculo está reproduzida no Anexo C – Requisitos de Negócio Definidos pela PT Comunicações, constituindo esta o conjunto preliminar de requisitos de negócio a considerar.

3.2.3 Definição Preliminar do Âmbito da Prova de Conceito

Analisando as tabelas do Anexo C – Requisitos de Negócio Definidos pela PT Comunicações, podemos identificar três processos de negócio suscetíveis de serem incluídos no âmbito da prova de conceito:

- Pesquisas no portal SAPO;
- Anúncios SAPO;
- Carregamentos de contas SAPO.

É, assim, necessário priorizar estes conjuntos de requisitos. A Sequent Computer Systems, citada por Kimball em [KIM08], propõe uma análise por quadrantes para o fazer, com a viabilidade no eixo horizontal e o impacto para o negócio no eixo vertical. Os processos de negócio prioritários serão aqueles que apresentem, simultaneamente, uma elevada viabilidade e um elevado impacto no negócio (quadrante superior direito).

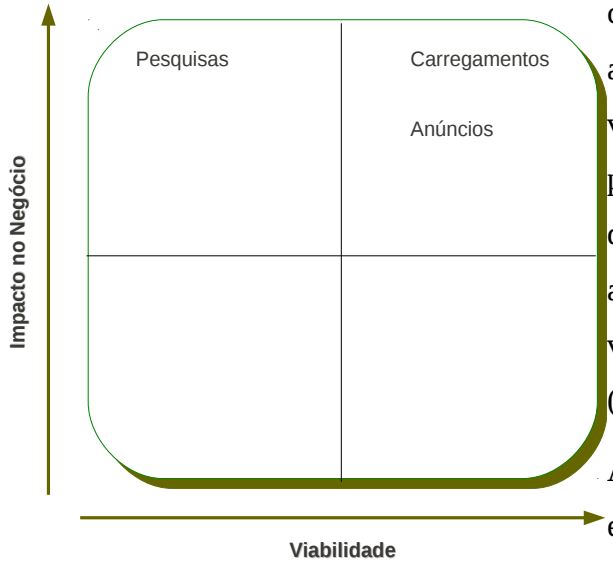


Figura 19: Matriz de Viabilidade de Requisitos

Analisando a figura 19, verificamos que, embora os processos relacionados com as pesquisas no portal tenham um elevado

impacto no negócio do serviço SAPO, a viabilidade da sua execução é substancialmente mais baixa, dado o elevado volume de dados envolvido. Assim, a prova de conceito incidiu sobre os carregamentos de contas SAPO e sobre os Anúncios SAPO.

O âmbito preliminar da prova de conceito realizado pode ser assim definido:

- Impressões e cliques de anúncios desde 2007, com granularidade diária;
- Carregamentos de contas SAPO desde 2007, com granularidade diária;
- Dimensões
    - Data (granularidade: dia)
    - Geografia (granularidade: localidade)
    - Anunciante (granularidade: anunciante)
    - Palavras (granularidade: palavra)
    - *Sites* (granularidade: *site*)
    - Tipo de Carregamento (granularidade: tipo)
- Utilizadores: Tratando-se de uma prova de conceito, não se prevê a sua entrada em produção, estando o acesso reservado a alguns utilizadores chave para efeitos de demonstração e avaliação. Caso a prova de conceito venha a evoluir para um projeto comercial concreto, ter-se-á que ter em conta a utilização da solução principalmente por utilizadores de negócio;
- Exclusões: Está excluída a utilização de fontes externas à base de dados operacional dos anúncios SAPO, nomeadamente, os *logs* JSON;
- Riscos: A não inclusão da informação contida nos *logs* JSON poderá levar à redução do âmbito da prova de conceito. O tempo disponível que a "equipa técnica" tem para dedicar à implementação da prova de conceito é limitado, pelo que existe o risco de não conclusão do protótipo no prazo previsto;
- Mitigação dos riscos: Eventual redução do âmbito da prova de conceito. Comunicação regular com os diferentes *stakeholders* acerca do cumprimento das tarefas inerentes à

realização da iniciativa.

3.2.4 Justificação de Negócio

Para além da motivação de negócio explicitada anteriormente, aquando da avaliação da prontidão organizacional para uma iniciativa de BI, num projeto propriamente dito seria ainda necessário fazer uma rigorosa análise custo/benefício que incluísse custos e benefícios financeiros e intangíveis. Esta análise teria que incluir investimentos e custos financeiros com a aquisição de *hardware*, serviços e formação, retornos e benefícios financeiros e o retorno do investimento estimado [KIM08].

Em relação a estes indicadores, como já foi referido, os custos serão virtualmente nulos, não sendo necessária a aquisição de *hardware* ou serviços. Em termos de benefícios, eles serão sobretudo intangíveis, na medida em que o objetivo desta iniciativa é, para além da vertente académica, demonstrar à organização as potencialidades e viabilidade de uma solução de BI *open source* na sua área de negócio.

Caso a prova de conceito venha a evoluir para um projeto em concreto, todas estas questões terão que ser detalhadamente equacionadas.

3.2.5 Plano de Projeto

Uma vez estabelecido o âmbito da prova de conceito, é chegado o momento de construir o plano de projeto. Num projeto concreto, começar-se-ia por constituir a equipa de projeto. Neste caso, a equipa de projeto é constituída apenas pelo mestrando, assumindo o co-orientador e alguns membros chave da organização o papel de patrocinadores. Kimball [KIM08] define um grande conjunto de papéis que deverão ser assumidos pelos diferentes membros da equipa de projeto, como gestor de projeto, patrocinador de negócio, analista de negócio, modelador de dados, administrador do SGBD do *data warehouse*, *designer* do sistema de ETL e *data staging*, responsáveis pelo desenvolvimento das aplicações para os utilizadores finais, e responsáveis por formar os utilizadores, entre outros. Na metodologia de Kimball estão definidas todas as responsabilidades inerentes a estes papéis, que podem ser assumidos por membros diferentes da equipa, ou, em caso de necessidade, um único membro poderá assumir mais que um papel.

Na presente situação, tratando-se de uma prova de conceito, uma única pessoa (o mestrando) assumiu todos estes papéis, à exceção dos de patrocinador.

Definida a equipa de projeto, foi construído o plano do projeto, que pode ser consultado no Anexo D - Plano de Projeto. O plano de projeto foi construído usando o Planner, uma aplicação de gestão de projetos *open source* para Linux, semelhante ao muito popular Microsoft Project.

3.3 Requisitos de Negócio

Nesta secção serão detalhados os requisitos de negócio que a prova de conceito deverá satisfazer.

Kimball [KIM08] propõe, na sua metodologia, dois métodos alternativos para levar a cabo o processo de levantamento de requisitos. São eles as entrevistas individuais ou em pequenos grupos com utilizadores e técnicos, ou sessões com um número mais alargado de participantes, dirigidas por um facilitador. Ambas têm os seus prós e contras, sendo que neste caso, dado o âmbito relativamente limitado da prova de conceito, se optou por realizar entrevistas individuais.

Kimball propõe um procedimento extremamente detalhado de preparação e condução de entrevistas, que se poderá justificar em projetos mais complexos, incluindo a preparação de um guião pormenorizado. Neste caso, a abordagem adotada não foi tão meticulosa, dado tratar-se de uma prova de conceito com um âmbito relativamente reduzido em termos de negócio. Ainda assim, para que os resultados fossem os melhores possíveis, foi feito algum trabalho preparatório.

3.3.1 Preparação

A preparação das entrevistas de levantamento de requisitos passou, numa primeira fase, por conversas informais com a equipa técnica do sistema operacional e por alguma investigação na Internet no sentido de compreender o negócio o melhor possível, bem como os dados atualmente disponibilizados pela plataforma.

Assim, para além da descrição genérica do negócio efetuada no início do capítulo, chegou-se às seguintes conclusões:

- O sistema dos Anúncios SAPO permite a importação de campanhas registadas no serviço Adwords do Google;
- Os Anúncios SAPO funcionam com um sistema de pré-pagamento;
- Existem redes afiliadas aos Anúncios SAPO (por exemplo, Anúncios IOL, Anúncios Cofina, Controlinveste, entre outras) que utilizam a plataforma do SAPO para colocar os seus anúncios. Estas redes possuem utilizadores com privilégios especiais (os *master users*), que gerem grupos de utilizadores e podem, entre outras coisas, realizar carregamentos em massa. Estes utilizadores requerem informação analítica específica, muitas vezes obtida a partir de *queries* complexas sobre a base de dados operacional;
- A faturação dos carregamentos é feita manualmente em folhas de cálculo;
- Em termos técnicos, a plataforma Anúncios SAPO foi desenvolvida utilizando MySQL, PHP, Linux, Perl, Python e C;
- A informação disponibilizada atualmente pelo *back-office* da plataforma é, resumidamente, a constante no Anexo E – Informação do Backoffice Anúncios SAPO.

3.3.2 Entrevistas

Para a condução das entrevistas, tomou-se o guião proposto por Kimball em [KIM08] como referência. Assim, discutiram-se os objetivos da iniciativa, determinaram-se as responsabilidades dos entrevistados no processo de tomada de decisão, identificaram-se indicadores necessários à tomada informada de decisões, os entrevistados expuseram brevemente qual a informação que utilizam e necessitam, como obtêm essa informação atualmente e quais as insuficiências na informação atual, em termos de tempo, qualidade e quantidade.

As entrevistas foram calendarizadas pelo co-orientador, tomando em conta as disponibilidades de todos os intervenientes.

3.3.3 Requisitos

Os entrevistados identificaram algumas insuficiências na informação fornecida atualmente. Nomeadamente,

- Não é possível, com a informação disponível, identificar os *sites* em que os cliques são efetuados, a não ser que o anúncio respetivo faça parte de uma campanha segmentada por *site*;
- Falta de informação detalhada nas campanhas segmentadas por palavras;
- A informação que existe sobre os *slots* em que os anúncios são impressos (normalmente, as páginas dispõem de três *slots* para colocação de anúncios. Dois *banners* no topo e no fundo da página, que garantem os maiores CTRs, e um *skyskraper* lateral, com um CTR normalmente menor) é insuficiente. Atualmente, é enviada por e-mail de forma pouco clara e trabalhada manualmente numa folha de cálculo;
- Falta de informação histórica anterior a 2006;
- Carregamentos feitos por *wallet* SAPO nem sempre condizem com os dados obtidos a partir da *gateway* de pagamentos.

Atualmente, a informação utilizada envolve a evolução horária de cliques, e outras métricas por redes e clientes, e diária de carregamentos por tipo de carregamento. As análises mais comuns relativas a carregamentos estão relacionadas com a recorrência dos mesmos, com a caracterização e segmentação dos anunciantes que efetuam carregamentos, com o tempo que o saldo das contas dos anunciantes demoram a ser consumidos e com a identificação dos clientes com contas ativas mas sem cliques.

Os entrevistados manifestaram preocupação com o carácter estático dos relatórios que recebem, o que obriga a trabalho adicional dos dados em folhas de cálculo. Manifestaram ainda a necessidade que sentem de criação de um sistema de alarmes que possa sinalizar eventos que requeiram a sua atenção imediata relacionados com situações anormais nas análises mencionadas.

Em conversas posteriores com a equipa técnica dos Anúncios SAPO, chegou-se à conclusão que algumas das insuficiências apontadas se devem a limitações na base de dados operacional, nomeadamente, a impossibilidade de saber em que *sites* são efetuados os cliques e em que *slots* os anúncios são impressos. Atualmente, tal já é possível, após desenvolvimentos adicionais efetuados pela referida equipa, mas não existem dados históricos

suficientes para alimentar um *data warehouse*. Assim, a solução de BI não deverá, a curto prazo, poder preencher esta lacuna.

É importante referir, no entanto, que grande parte da informação não disponível nos relatórios atuais está registada nos *logs* do sistema, num formato *standard* (JSON), pelo que poderá ser considerada em iterações subsequentes à realização da prova de conceito.

Podemos sumarizar estes requisitos num conjunto de questões analíticas e suas variações, a que o sistema deverá ser capaz de responder, de forma interativa e intuitiva:

- Quais as redes que mais anúncios serviram por período de tempo?
- Qual a evolução das várias redes ao longo do tempo?
- Qual o top 5 de palavras em termos de impressões, cliques, CTR e CPC médio por unidade de tempo?
- Qual o top 5 de anunciantes nas diferentes redes por período de tempo?
- Quais os distritos cujos anunciantes deram origem a mais impressões em 2009?
- Quais os meios de pagamento mais usados por período de tempo?
- Como evoluiu a utilização dos diversos tipos de carregamento ao longo do tempo na rede SAPO?
- Qual a rede para além do SAPO com mais carregamentos por período de tempo?
- Qual a evolução ao longo do tempo dos carregamentos nas diversas redes?
- Qual o top 5 de carregamentos por anunciantes e por rede?

Após o levantamento de requisitos e, concluíndo-se não ser possível a utilização da dimensão *site* na prova de conceito devido a limitações no sistema operacional, tomou-se a decisão de reduzir o âmbito da prova de conceito definido em 3.2.3 - Definição Preliminar do Âmbito da Prova de Conceito. Assim, retirou-se a dimensão *site* do conjunto de dimensões a considerar nesta fase, sem prejuízo da possibilidade de vir a ser incluída no futuro.

3.4 Modelação Dimensional

Kimball [KIM08] define modelação dimensional como sendo uma técnica de desenho lógico que procura apresentar os dados através de um modelo intuitivo, permitindo um rápido acesso aos dados, utilizando o modelo relacional com algumas restrições importantes. Cada modelo dimensional é composto por uma tabela com uma chave primária composta, denominada de tabela de factos, e um conjunto de tabelas mais pequenas, designadas de tabelas de dimensões. Cada dimensão tem uma chave primária simples, que corresponde a um dos componentes da chave composta da tabela de factos. Esta estrutura é conhecida como esquema em estrela, utilizada, com algumas variantes, pela maioria das soluções de BI. Cada esquema em estrela constitui o que se chama de *data mart*, e o conjunto articulado de todos os *data marts* constitui o modelo lógico do *data warehouse*.

Na medida em que as dimensões podem sofrer alterações, Kimball [KIM08] define diferentes tipos de dimensões, consoante o tipo de alteração que é necessário fazer:

- Tipo I – A alteração é registada na tabela através de uma instrução UPDATE. Utilizado nos casos em que não é necessário guardar o histórico da alteração;
- Tipo II – É criado um novo registo na tabela para guardar o histórico da alteração. Útil para quando se pretende guardar o histórico das alterações efetuadas;
- Tipo III – As alterações na dimensão são guardadas no mesmo registo da tabela, coexistindo ambas as versões. Deve ser evitado, na medida em que implica alterações na estrutura da tabela a nível físico.

	Impressões	*Carregamentos*
Data	X	X
Geografia	X	X
Palavras	X	
Anúncios	X	
Anunciantes	X	X
Tipo de Carregamento		X

Tabela 5: Matriz do Data Warehouse

Como ponto de partida, Kimball [KIM08] propõe a construção de uma matriz do *data warehouse*, em que factos e dimensões são cruzados, dando origem à matriz da tabela 5. A partir desta tabela, verifica-se a necessidade de construir duas tabelas de factos (Impressões e Carregamentos), com três das seis dimensões de análise comuns (Data, Geografia e Anunciantes), duas dimensões específicas das Impressões (Anúncios e Palavras) e uma outra específica dos Carregamentos (Tipo de Carregamento).

Para a fase de desenho lógico do modelo de dados do *data warehouse* foi utilizado o SQL Power Architect, uma aplicação de modelação de dados *open source* disponibilizada pelo SQL Power Group.

3.4.1 Impressões (Facto)

Cada registo da tabela de impressões conterá o número de impressões e cliques recebidos por anúncio, anunciante, dia, localidade (referente ao anunciante) e palavra. Esta será a sua granularidade, tratando-se, portanto, do que Kimball define como um instantâneo periódico (*snapshot*) [KIM08], uma vez que o registo de cliques e impressões individuais implicaria um volume de dados incomportável. Adicionalmente, serão registadas as métricas de CPC Médio, RPM, Custo Total, CTR, Conversões e Posição Média para cada registo.

Estará associada às dimensões Data, Geografia, Palavras, Anúncios e Anunciantes.

FactImpressoes
IdData: Integer (PK)
IdAnuncio: Integer (PK)
IdAnunciante: Integer (PK)
IdGeografia: Integer (PK)
IdPalavra: Integer (PK)
Cliques: Integer
Impressoes: Integer
CPCMedio: Float
RPM: Float
CustoTotal: Float
CTR: Float
Conversoes: Integer
PosicaoMedia: Float

Tabela 6: Desenho lógico da tabela de factos de Impressões

O seu desenho lógico está esquematizado na tabela 6.

3.4.2 Carregamentos (Facto)

A granularidade da tabela de carregamentos corresponde ao valor total dos carregamentos efetuados diariamente por anunciante, respetiva localidade e por tipo de carregamento, constituindo assim um instantâneo periódico.

Estará, assim, associada às dimensões Data, Anunciante, Geografia e Tipo de Pagamento.

FactCarregamentos
IdTempo: Integer (PK)
IdAnunciante: Integer (PK)
IdGeografiaAnunciante: Integer (PK)
IdTipoPagamento: Integer (PK)
Valor: Float

Tabela 7: Desenho lógico da tabela de factos de Carregamentos

O seu desenho lógico consta da tabela 7.

3.4.3 Data (Dimensão)

Utilizando os conhecimentos transmitidos pela Professora Doutora Maria José Trigueiros nas suas aulas de Sistemas Integrados de Apoio à Decisão, foi construído o diagrama da figura 20, onde são visíveis os atributos individuais da tabela de dimensão Data, bem como as hierarquias por eles formadas. Trata-se de uma forma mais clara de mostrar detalhadamente a articulação dos atributos individuais de uma dimensão, bem como as hierarquias por eles formadas, semelhante à proposta por Kimball [KIM08].

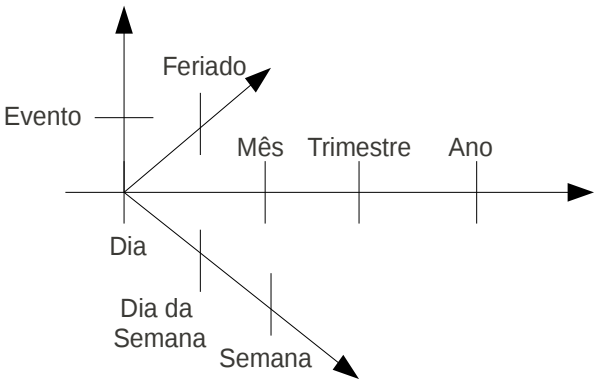


Figura 20: Dimensão Data - Hierarquias

Esta dimensão apresenta várias hierarquias, sendo as principais a formadas pelos atributos dia – mês – trimestre – ano, e também a formada por dia – dia da semana – semana – ano.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdData : Integer (PK)	Chave substituta	1800	N/A	N/A
Data : DateTime	Chave natural	1800	I	01-01-10
Dia : Integer	Dia	31	I	De 1 a 31
DiaSemana: Varchar	Dia da semana	7	I	Segunda, Terça...
Feriado: Varchar	Feriado	15	I	Páscoa, Natal...
Evento: Varchar	Evento especial	N/D	I	Final da Taça
Semana: Integer	Nº da semana	52	I	De 1 a 52
Mes: Varchar	Mês	12	I	Janeiro, Fevereiro...
Trimestre: Integer	Nº do trimestre	4	I	De 1 a 4
Ano: Integer	Ano	5	I	2007, 2008...

Tabela 8: Detalhes lógicos da tabela Data

A tabela 8 sumariza o desenho lógico da tabela, bem como alguma informação adicional acerca de cada atributo.

3.4.4 Geografia do Anunciante (Dimensão)

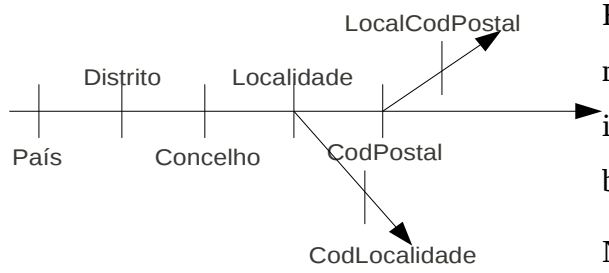


Figura 21: Dimensão Geografia - Hierarquias

Recorrendo à mesma abordagem, a figura 21 mostra os atributos individuais mais importantes da tabela da dimensão Geografia, bem como as hierarquias por ela constituídas.

Na tabela 9 podemos encontrar o desenho lógico desta tabela, e alguma informação adicional acerca dos diversos atributos.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdGeografia : Integer (PK)	Chave substituta	187000	N/A	N/A
codLocalidade : Varchar	Chave natural	187000	I	0101249
Pais : Varchar	País	1	I	Portugal
Distrito: Varchar	Distrito	18	I	Lisboa, Setúbal...
Concelho: Varchar	Concelho	300	I	Oeiras, Seixal...
Localidade: Varchar	Localidade	20000	I	Sassoeiros, Corroios...
CodPostal: Varchar	Código Postal	187000	I	2855-424
LocalCodPostal: Varchar	Descrição do código postal	3700	I	LISBOA, CORROIOS

Tabela 9: Detalhes lógicos da tabela Geografia

3.4.5 Anunciante (Dimensão)

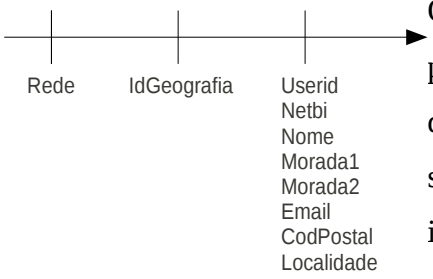


Figura 22: Dimensão Anunciante - Hierarquias

O serviço Anúncios SAPO tem milhares de utilizadores, que podem ser hierarquizados por rede e área geográfica (como definido na dimensão Geografia). Os restantes atributos serão utilizados para caracterizar cada anunciante individualmente, tal como se depreende da figura 22.

Na tabela 10 está descrito o desenho lógico da tabela correspondente à dimensão Anunciante.

Note-se que esta dimensão, no que respeita à classificação quanto ao tipo de SCD, é híbrida, na medida em que será registado o histórico das alterações aos campos relativos à morada e rede de afiliação, enquanto que este registo não é considerado relevante para os restantes atributos. Adicionalmente, tratando-se esta iniciativa uma prova de conceito, apenas nesta

dimensão serão tratados atributos SCD tipo II.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdAnunciante : Integer (PK)	Chave substituta	15500	N/A	
user_id : Integer	Código operacional	15500	I	
netbi : Varchar	Chave natural	15500	I	*Usernames*
Nome : Varchar	Nome	15500	I	Filipe, António...
Morada1 : Varchar	Primeira linha da morada	15500	II	Rua da Betesga, 1
Morada2 : Varchar	Segunda linha da morada	2000	II	Rossio
CodPostal : Varchar	Código Postal	8000	II	2855-424
Localidade : Varchar	Localidade	2000	II	Lisboa
email: Varchar	e-mail	12000	I	Mario.silva@sapo.pt
IdGeografia: Integer (FK)	Ligação à dimensão Geografia	5500	II	12400
Rede : Varchar	Rede à qual o utilizador está afiliado	10	II	SAPO, Cofina
Tipo : Varchar	Particular ou corporativo	2	I	Particular, Empresarial

Tabela 10: Detalhes lógicos da tabela Anunciante

3.4.6 Anúncio (Dimensão)

Existem centenas de milhar de anúncios atualmente registados na base de dados operacional dos Anúncios SAPO. Como explicitado na figura 23, estes podem ser hierarquizados descendentemente por anunciante, campanha, grupo e anúncio individual. Os diferentes atributos colocados em cada nível têm como objetivo a caracterização de, respetivamente, cada campanha, grupo de anúncios e anúncio individual.

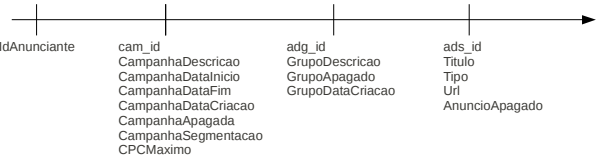


Figura 23: Dimensão Anúncio - Hierarquias

A tabela 11 resume o desenho lógico da dimensão Anúncio, fornecendo informação adicional sobre os seus atributos.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdAnuncio : Integer (PK)	Chave substituta	400000	N/A	
ads_id : Integer	Chave natural	400000	I	
IdAnunciante : Integer (FK)	Ligação à dimensão anunciante	15500	I	
cam_id : Integer	Código de campanha	26000		
CampanhaDescricao : Varchar	Nome da campanha	15500	I	Campanha 1

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
CampanhaDataInicio : Date	Data de início da campanha	12300	I	2007-10-10
CampanhaDataFim : Date	Data de fim da campanha	12300	I	2008-10-10
CampanhaDataCriacao : Date	Data de criação da campanha	25000	I	2007-10-09
CampanhaApagada : Integer	Indica se a campanha foi apagada pelo anunciante	2	I	0, 1
CampanhaSegmentacao : Varchar	Tipo de segmentação da campanha	2	I	Sites, Palavras
CPCMaximo : Float	CPC Máximo para a campanha	200	I	0,10; 0,50
adg_id : Integer	Código de grupo de anúncios	125000	I	
GrupoDescricao : Varchar	Nome do grupo de anúncios	83000	I	Grupo 1
GrupoApagado : Integer	Indica se o grupo foi apagado pelo anunciante	2	I	0, 1
GrupoDataCriacao : Date	Data de criação do grupo de anúncios	50000	I	2008-11-10
Titulo : Varchar	Título do anúncio	35000	I	Vende-se carro
Tipo : Varchar	Tipo de anúncio	2	I	Texto, Imagem
Url : Varchar	URL do site promovido	20000	I	www.sapo.pt
AnuncioApagado : Integer	Indica se o anúncio foi apagado pelo anunciante	2	I	0, 1
AnuncioDataCriacao : Date	Data de criação do anúncio	68500	I	2008-11-10

Tabela 11: Detalhes lógicos da tabela Anúncio

3.4.7 Palavras (Dimensão)

A única hierarquia da dimensão Palavras tem apenas um nível (as próprias palavras), pelo que não será necessário mostrá-la graficamente. Poder-se-iam hierarquizar as palavras por tamanho, mas tal não teria relevância em termos analíticos. A tabela 12 resume o desenho lógico da dimensão Palavras.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdPalavra : Integer (PK)	Chave substituta	1200000	N/A	
word : Varchar	Chave natural. Palavra no formato Base64.	1200000	I	03BlcmE=
Tamanho : Integer	Número de palavras na expressão	44	I	1, 2, 3...
Palavra : Text	Texto da expressão (palavra)	1200000	I	Opera, carro

Tabela 12: Detalhes lógicos da tabela Palavras

3.4.8 Tipo de Pagamento (Dimensão)

À semelhança da dimensão Palavras, esta dimensão possui uma única hierarquia com um

único nível. Comparativamente às outras dimensões, esta terá uma cardinalidade muito baixa, como fica evidente na tabela 13, em que se apresenta o desenho lógico da dimensão.

Atributo	Descrição	Cardinalidade	SCD	Valores Típicos
IdTipo : Integer (PK)	Chave substituta	4	N/A	
payment_type : Integer	Chave natural	4	I	1, 2, 3...
Tipo : Varchar	Descrição do tipo de pagamento	4	I	Wallet, PayPal...

Tabela 13: Detalhes lógicos da tabela TipoPagamento

3.4.9 Modelo Lógico

Como já foi referido, o processo de modelação dimensional exposto nas secções anteriores tomou como referência a metodologia *Business Dimensional Lifecicle*, de Kimball [KIM08], embora com uma abordagem um pouco mais superficial que a proposta pelo autor, devido ao facto de se tratar de uma prova de conceito, com um âmbito mais reduzido relativamente ao que é usual em projetos concretos.

Como entregáveis desta fase, foi produzida a matriz completa do *data warehouse*, que consta do Anexo F – Matriz do Data Warehouse. A matriz foi construída usando a aplicação de folha de cálculo OpenOffice Calc, uma popular folha de cálculo *open source*, parte do pacote de produtividade OpenOffice.org, cujas funcionalidades rivalizam com a maioria dos pacotes comerciais.

A figura 24 concretiza esta matriz no modelo lógico propriamente dito do *data warehouse*, construído com a ferramenta *open source* de modelação de dados SQL Power Architect, uma aplicação Java do SQL Power Group.

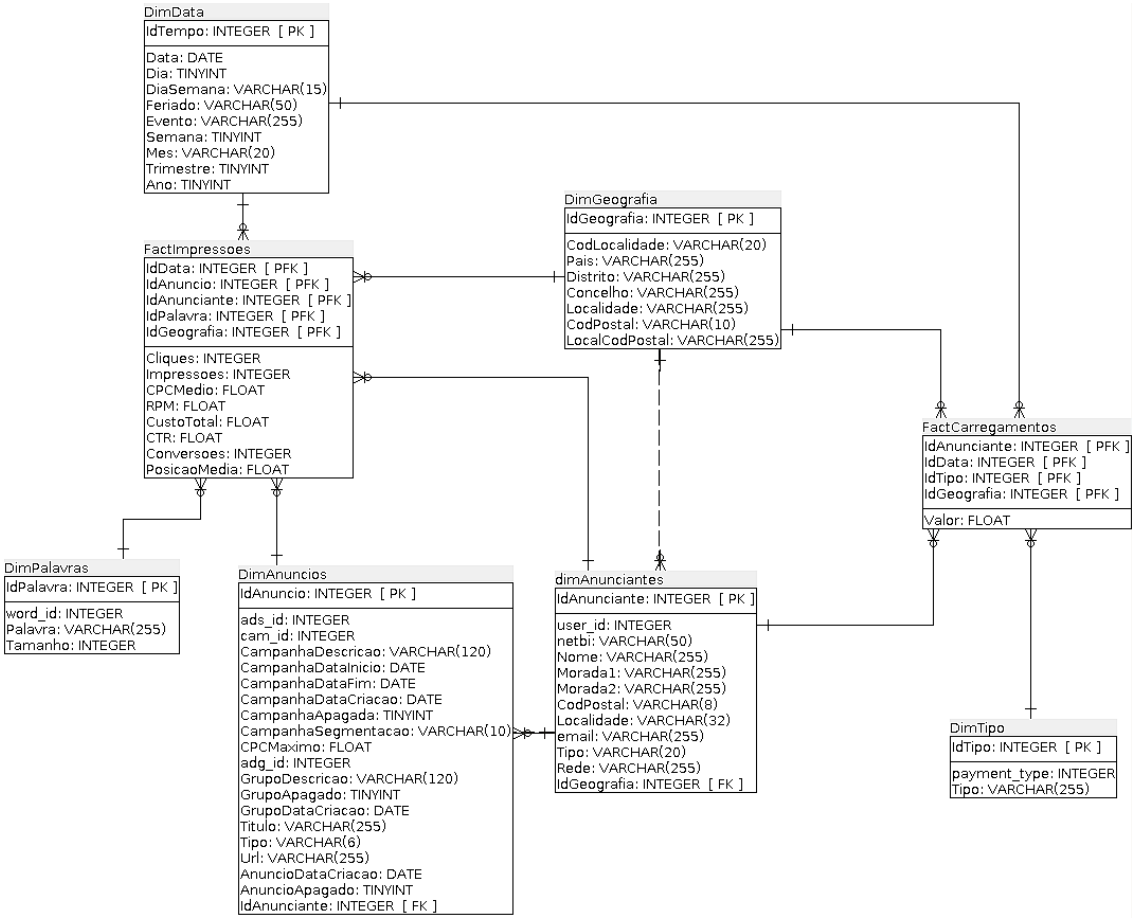


Figura 24: Modelo lógico do Data Warehouse

3.5 Arquitetura Técnica

Neste sub-capítulo far-se-á uma descrição de muito alto nível da arquitetura técnica usada na implementação da prova de conceito. Ao contrário do que acontece tipicamente nos grandes projetos de BI, grande parte das decisões inerentes a esta arquitetura estão fora do controlo da "equipa" técnica, pelo que este ponto não será abordado detalhadamente, à exceção dos componentes que dependem, efetivamente, da referida "equipa".

Nomeadamente, não foram criados planos detalhados de arquitetura e de infraestrutura, que são entregáveis importantes na metodologia de Kimball [KIM08], uma vez que todos os componentes da infraestrutura e arquitetura serão executados na mesma máquina.

3.5.1 Hardware

Para a implementação da presente prova de conceito, a PT Comunicações disponibilizou um computador de secretária com as seguintes características base:

- Fujitsu-Siemens Esprimo
- Processador Intel Pentium D 3 GHz
- RAM: 3 GB
- Disco: 160 GB

Todos os componentes da arquitetura adotada funcionam em simultâneo neste computador.

3.5.2 Sistemas Fonte

Na presente prova de conceito será usado apenas um sistema fonte, uma cópia da base de dados operacional da plataforma Anúncios SAPO, que será executada no computador acima descrito, em simultâneo com os restantes componentes. Trata-se de uma base de dados MySQL com 87 tabelas e um volume de dados e índices total de cerca de 50 Gbytes.

Existem dados que podem ser relevantes para próximas iterações, nomeadamente, dados relativos aos pedidos de anúncios e pesquisas ao servidor, que são armazenados em *logs*, o que, embora estejam num formato estruturado e acessível (JSON), representaria um risco acrescido para o sucesso desta iniciativa, bem como um volume de dados incomportável com os recursos disponíveis.

Foram ainda utilizados alguns dados externos, nomeadamente, uma folha de cálculo para preenchimento da dimensão Data e um ficheiro de texto em formato CSV proveniente do *site* dos Correios[6] com a lista completa de localidades e códigos postais existentes em Portugal, para preenchimento da dimensão Geografia.

3.5.3 Área de *Staging*

A área de *staging* pode ser definida como o local onde ocorrem a maior parte dos processos de transformação de dados que foram extraídos dos sistemas fonte antes de serem carregados no

6 http://www2.ctt.pt/fewcm/wcmservlet/system/galleries/download/servicosonline/conteudosextra/0502_todos_cp.zip

data warehouse [KIM08]. A título de exemplo, inclui processos tão importantes como a criação de chaves substitutas para as dimensões.

Na presente prova de conceito, a maioria dos processos que ocorrem na área de *staging* ocorrem diretamente sobre a base de dados do *data warehouse*, com recurso a alguns ficheiros temporários. É importante notar, no entanto, que esta não é uma boa prática em sistemas mais complexos, devendo existir uma base de dados própria para esse efeito que centralize, por exemplo, o processo de criação de chaves substitutas. No presente caso, ao contrário do que acontece num projeto típico, existe apenas uma fonte de dados, não considerando os pequenos ficheiros externos, sendo que a própria ferramenta de ETL adotada possui funcionalidades que facilitam a geração e gestão de chaves substitutas diretamente sobre as tabelas de dimensão, pelo que se adotou esta solução para acelerar o desenvolvimento. Num sistema destinado a utilização real, a opção teria sido substancialmente diferente.

3.5.4 Área de Apresentação de Dados

A área de apresentação de dados, ou seja, o *data warehouse* propriamente dito, é constituída por uma base de dados MySQL que segue, tanto quanto possível, o modelo lógico apresentado em 3.4.9 - Modelo Lógico, com a adição de tabelas com agregações. Esta base de dados está implementada no servidor MySQL instalado no computador acima descrito.

O desenho físico será, no entanto, diferente, uma vez que os *storage engines* utilizados (MyISAM e Memory) não suportam relações com chaves estrangeiras entre as tabelas. Quando forem abordados os critérios de seleção de ferramentas utilizados, serão expostas as razões que presidiram a esta opção.

Além disso, a ferramenta de ETL adotada (Pentaho Kettle) requer a existência de campos adicionais nas tabelas de dimensão para a manutenção de SCDs tipo II.

3.5.5 Seleção de Ferramentas

Nesta secção é descrita a abordagem genérica que foi adotada para a seleção das ferramentas a utilizar. De acordo com Kimball [KIM08], existem quatro domínios em que é necessária uma avaliação técnica:

- Plataformas de hardware;

Open Source Business Intelligence

- Sistemas de gestão de bases de dados (SGBDs);
- Ferramentas de acesso aos dados (Plataforma de BI) – Foram consideradas as duas plataformas de BI *open source* mais populares a nível mundial, a JasperSoft BI e a Pentaho BI Suite.
- Ferramenta de ETL – Foram consideradas as ferramentas incluídas nas plataformas de BI consideradas e o Talend Open Studio;

Kimball propõe uma análise bastante detalhada à fase de seleção de ferramentas, incluindo a construção de matrizes de requisitos das soluções a avaliar, serviços de suporte dos fornecedores, análise de custos e, eventualmente, a construção de protótipos.

Tratando-se a presente prova de conceito, ela própria, um protótipo, foi adaptada uma abordagem de mais alto nível, obedecendo a critérios mais genéricos. É de realçar, no entanto, que em projetos concretos esta fase deverá ser alvo de maior atenção, recomendando a Gartner [BIT08] que se usem os mesmos critérios de avaliação (funcionalidade, fiabilidade, suporte e estabilidade) para as soluções *open source* que os que se usam para qualquer outro tipo de solução, e que as análises efetuadas tenham o mesmo nível de detalhe.

3.5.5.1 Plataforma de *Hardware*

Neste domínio, não foi necessária uma avaliação de diferentes plataformas, na medida em que foi necessário trabalhar com os recursos disponibilizados, um computador de secretária com sistema operativo Linux. Esta plataforma é suficiente para construir um pequeno protótipo, mas seria obviamente inadequada para um sistema em produção típico.

3.5.5.2 Sistemas de Gestão de Bases de Dados (SGBDs)

Foram considerados diferentes *storage engines* MySQL, e um *storage engine* externo, o Brighthouse da Infobright.

Não foram considerados outros SGBDs para além do MySQL, dado que a própria base de dados operacional da plataforma Anúncios SAPO está construída sobre um SGBD MySQL (o que vem, mais uma vez, evidenciar a capacidade de suporte a aplicações críticas de negócio por parte de ferramentas *open source*). Além de que existe na organização um *know-how* considerável relativamente a este SGBD, o MySQL tem capacidade para suportar soluções de

BI através dos seus diversos *storage engines* [MYS07].

Assim, o MySQL tem ao dispor dos utilizadores diversas formas de armazenamento de dados, designadas por *storage engines*:

- MyISAM – Segundo a MySQL, este *storage engine* executa *queries* e *inserts* rapidamente, é não transacional, dispõe de *locks* a nível de tabela e suporta vários tipos de índices, o que o torna numa boa escolha para soluções de *data warehousing*;
- Archive – Permite uma economia de espaço de até 80% em relação a outros *storage engines*, mas não permite *updates* nem *deletes* seletivos. Pode ser uma boa opção para arquivar dados antigos pouco utilizados, ou para guardar informação sensível, como tabelas de auditoria;
- Memory – Este *storage engine* tem a particularidade de guardar todos os dados em memória, o que lhe garante um tempo de resposta muito rápido. É ideal para tabelas de dimensão, mas há alguns tipos de dados que não são suportados;
- CSV – Permite acesso direto do servidor a ficheiros de texto em formato CSV;
- Federated – Permite criar uma única base de dados a nível lógico a partir de várias bases de dados físicas em vários servidores.
- InnoDB – *Storage engine* transacional, mais adequado ao suporte a sistemas operacionais;
- NitroEDB – Permite lidar com grandes volumes de dados e tem um ótimo desempenho, mas não é *open source*, pelo que não foi considerado;
- Brighthouse – Resultado da pareceria entre a MySQL e a Infobright, é orientado por colunas que permite uma grande redução do espaço em disco utilizado, e é aclamado pela MySQL como a escolha perfeita para um *data warehouse*.

A solução da Infobright, descrita em 2.7.1.9 - Infobright, seria, pelas características descritas, como a primeira escolha para o desenvolvimento da presente prova de conceito. Tem, no entanto, uma limitação importante. Na sua versão *open source*, os comandos SQL *insert, update* e *delete* não são suportados, sendo possível apenas o carregamento por *bulk loading*, o que obrigaria a manter uma base de dados idêntica usando outro *storage engine* e carregar a

base de dados Infobright por *bulk loading*, tabela a tabela, a partir da primeira.

Assim, foi selecionado o *storage engine* MyISAM para todas as tabelas. O *storage* engine Memory foi considerado para as tabelas de dimensão, mas não foi adotado, uma vez que implicaria o recarregamento de todas as tabelas após cada reinício do servidor [ORA10]. Embora tenham um melhor desempenho, esta desvantagem determinou a sua não adoção.

3.5.5.3 Plataforma de BI

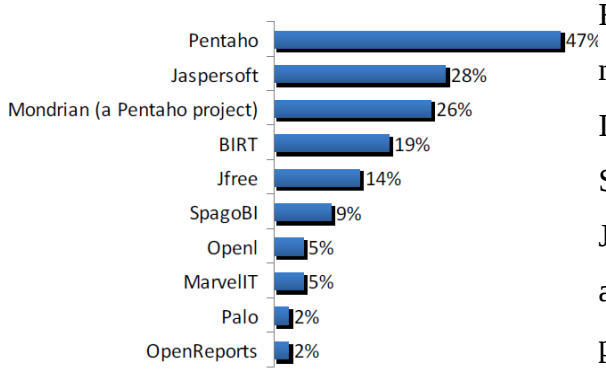


Figura 25: Plataformas BI open source *mais utilizadas [MAD09]*

Foram consideradas as duas plataformas de BI mais populares a nível mundial, a Business Intelligence Suite da Jaspersoft e a Pentaho BI Suite [MAD09]. Segundo o Gartner Group, a Jaspersoft e a Pentaho emergiram nos últimos anos como soluções viáveis no mercado de plataformas de BI. Ambas as soluções disponibilizam um conjunto abrangente de funcionalidades comparáveis às dos vendedores tradicionais de plataformas de BI [RIC09].

A escolha recaiu sobre a Pentaho porque, como se pode ver desde logo pela figura 25, se trata da plataforma de BI *open source* mais utilizada a nível global.

Adicionalmente, ambas as plataformas incluem ferramentas de ETL, servidor OLAP, relatórios e ferramentas analíticas, mas falta uma ferramenta de *data mining* à solução da Jaspersoft, o que faz da Pentaho BI suite uma plataforma mais completa. Esta lacuna não é determinante na implementação da prova de conceito, uma vez que não se efetuaram análises de *data mining* e, em caso de necessidade, poder-se-ia utilizar qualquer outra ferramenta, mas foi tida em consideração, dado o potencial que este tipo de ferramentas apresenta e a necessidade deste tipo de análises nos Anúncios SAPO.

Outra razão da escolha prende-se com o servidor OLAP utilizado por ambas as soluções, o Mondrian, desenvolvido pela Pentaho. Assim, fosse qual fosse a solução selecionada de entre as duas em consideração, um dos principais componentes da plataforma seria Pentaho, pelo que se tomou a decisão de usar a plataforma da Pentaho. De notar também que a ferramenta

de ETL incluída na plataforma da Jaspersoft também não foi desenvolvida pela própria, tratando-se de uma versão personalizada do Talend Open Studio.

Trata-se, obviamente, de um critério subjetivo, que não pretende retirar valor à solução da Jaspersoft, até porque, como se pode ver na tabela 14, as duas plataformas são muito semelhantes.

Módulos	Pentaho	Jaspersoft
Servidor Aplicacional	JBoss	JBoss
Autenticação	Acegi	Acegi
Dashboard	JFreeChart	JFreeChart
Data Mining	Weka	ND
SGBD	MySQL, Oracle, SQL Server, PostgreSQL, etc.	MySQL, Oracle, SQL Server, PostgreSQL, etc.
ETL	Pentaho Data Integration (Kettle)	JasperETL (Talend Open Studio)
Geo-Referenciação	Google Maps	Google Maps
Scheduler	Quartz	Quartz
OLAP	Mondrian, JPivot	Mondrian, JPivot
Portal	JBoss Portal	Liferay
Relatórios	Pentaho Report Designer, JasperReport, BIRT	JasperReport
Single Sign On	Acegi	CAS
Servidor Web	Tomcat	Tomcat

Tabela 14: Módulos das plataformas consideradas [GOL09]

3.5.5.4 Ferramenta de ETL

Para além das ferramentas de ETL incluídas nas plataformas de BI consideradas (Pentaho Data Integration e JasperETL), seria ainda de ter em conta, neste capítulo, o Talend Open Studio, uma das ferramentas de ETL *open source* mais populares.

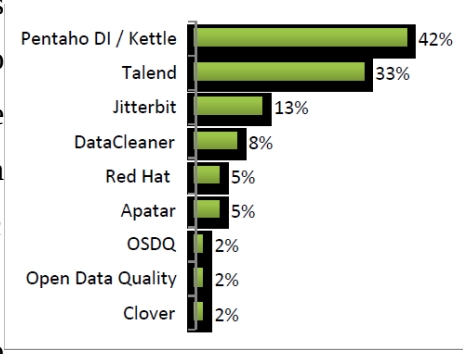


Figura 26: Ferramentas ETL open source mais utilizadas [MAD09]

Na realidade, o JasperETL é uma versão personalizada do Talend Open Studio, pelo que a escolha recaiu na ferramenta Pentaho Data Integration, para além do facto de ser a mais

utilizada mundialmente, como é visível na figura 26.

3.6 Desenho Físico

Uma vez decididas as ferramentas a utilizar, é possível agora efetuar o desenho físico da base de dados.

Na metodologia de Kimball [KIM08], esta fase decorre em paralelo com a fase de seleção de ferramentas, como se pode verificar na figura 1 da página 16. É, contudo, necessário algum cuidado, uma vez que as ferramentas e SGBDs selecionados podem ter um forte impacto no desenho físico do *data warehouse*, como se verá. A cautela aconselha assim que a fase de desenho físico tenha início quando já existir alguma informação acerca das ferramentas selecionadas que possa ter impacto no modelo físico do *data warehouse*.

No presente caso, a decisão pela utilização de tabelas MyISAM implicou a ausência de relações físicas entre tabelas, uma vez que este *storage engine* não as suporta [RIC09]. Esta situação seria provavelmente inaceitável num sistema que não tivesse outros meios de garantir a integridade referencial da base de dados e não seria, mesmo assim, uma boa prática. No entanto, um *data warehouse* é um ambiente mais estável do ponto de vista da volatilidade da informação (é, por definição, uma base de dados não volátil), no sentido em que só excecionalmente são apagados ou alterados dados. Adicionalmente, uma das funções dos processos de ETL a desenhar é, precisamente, garantir a integridade referencial do *data warehouse* [KIM08], pelo que será o subsistema de ETL o responsável por lidar com esta questão, e não a própria base de dados.

A alternativa seria utilizar o *storage engine* InnoDB, que, com algum esforço de otimização, pode ser utilizado nalguns contextos, mas que, em termos genéricos, não é adequado para *data warehouses*, dado o seu carácter transacional.

Também a escolha da ferramenta de ETL (Pentaho Data Integration) teve impacto no desenho físico. Para que a gestão das *slowly changing dimensions* pudesse ser automatizada, foi necessário criar campos adicionais nas tabelas de dimensão:

- versao – para indicar a versão de cada registo da tabela de dimensão;
- inicioValidade – para indicar a data de início de validade de cada registo da

tabela;

- fimValidade - para indicar a data de final de validade cada registo da tabela.

Estes campos devem existir sempre em SCDs tipo II, qualquer que seja a ferramenta usada, mas a ferramenta de ETL selecionada exige também a presença destes campos em SCDs tipo I, o que, normalmente, não é necessário. A alternativa seria gerir as SCDs tipo I não utilizando a funcionalidade de gestão de SCDs da ferramenta, o que não parece justificável.

Desta forma, o modelo lógico introduzido em 3.4.9 - Modelo Lógico traduz-se no seguinte modelo físico da figura 27, sem tabelas de agregação, que serão criadas mais tarde. Neste modelo foram mantidas as designações PFK (*Primary Foreign Key*) e FK (*Foreign Key*) apenas por uma questão de legibilidade do modelo, uma vez que, na base de dados, os campos indicados como PFK são apenas PK (*Primary Key*) e os campos FK são campos normais.

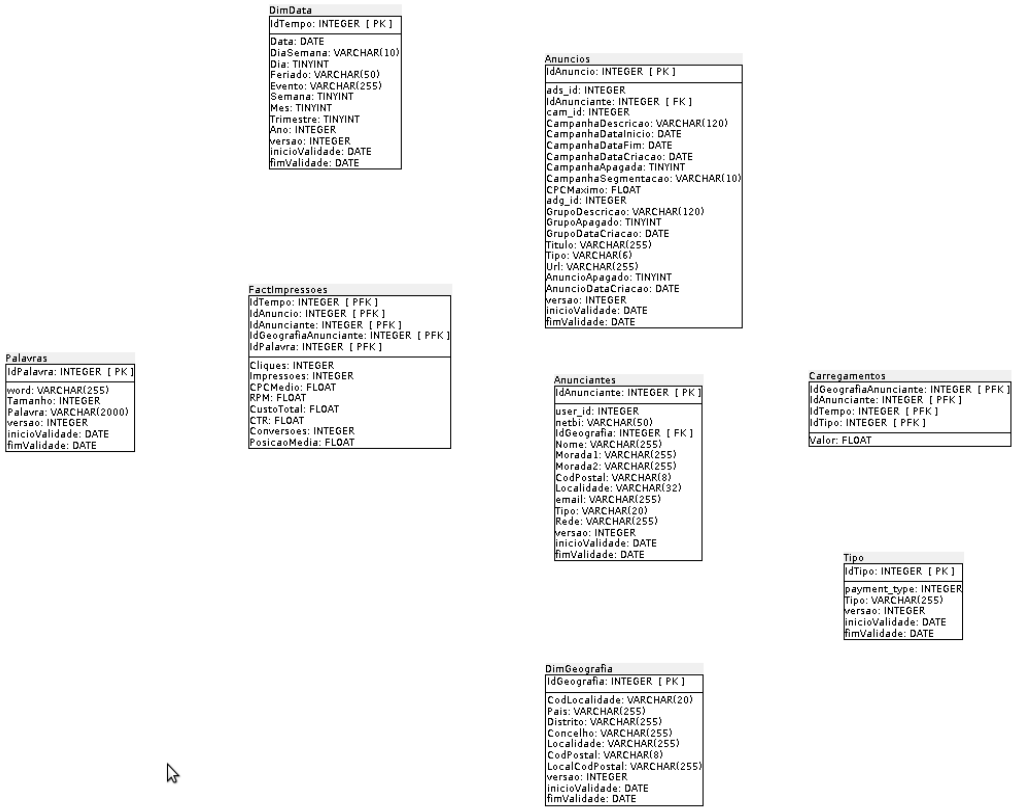


Figura 27: Modelo físico (sem agregações)

O desenho e desenvolvimento dos processos de ETL são o principal desafio na implementação de qualquer solução de BI. Nesta secção serão descritos os principais passos

tomados nesta fase.

3.7 Desenho e Desenvolvimento do Data Staging *e Processos ETL*

Após a conclusão do desenho físico do *data warehouse*, estão criadas as condições para o desenho e desenvolvimento da área de *staging* de dos processos de ETL, como se depreende da figura 1 na página 16. Este assunto é abordado na presente secção.

3.7.1 Mapa Lógico

Antes da implementação do subsistema de extração, é necessária a criação de um mapa lógico que regista a relação entre os dados dos sistemas fonte e os campos de destino no *data warehouse* [KIM04].

Através do mapa lógico identifica-se claramente a fonte, operacional ou não, mais provável para cada campo do Data Warehouse. A tabela 15 apresenta um resumo do mapa lógico assim construído. A versão completa encontra-se no Anexo G – Mapa Lógico do Data Warehouse. A sua elaboração obedeceu à estrutura proposta por Kimball em [KIM04].

Tabela do Data Warehose	**Tabela Fonte**
DimData	Ficheiro Excel (fonte externa)
DimPalavras	wrd_words
DimGeografia	Ficheiros CSV 'distritos', 'concelhos' e 'todos_cp' (fonte externa - CTT)
DimAnunciantes	usr_users, ads_portals
DimAnuncios	cam_campaigns, adg_adgroups, ads_ads
DimTipo	DATA GRID (fonte externa)
FactCarregamentos	usr_payments
FactImpressoes	usr_impressions_report

Tabela 15: Mapa Lógico do data Warehouse - Resumo

Para a elaboração do mapa lógico assim construído, foi necessário, em primeiro lugar, efetuar uma análise à qualidade dos dados nas tabelas fonte.

3.7.2 Data Profiling

Jack Olson, citado por Kimball em [KIM04], define *data profiling* como sendo o processo de análise de dados que tem como objetivo obter um conhecimento detalhado dos dados no que diz respeito ao seu conteúdo, estrutura e qualidade. Em casos extremos, quando as fontes de

dados têm falhas graves e que não podem ser usadas na persecução dos requisitos de negócio, a fase de *data profiling* pode levar ao cancelamento do projeto [KIM04].

Assim, as tarefas inerentes a esta fase devem ser executadas ainda na fase de planeamento do projeto, no sentido de determinar prazos de desenvolvimento realistas, identificar limitações nas fontes de dados e implementar procedimentos alternativos de captura de dados nas fontes [KIM04].

O SQL Power Architect foi a ferramenta utilizada para produzir os relatórios de *data profiling* que constam do Anexo H – Relatórios de Data Profiling, nos quais se baseou a construção do mapa lógico descrito na secção anterior.

Nesta subsecção apresenta-se uma síntese dos problemas encontrados em cada tabela, relativamente à qualidade dos dados, tendo em conta as fontes identificadas no mapa lógico, bem como algumas sugestões para os solucionar ou minorar.

- *Wrd_words*: Não há problemas a assinalar;
- *usr_users:* Existem diversos problemas de qualidade de dados na tabela de utilizadores, o que leva a crer que o processo de registo dos utilizadores pode ser melhorado, incluindo validações que garantam uma maior qualidade de dados:
  - Campo *usr_netbi*: 16 utilizadores não têm *username*;
  - Campo *usr_name:* Este campo usa códigos HTML para os caracteres acentuados e para caracteres especiais. O sistema de ETL terá de substituir estes códigos pelo caracter correspondente. Adicionalmente, existem muitos nomes repetidos (ex. Existem 104 utilizadores chamados “Paulo Roberto”);
  - Campo *usr_primarymail*: 5% dos registos têm este campo vazio;
  - Campo *usr_addr1:* 30% dos utilizadores não indicaram qualquer morada no seu registo, sendo que, na sua maioria indicam apenas a localidade, ou mesmo moradas inválidas;
  - Campo *usr_addr2:* 85% dos registos da tabela de utilizadores têm este campo por preencher;

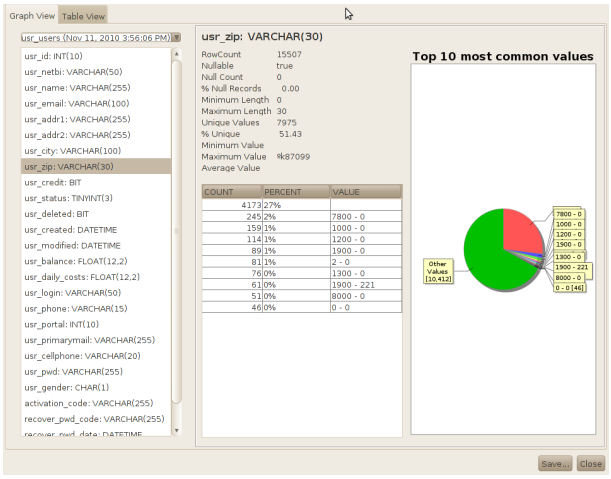


Figura 28: Data profiling – usr_users.usr_zip

- Campo *usr_city:* 25% dos registos têm este campo por preencher;
- Campo *usr_zip:* 27% dos registos têm este campo por preencher, e menos de metade contêm códigos postais válidos. Num projeto concreto, tal tornaria este campo inutilizável para efeitos de análise. Visto tratar-se de um protótipo, considera-se útil prosseguir, salientando que o processo de recolha de dados dos utilizadores necessita de ser melhorado no sistema operacional;

- *ads_portals*: Não há problemas a registar;
- *cam_campaigns*
  - Campo *cam_description*: 14% das campanhas têm “Campanha 1” como descrição;
  - Campo *cam_start_date:* 14% das campanhas têm a mesma data de início (2009-12-30);
- *adg_adgroup*
  - Campo *cam_id*: 34% dos grupos de anúncios dizem respeito à mesma campanha. Investigou-se a que anunciante pertencia esta campanha, e constatou-se que se trata de um utilizador interno da equipa comercial dos Anúncios SAPO, pelo que se chegou à conclusão que há necessidade, por parte da equipa do sistema operacional, de separar os utilizadores internos dos restantes de uma forma clara na base de dados;
  - Campo *adg_description:* 6% dos grupos de anúncios têm “Grupo 1” como descrição. A segunda descrição mais comum, “Grupo”, não ultrapassa 1%;
- *ads_ads*

- Campo *usr_id*: 36% dos anúncios pertencem ao utilizador de nome “fdgdfg”, que se presume ser um utilizador interno ou de testes. Mais de metade dos anúncios pertencem a utilizadores internos ou de testes, o que reforça a necessidade de os separar dos restantes a nível operacional;
- Campo *ads_title*: 13% dos anúncios não têm o campo correspondente ao título preenchido;
- Campo *ads_url_display*: 15% dos anúncios não têm este campo preenchido;

- *usr_payments:* Não há problemas relevantes a assinalar;
- *usr_impressions_report*: Não há problemas relevantes a assinalar.

De notar, ainda, que a descrição dos tipos de pagamento existentes, a incluir na tabela DimTipo, não estão presentes na base de dados operacional, pelo que o processo de ETL terá que os criar com base em informação fornecida pela equipa do sistema operacional.

3.7.3 Processos de ETL

Nesta secção apresenta-se uma descrição breve dos processos ETL desenvolvidos com recurso à ferramenta Pentaho Data Integration, a ferramenta selecionada para esta tarefa, como mencionado anteriormente. Trata-se de uma ferramenta de utilização muito semelhante à maioria das aplicações de ETL comerciais, e de funcionalidades, grosso modo, equivalentes. Embora o desenvolvimento seja, sobretudo, gráfico, os ficheiros resultantes são em formato XML.

Foram criados dois tipos de processos; as transformações, destinadas a povoar tabelas individuais no *data warehouse*, e os *jobs* através dos quais uma ou mais transformações são executadas, bem como outros *jobs* mais atómicos.

3.7.3.1 TransData

Através da transformação TransData, é preenchida a tabela DimData do *data warehouse*. A figura 29 apresenta esta transformação de uma forma esquemática.

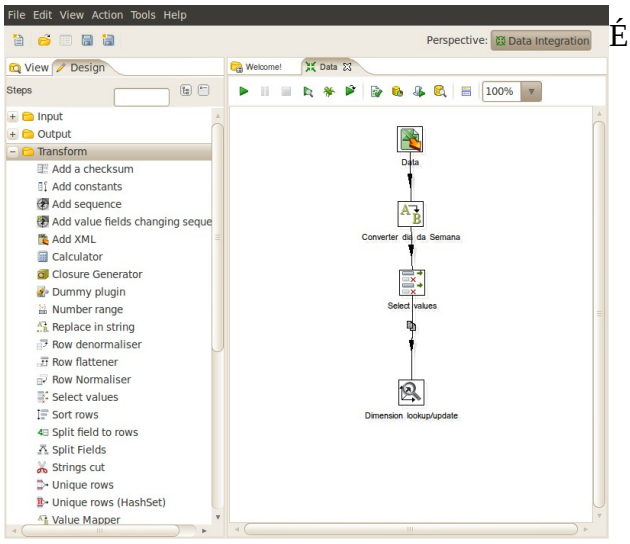


Figura 29: Transformação TransData

É constituída pelos seguintes passos:

1. 'Data' - extração dos dados de uma folha de cálculo desenvolvida para o efeito, usando o pacote de produtividade *open source* OpenOffice.org, com a estrutura da figura 30.
2. 'Converter dia da semana' – Conversão do dia da semana do formato numérico da folha de cálculo para texto;

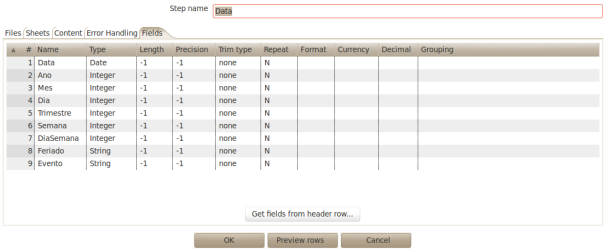


Figura 30: Estrutura da folha de cálculo 'Tempo.xls'

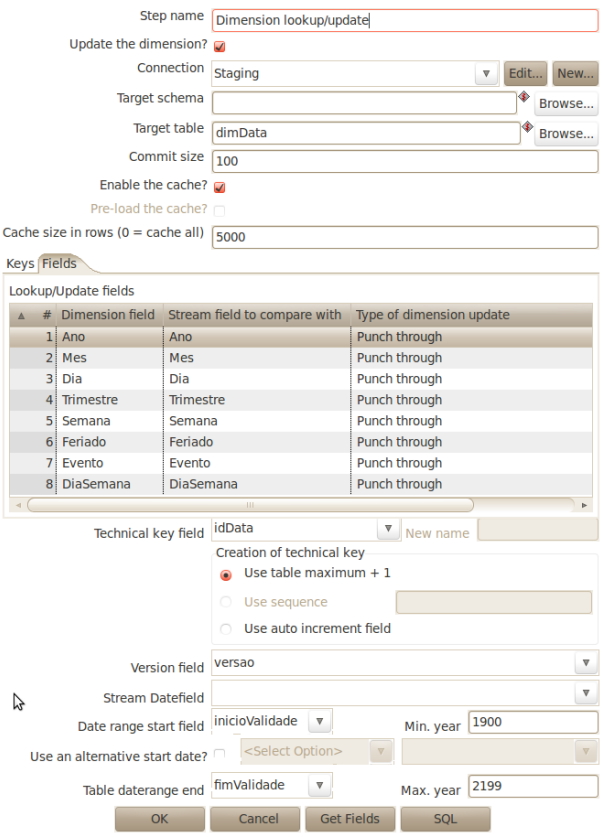


Figura 31: TransData - Dimension lookup/update

3. 'Select values' – selecionar apenas os campos a incluir na tabela, renomeando-os adequadamente e descartando os restantes.'Dimension lookup/update' – Povoamento da tabela DimData, incluindo a geração de chaves de substituição. A designação 'Punch Through', na figura 31, indica uma SCD tipo I. Também são visíveis os campos relativos à versão e datas de início e fim de validade que, teoricamente, seriam necessários apenas para SCD tipo II, mas que com este objeto do

Pentaho Integration Designer são requeridos em todos os casos.

3.7.3.2 TransPalavras e TransPalavras2

Estas transformações destinam-se a preencher a tabela DimPalavras do *data warehouse*. O preenchimento desta tabela é realizado com recurso a duas transformações devido a restrições de memória do *hardware* utilizado, que impõem a criação de um ficheiro intermédio para manipulação de alguns dados. Originalmente, a tabela DimPalavras era preenchida com recurso a uma única transformação, que, em testes, se mostrou conceptualmente correta, mas que abortava por falta de memória.

Estas transformações compreendem os seguintes passos:

1 TransPalavras

1.1 'Extrair palavras' - Extração dos dados fonte da tabela 'wrd_words' do sistema operacional;

1.2 'Text file output' – Armazenamento temporário dos dados extraídos num ficheiro CSV;

2 TransPalavras2

2.1 'Text file input' – Leitura dos dados do ficheiro CSV criado no passo anterior;

2.2 'Mudar nomes de campos' – Renomeação dos campos da *stream* para os nomes constantes na tabela DimPalavras do *data warehouse*;

2.3 'Carregar Dim Palavras' – Carregamento da tabela DimPalavras do *data warehouse*, com recurso ao objeto 'Dimension lookup/update' do Pentaho Data Integration.

3.7.3.3 TransGeografia

Tal como o nome indica, esta transformação destina-se ao povoamento da tabela DimGeografia do *data warehouse*. É composta pelos seguintes passos:

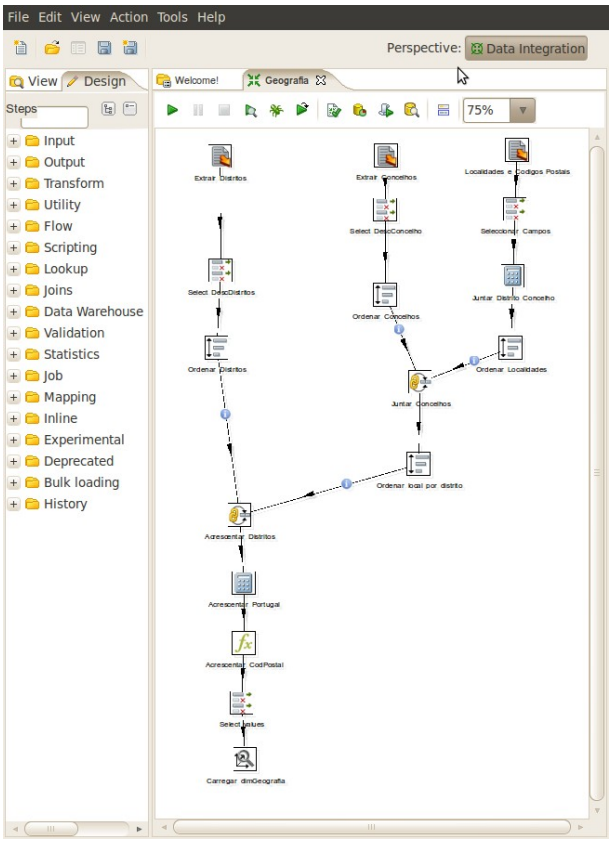


Figura 32: TransGeografia

1 'Extrair Distritos' – Extração de dados relativos aos distritos a partir do ficheiro 'distritos.txt', fornecido pelos CTT;

2 'Extrair Concelhos' - Extração de dados relativos aos concelhos a partir do ficheiro 'concelhos.txt', fornecido pelos CTT;

3 'Localidades e Codigos Postais' - Extração de dados relativos aos códigos postais e localidades a partir do ficheiro 'todos_cp.txt', fornecido pelos CTT;

4 'Select DescDistritos' – Remoção de campo redundante da *stream* proveniente do passo 1;

5 'Select DescConcelho' - Remoção de campos redundantes da *stream* proveniente do passo 2;

6 'Seleccionar Campos' – Seleção dos campos relevantes da *stream* proveniente do passo 3;

7 'Juntar Distrito Concelho' – Concatenação dos códigos de distrito e concelho, para correspondência posterior com os códigos de concelho provenientes do ficheiro 'concelhos.txt';

8 'Ordenar Distritos' – Ordenação da *stream* de distritos por código de distrito;

9 'Ordenar Concelhos' - Ordenação da *stream* de concelhos por código de concelho;

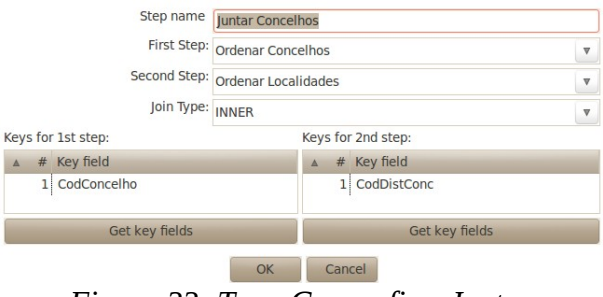


Figura 33: TransGeografia - Juntar Concelhos

10 'Juntar Concelhos' – INNER JOIN das *streams* de concelhos e de localidades e códigos postais por código de concelho (figura 33);

11 'Ordenar local por distrito' – Ordenação da lista de localidades e códigos postais obtida no passo 11 por código de distrito;

12 'Acrescentar Distritos' – INNER JOIN da *stream* de distritos com a lista de localidades e códigos postais, por código de distrito, resultante numa única *stream*;

13 'Acrescentar Portugal'

13.1 Acrescentar o campo 'País' à *stream*, com o valor 'Portugal' por omissão;

13.2 Calcular o código de localidade, concatenando os campos de código de distrito, código de concelho e código de localidade;

14 'Acrescentar CodPostal' – Calcular o código postal, concatenando os campos CP4 e CP3, correspondentes, respetivamente, aos primeiros 4 dígitos e aos últimos 3 do código postal;

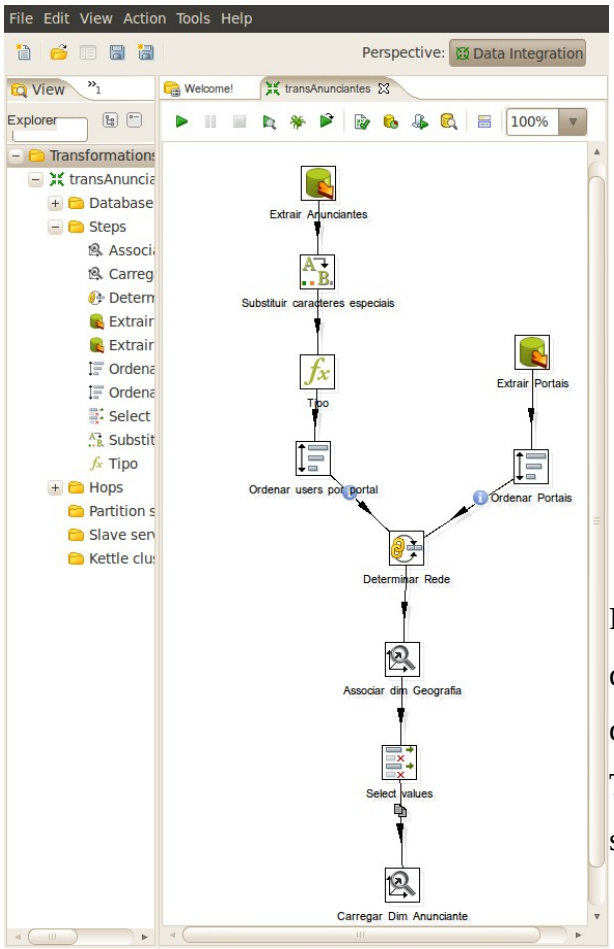


Figura 34: TransAnunciantes

15 'Select Values' – Seleccionar os campos relevantes para povoamento da dimensão Geografia

16 'Carregar dimGeografia' – Carregamento da tabela DimGeografia do *data warehouse*.

3.7.3.4 TransAnunciantes

Para o povoamento da tabela correspondente à dimensão Anunciantes no *data warehouse*, foi desenvolvida a transformação TransAnunciantes, que compreende os seguintes passos:

1. 'Extrair Anunciantes' – Extração dos campos relevantes da tabela 'usr_users';

2. 'Substituir caracteres especiais' – Substituição dos códigos HTML para caracteres especiais pelos caracteres propriamente ditos nos campos 'usr_addr1', 'usr_addr2', 'usr_city' e 'usr_name'.
3. 'Tipo' – Conversão do tipo de anunciante de valores numéricos para texto; cálculo do código postal a partir do campo 'usr_zip';
4. 'Ordenar users por portal' – Ordenação dos anunciantes extraídos por código de portal;
5. 'Extrair portais' – Extração das designações e códigos dos portais da tabela 'ads_portals';
6. 'Ordenar portais' – ordenação dos portais por código;
7. 'Determinar Rede' – INNER JOIN das duas *streams* por código de portal, daí resultante uma única *stream*, com a designação da rede do anunciante;
8. 'Associar dim Geografia' – Consulta à tabela DimGeografia do *data warehouse* para obter a chave substituta correspondente ao código postal do anunciante, associando assim as tabelas DimAnunciantes e DimGeografia. Há que assinalar que o campo 'usr_zip' da tabela 'usr_users' tem sérios problemas de qualidade de dados, pelo que num projeto concreto teria que se encontrar outra solução (por exemplo, determinar a localização do anunciante pelo seu endereço IP);
9. 'Select values' – Seleção e renomeação dos campos relevantes;
10. 'Carregar Dim Anunciante' – Carregamento da tabela DimAnunciantes do *data warehouse*'. Na figura 35 a indicação 'Punch Through' indica um campo SCD tipo I, enquanto que 'Insert' indica um campo SCD tipo II. Trata-se, assim, de uma SCD

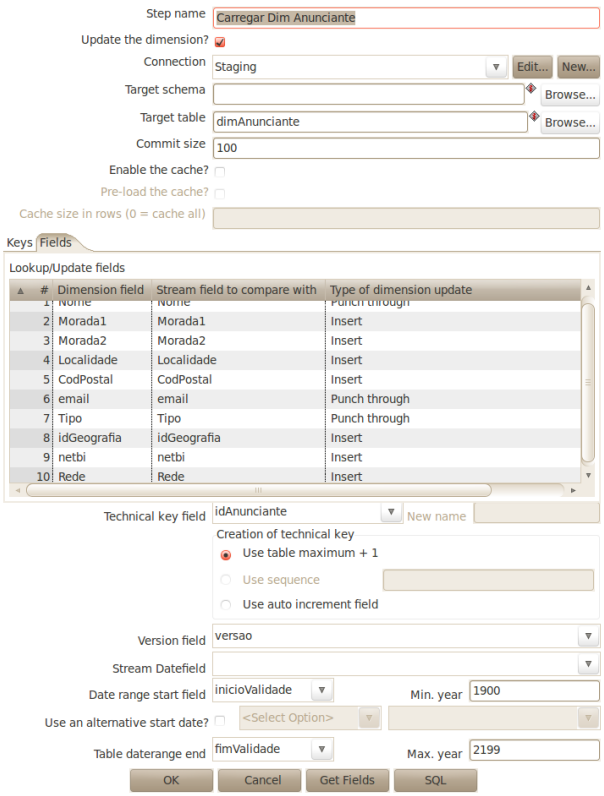


Figura 35: TransAnunciantes - Carregar Dim Anunciante

híbrida, em que a gestão de versões é feita automaticamente pelo Pentaho Data Integration.

3.7.3.5 TransAnuncios

A transformação TransAnuncios destina-se a povoar a dimensão anúncios, ou seja, a tabela DimAnuncios do Data Warehouse. Os passos descritos abaixo compõem esta transformação:

1. 'Extrair Anuncios' – Extração dos campos relevantes da tabela 'ads_ads';
2. 'Extrair Grupos' – Extração dos campos relevantes da tabela 'adg_adgroups';

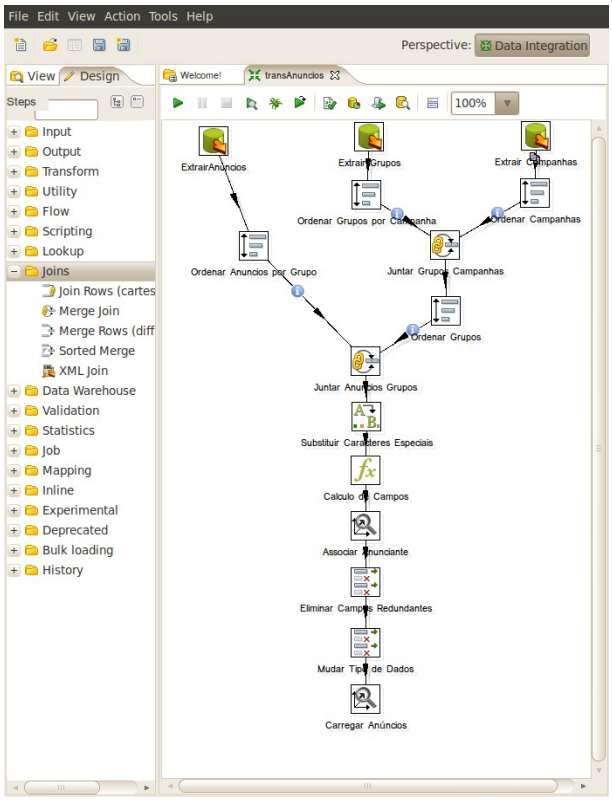


Figura 36: TransAnuncios

3. 'Extrair Campanhas' – Extração dos campos relevantes da tabela 'cam_campaigns';
4. 'Ordenar Anuncios por Grupo' – Ordenação dos anúncios extraídos por código de grupo;
5. 'Ordenar Grupos por Campanha' – Ordenação dos grupos de anúncio por código de campanha;
6. 'Ordenar Campanhas' – Ordenação das campanhas por código;
7. 'Juntar Grupos Campanhas' – INNER JOIN entre as *streams* resultantes dos passos 5 e 6, por código de campanha, daí resultando uma *stream* agregada de grupos de anúncios e campanhas;
8. 'Ordenar Grupos' – Ordenação da *stream* de grupos e campanhas, resultante do passo anterior, por código de grupo;
9. 'Juntar Anuncios Grupos' – INNER JOIN das *streams* resultantes dos passos 4 e 8 numa única, por código de grupo, daí resultante uma única *stream* com informação de anúncios, grupos e campanhas;

10. 'Substituir Caracteres Especiais' – Substituição dos códigos HTML para caracteres especiais pelos caracteres propriamente ditos para o campo 'ads_title' da tabela 'ads_ads'

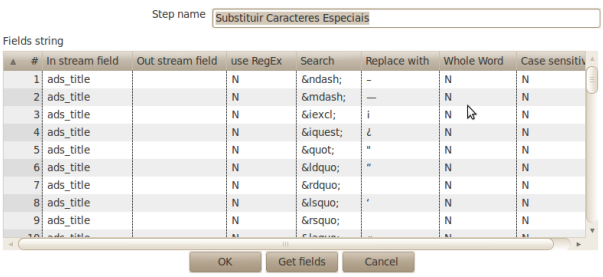


Figura 37: TransAnuncios - Substituir Caracteres Especiais

11. 'Calculo de Campos' – Conversão dos valores numéricos dos campos 'ads_type' e 'site_targeted' para texto;
12. 'Associar Anunciante' - Consulta à tabela DimAnunciante do *data warehouse* para obter a chave substituta correspondente ao user_id, associando assim as tabelas DimAnuncios e DimAnunciantes;
13. 'Eliminar Campos Redundantes' – Seleção e renomeação dos campos da *stream* a carregar no *data warehouse*;
14. 'Mudar Tipo de Dados' – Conversão de alguns campos para o tipo de dados adequado à tabela DimAnuncios;
15. 'Carregar Anuncios' – Carregamento da tabela de dimensão DimAnuncios do *data warehouse*'. A gestão das alterações na dimensão é gerida automaticamente pelo Pentaho Data Integration.

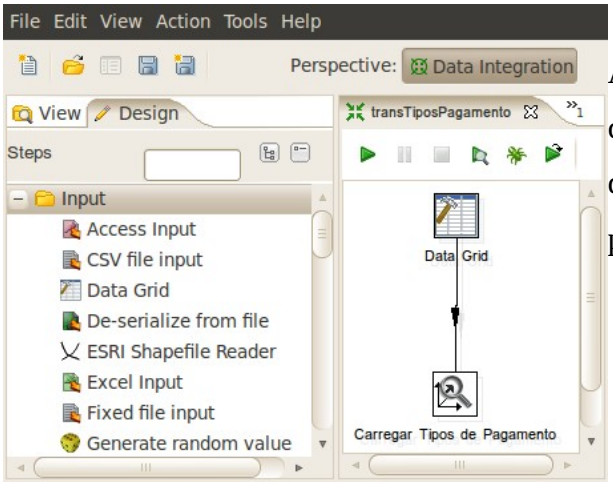


Figura 38: TransTiposPagamento

3.7.3.6 TransTiposPagamento

A transformação TransTiposPagamento, tal como o nome indica, povoa a tabela de dimensão 'DimTipo', sendo composta pelos passos seguintes:

1. 'Data Grid' – Uma vez que as descrições dos diferentes tipos de pagamento não estão disponíveis na base de dados operacional, recorreu-se ao objeto *Data Grid* do Pentaho Data

Integration, tal como consta da figura 39;

2. 'Carregar Tipo de Pagamento' – Carregamento da tabela de dimensão 'DimTipo' do *data warehouse*.

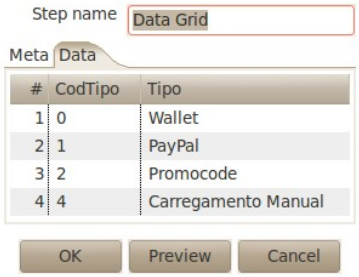


Figura 39: TransTiposPagamento - Data Grid

3.7.3.7 TransFactos

Esta transformação destina-se ao povoamento da tabela de factos do *data warehouse* 'FactImpressões', e compreende os seguintes abaixo. Assinala-se que o objeto utilizado nas consultas às dimensões (Dimension lookup/update), nesta e noutras transformações, acrescenta a chave substituta da versão mais atual do registo da dimensão. Existe também uma opção para obter a chave substituta de um registo em vigor numa data especificada [CAS10].

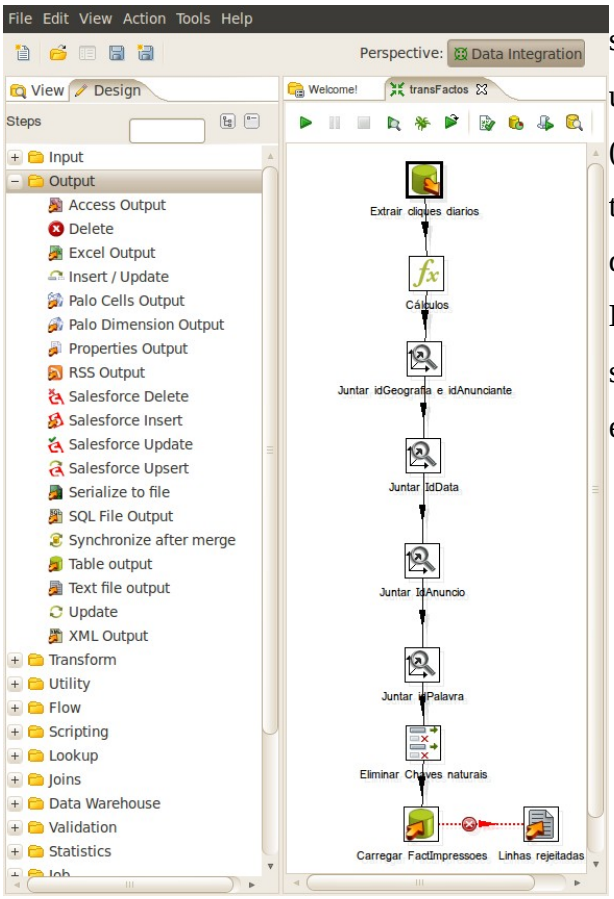


Figura 40: TransFactos

1 'Extrair Cliques Diarios' – Extração dos campos relevantes da tabela 'usr_impressions_report';

2 'Cálculos' – Cálculo dos valoresCPC Médio, RPM e CTR, de acordo com as fórmulas constantes na figura 41;

3 Surrogate Key Pipeline [KIM04]

3.1 'Juntar idGeografia e idAnunciante' – Consulta à tabela DimAnunciante do *data warehouse* para obtenção das chaves substitutas idGeografia e idAnunciante;

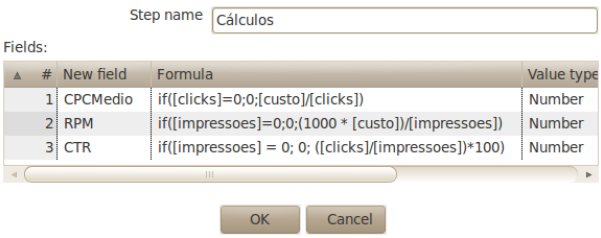


Figura 41: TransFactos - Cálculos

3.2 'Juntar idData' – Consulta à tabela DimData do *data warehouse* para obtenção da chave substituta idData;

3.3 'Juntar idAnuncio' - Consulta à tabela DimAnuncios do *data warehouse* para obtenção da chave substituta idAnuncio;

3.4 'Juntar idPalavra' – Consulta à tabela DimPalavras do *data warehouse* para obtenção da chave substituta idData;

4 'Eliminar Chaves naturais' – Seleção dos campos da *stream* a carregar na tabela de factos 'FactImpressoes';

5 'Linhas rejeitadas' – Registo das linhas rejeitadas pelo processo ETL num ficheiro de texto CSV.

3.7.3.8 TransFactosPagamentos

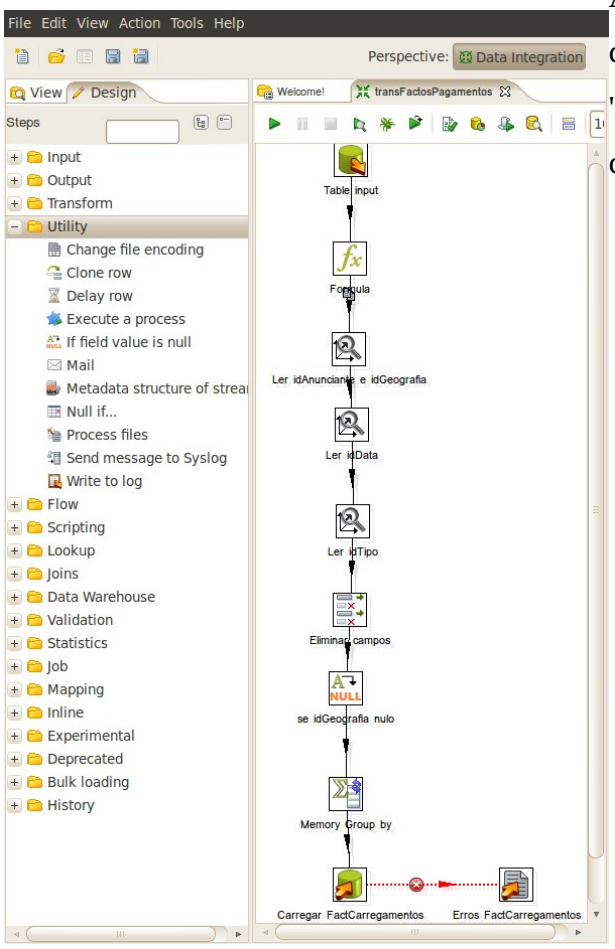


Figura 42: TransFactosPagamentos

A transformação TransFactosPagamentos tem como objetivo povoar a tabela de factos 'FactCarregamentos' do *data warehouse*, sendo constituída pelos passos abaixo:

1 'Table input' – Extração dos campos relevantes da tabela 'usr_payments';

2 'Formula' – Cálculo do tipo e data de pagamento;

3 'Surrogate Key Pipeline' [KIM04]

3.1 'Ler idAnunciante e idGeografia' - Consulta à tabela DimAnunciante do data warehouse para obtenção das chaves substitutas idGeografia e idAnunciante;

3.2 'Ler idData' - Consulta à tabela DimData do *data warehouse* para

obtenção da chave substituta idData;

3.3 'Ler idTipo' - Consulta à tabela DimTipo do *data warehouse* para obtenção da chave substituta idTipo;

4 'Eliminar campos' – Seleção e renomeação dos campos relevantes a carregar na tabela do *data warehouse*;

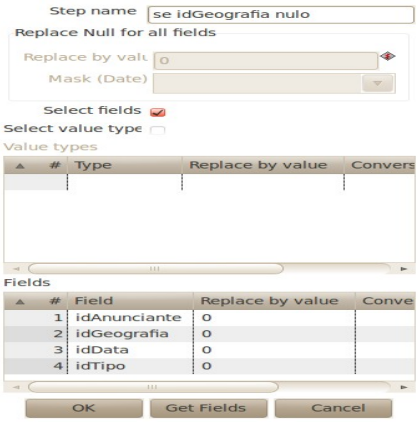


Figura 43: TansFactosPagamentos - se idGeografia nulo

5 'Se idGeografia nulo' – Tratamento de eventuais valores nulos (substituição por 0) provenientes das consultas às tabelas de dimensão na *surrogate key pipeline* (figura 44);

6 'Memory Group By' – Agregação (soma) do valor dos pagamentos agrupados por anunciante, data, geografia e tipo (figura 44);

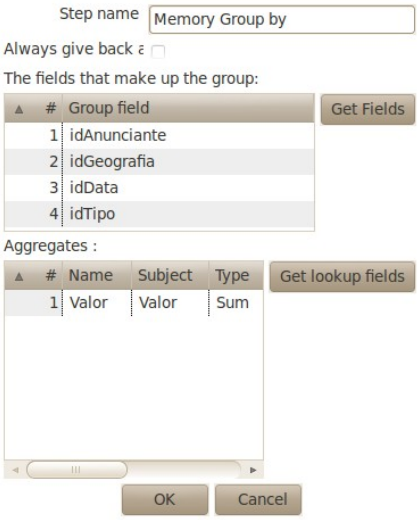


Figura 44: TransFactosPagamentos - Memory Group by

7 'Carregar FactCarregamentos' – Povoamento da tabela 'FactCarregamentos' do *data warehouse*;

8 'Erros Factos Carregamentos' – Registo, num ficheiro de texto CSV, das linhas rejeitadas pelo processo de ETL relativas a esta transformação.

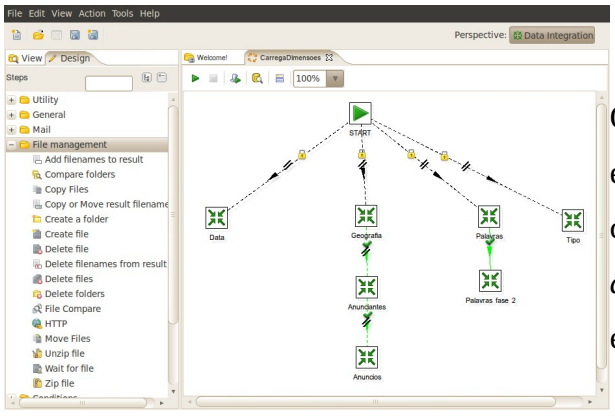


Figura 45: CarregaDimensoes

3.7.3.9 CarregaDimensoes

O *job* CarregaDimensoes é responsável pela execução das transformações envolvidas no carregamento das tabelas de dimensão do *data warehouse*, executando-as pela ordem especificada na figura 45.

Como se pode observar na figura, algumas das transformações são executadas em paralelo, no sentido de melhorar o desempenho do processo, enquanto que outras, que só podem ser executadas após transformações anteriores, são-no em sequência.

3.7.3.10 CarregaDataMart

Por último, o *job* CarregaDataMart desencadeia todo o processo de ETL acima descrito, executando o *job* CarregaDimensoes e, seguidamente e em paralelo, as duas transformações responsáveis pelo povoamento das tabelas de factos, como é visível na figura 46.

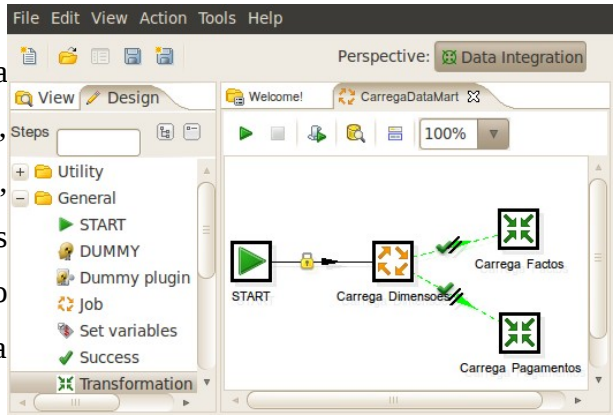


Figura 46: CarregaDataMart

3.8 Especificação Analítica da Aplicação

Neste ponto, uma vez executados os processos de ETL, o *data warehouse* está povoado, mas ainda não está pronto para ser aberto aos utilizadores. Para isso, é necessário especificar, construir e/ou configurar as aplicações finais de exploração de dados.

3.8.1 Configuração e Especificação do Esquema

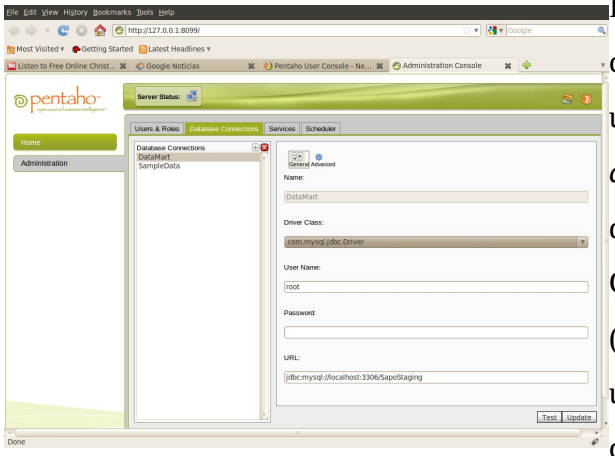


Figura 47: Pentaho Administration Console

Para atingir este objetivo, começou-se por configurar a plataforma de BI Pentaho com uma nova fonte de dados, correspondente ao *data warehouse*. Para esse efeito, a plataforma disponibiliza a Pentaho Administration Console, onde, para além de fontes de dados (figura 47), é possível configurar contas de utilizadores, definir níveis de segurança e calendarizar a execução de *jobs* ETL [PEN09].

Uma vez configurada a fonte de dados para o *data warehouse*, é já possível produzir relatórios *ad-hoc* através da User Console disponibilizada pela plataforma. Só será, no entanto, possível realizar análises OLAP com a especificação de um esquema (*schema*). Um esquema é, no

Mondrian, o servidor OLAP usado pela plataforma, a definição de uma base de dados multi-dimensional. Contém um modelo lógico constituído por um ou mais hipercubos (correspondendo, grosso modo um cubo a um esquema em estrela), hierarquias e níveis, bem como o mapeamento deste modelo para o modelo físico do *data warehouse*[HYD09]. O modelo lógico é, assim, constituído pelas estruturas usadas em *queries* MDX, uma linguagem semelhante ao SQL, mas utilizada em *queries* OLAP [WIK10].

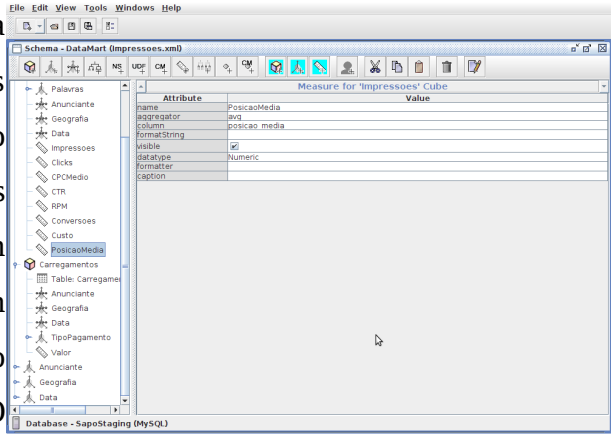


Figura 48: Schema Workbench

O esquema é especificado usando uma complexa estrutura em XML, cujos detalhes podem ser consultados em [HYD09]. No sentido de simplificar a tarefa de especificação do esquema, a Pentaho disponibiliza a ferramenta Schema Workbench, uma ferramenta visual que permite a especificação e teste de esquemas Mondrian de uma forma mais intuitiva [WOO07]. A figura 48 ilustra um aspeto da especificação do esquema usando o Schema Workbench. Um tutorial completo pode ser consultado em [WOO07]. O Anexo I – Esquema do Mondrian, apresenta a especificação completa do esquema desenvolvido, sem agregações.

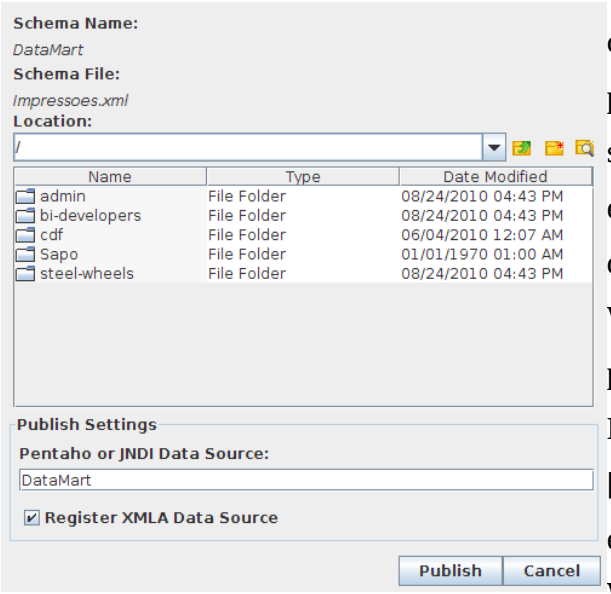


Figura 49: Schema Workbench - Publicação de um esquema

Para que o esquema assim criado fique disponível para a realização de análises OLAP, é ainda necessário publicá-lo na plataforma de BI. A figura 49 ilustra o processo. A documentação *online* da Pentaho contém um tutorial em [BAK10] onde este procedimento é detalhado.

Após publicação bem sucedida do esquema, este passou a estar disponível na plataforma para

realização de análises.

3.8.2 Agregações

Ao contrário de outros servidores OLAP, o Mondrian não armazena dados em disco; apenas trabalha com os dados do SGBD, constituindo um motor ROLAP (OLAP Relacional) puro. Uma vez lido um segmento de dados, este é armazenado na cache. Isto simplifica o processo de instalação do Mondrian, mas limita o seu desempenho quando tem que lidar com grandes volumes de dados, como é o caso do cubo relativo aos cliques e impressões, cujas análises mostram um claro problema de desempenho com o *hardware* utilizado. Nestes casos, é conveniente utilizar tabelas de agregações, ou seja, um conjunto de dados sumarizados pré-calculados [EMB08].

Assim, decidiu-se pela criação de um conjunto de tabelas de agregação para o cubo Impressões. Podem ser construídas centenas ou milhares de tabelas de agregações, com todas as combinações possíveis de dimensões, hierarquias e níveis, pelo que é impossível construir todas. O ideal seria gerar tabelas que correspondessem às análises mais frequentes, mas não existe experiência de utilização que permita determiná-las.

Relativamente a este último ponto, o Mondrian possui a funcionalidade "Aggregation Generator" (AggGen), que gera código SQL para o suporte e criação de tabelas de agregação, isto é, para uma determinada *query* MDX executada pelo Mondrian, o AggGen gera as instruções CREATE e INSERT ótimas para essa *query*, que podem ser posteriormente usadas para a criação e manutenção de uma tabela de agregação, caso tal se justifique [WOH07].

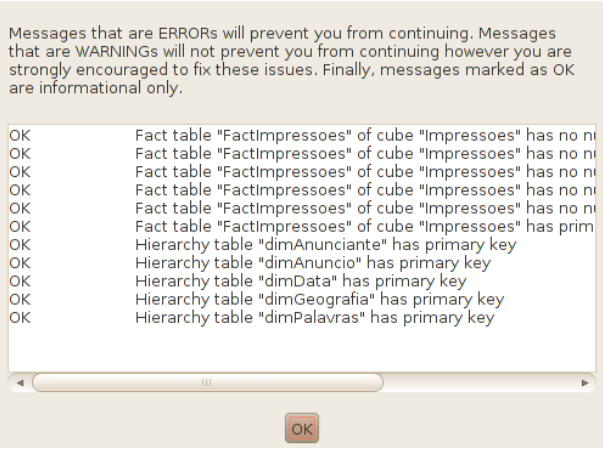


Figura 50: Aggregation Designer - Validação do cubo 'Impressões'

No sentido de endereçar este problema, a plataforma Pentaho disponibiliza a aplicação Aggregation Designer, que permite definir as agregações a produzir. Tem também a funcionalidade "Advisor" que, através de um algoritmo "inteligente" que analisa a relação custo/benefício de cada agregação, faz recomendações acerca das agregações a criar

[PEN08].

Decidiu-se, assim, pela criação de um comportável número de 10 tabelas de agregação, construídas com a funcionalidade “Advisor” do Pentaho Aggregation Designer para o cubo “Impressões”. Considera-se que o cubo relativo aos carregamentos não necessita de agregações, uma vez que tem um volume de dados mais reduzido.

Ao iniciar o processo, o Aggregation Designer faz a validação automática da especificação do cubo selecionado, como é visível na figura 50.

Na opção “Adviser” da ferramenta especifica-se o número máximo de agregações a gerar, e/ou o tempo máximo de execução do algoritmo (figura 51).

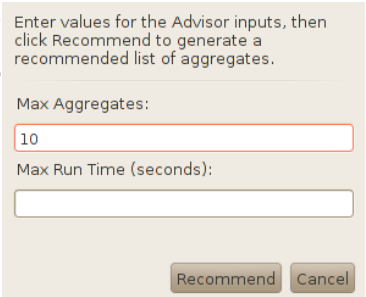


Figura 51: Aggregation Designer - Advisor

Após execução, é sugerido um conjunto de tabelas de agregação, segundo a análise custo/benefício que consta da figura 52.

A opção “Export” do Aggregation Designer permite exportar as instruções SQL para criação e povoamento das tabelas, ou executá-las diretamente sobre o *data warehouse*, bem como guardar e publicar o esquema atualizado no Mondrian.

O Mondrian tem em conta um conjunto de convenções relativas aos nomes das tabelas de agregação, bem como das suas colunas. As tabelas que não obedeçam a estas convenções não serão utilizadas, a menos que tenham sido explicitamente declaradas na especificação do cubo [EMB08]. O Aggregation Designer gera tabelas que não seguem estas convenções, e, embora a especificação do cubo seja de facto atualizada, não foi possível conseguir que o Mondrian as reconhecesse. Foi, assim, necessário renomear as tabelas, bem como as suas colunas, de forma a que ficassem conformes às convenções estabelecidas, e ainda republicar o esquema antigo, sem declarações explícitas de agregações.

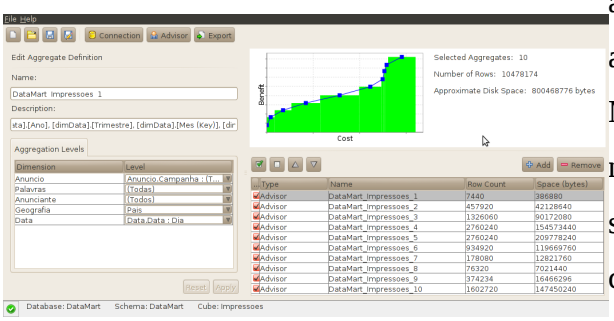


Figura 52: Aggregation Designer - Recomendações

Em testes, ainda assim, persistiram algumas falhas nas *queries ad-hoc*, que nos *logs* não eram

claras, que mais tarde se verificou tratarem-se de erros na especificação do esquema. Uma vez corrigidos, obteve-se uma melhoria significativa no desempenho da execução das *queries* MDX, devido à utilização das tabelas de agregação. No entanto, dadas as limitações impostas pelo *hardware* utilizado, elas permaneceram com um desempenho aquém do aceitável, não sendo possível prever o seu impacto numa máquina mais potente.

3.8.3 Análises Pré-Definidas

Neste ponto, definem-se as análises que irão dar resposta às questões de negócio referidas em 3.3.3 - Requisitos.

Na prática, tratou-se apenas de navegar intuitivamente pelos dados usando o navegador OLAP do Mondrian até se obter a resposta, e gravar a *query* MDX resultante. Foi, portanto, um processo linear e simples, que serviu, não só para desenvolver as análises que satisfizessem os requisitos estabelecidos, mas para validar os resultados junto dos utilizadores de negócio.

Assim, as respostas às questões enunciadas anteriormente resultam nas análises pré-definidas que se seguem, referidas a título de exemplo.

3.8.3.1 Quais as redes que mais anúncios serviram por período de tempo?

Para esta análise, no navegador OLAP seleciona-se a métrica 'Impressões' e as dimensões 'Data' e 'Rede', com o intervalo de datas para o qual existem dados e para as redes válidas (não 'NULL'). A figura 53 ilustra a análise. De notar que as designações das redes, à exceção do SAPO, estão ocultadas por razões de confidencialidade.

A gravação da análise é efetuada num ficheiro *.xaction*, num formato XML, no qual estão incluídos, para além da *query* MDX abaixo, outros elementos relativos à visualização da

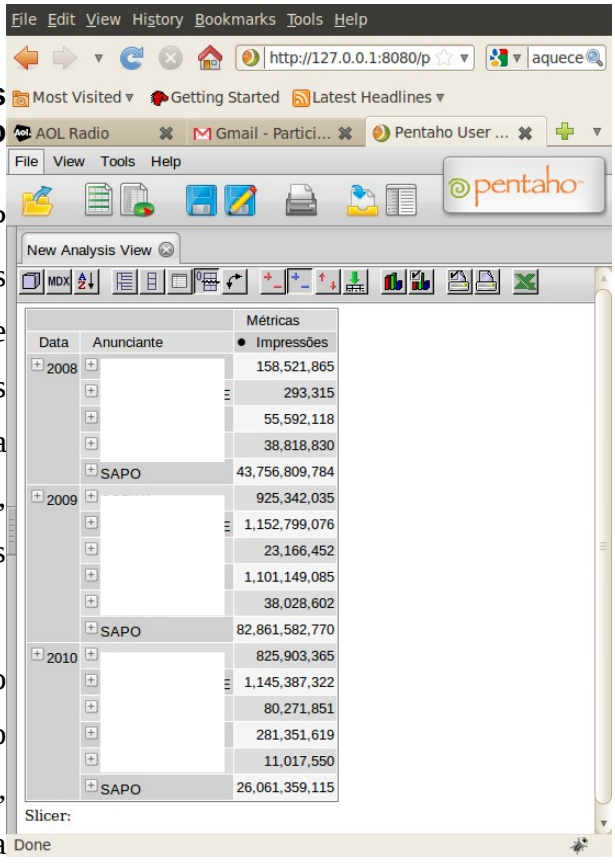


Figura 53: Evolução de impressões por rede

informação.

```
select NON EMPTY {[Measures].[Impressoes]} ON COLUMNS,
  NON EMPTY Crossjoin({[Data].[2007], [Data].[2008], [Data].[2009], [Data].[2010]},
{[Anunciante].[###], [Anunciante].[###], [Anunciante].[###], [Anunciante].[###], [Anunciante].
[###], [Anunciante].[###], [Anunciante].[SAPO]}) ON ROWS
from [Impressoes]
```

De notar que esta informação pode também ser visualizada sob a forma de gráfico e exportada para vários formatos, incluindo .xls, como é visível na figura.

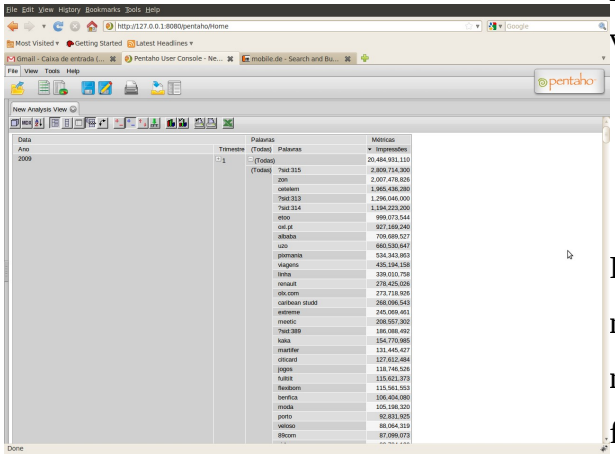


Figura 54: Impressões por palavra e data

3.8.3.2 Qual o top 5 de palavras em termos de impressões, cliques, CTR e CPC médio em 2009?

Para esta análise são relevantes as métricas referidas e as dimensão 'Data' e 'Palavras'. O resultado da análise para impressões consta da figura 54.

3.8.3.3 Qual o top 5 de anunciantes nas diferentes redes por período de tempo?

Esta análise cruza as dimensões 'Data' e 'Anunciante', resultante na análise na figura 55, em que são comparados os cliques e impressões por anunciante em 2009 (na imagem, apenas é visível o SAPO, por motivo de espaço). Por razões de confidencialidade, os nomes dos anunciantes individuais foram ocultados na imagem.

Esta análise é produzida pela *query* MDX abaixo.

```
select NON EMPTY {[Measures].[Impressoes],
[Measures].[Clicks]} ON COLUMNS,
```

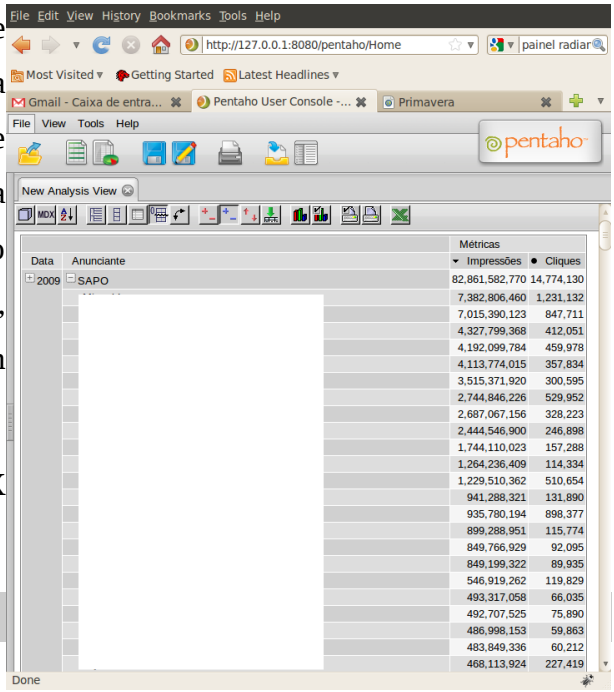
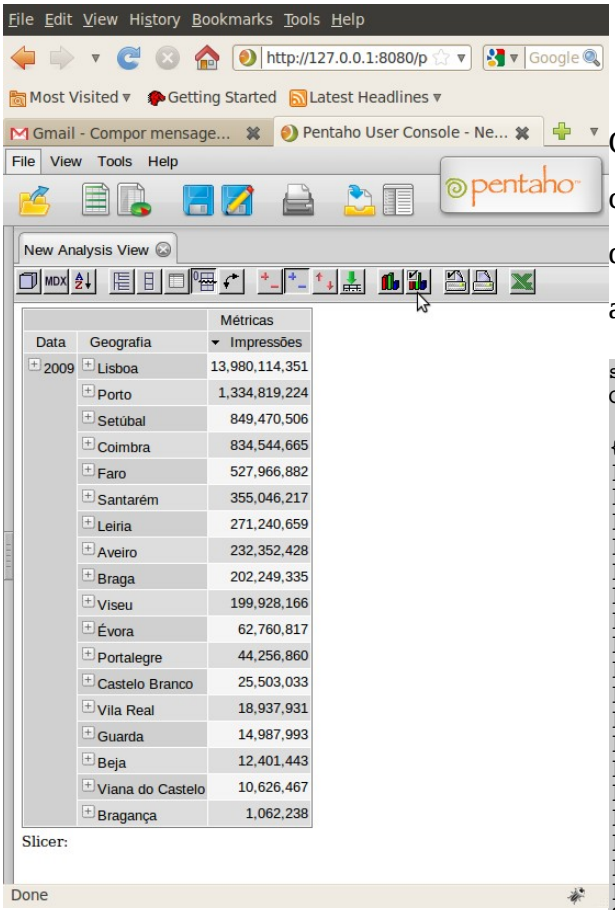


Figura 55: Impressões e cliques por rede e data

```
  NON EMPTY Hierarchize(Union(Union({([Data].[Todas], [Anunciante].[Todos])},
Crossjoin([Data].[Todas].Children, {[Anunciante].[Todos]})), Crossjoin({[Data].[2009]},
[Anunciante].[Todos].Children))) ON ROWS
from [Impressoes]
```



		Métricas
Data	Geografia	Impressões
2009	Lisboa	13,980,114,351
	Porto	1,334,819,224
	Setúbal	849,470,506
	Coimbra	834,544,665
	Faro	527,966,882
	Santarém	355,046,217
	Leiria	271,240,659
	Aveiro	232,352,428
	Braga	202,249,335
	Viseu	199,928,166
	Évora	62,760,817
	Portalegre	44,256,860
	Castelo Branco	25,503,033
	Vila Real	18,937,931
	Guarda	14,987,993
	Beja	12,401,443
	Viana do Castelo	10,626,467
	Bragança	1,062,238

Figura 56: Impressões por distrito em 2009

3.8.3.4 Quais os distritos cujos anunciantes deram origem a mais impressões em 2009?

Como o título indica, nesta análise são combinadas as dimensões 'Data' e 'Geografia', dando origem à figura 56 e à *query* MDX abaixo.

```
select NON EMPTY {[Measures].[Impressoes]} ON
COLUMNS,
  NON EMPTY Order(Crossjoin({[Data].[2009]},
{[Geografia].[Portugal].[Aveiro
], [Geografia].[Portugal].[Beja
], [Geografia].[Portugal].[Braga
], [Geografia].[Portugal].[Bragança
], [Geografia].[Portugal].[Castelo Branco
], [Geografia].[Portugal].[Coimbra
], [Geografia].[Portugal].[Évora
], [Geografia].[Portugal].[Faro
], [Geografia].[Portugal].[Guarda
], [Geografia].[Portugal].[Leiria
], [Geografia].[Portugal].[Lisboa
], [Geografia].[Portugal].[Portalegre
], [Geografia].[Portugal].[Porto
], [Geografia].[Portugal].[Santarém
], [Geografia].[Portugal].[Setúbal
], [Geografia].[Portugal].[Viana do Castelo
], [Geografia].[Portugal].[Vila Real
], [Geografia].[Portugal].[Viseu
]}), [Measures].[Impressoes], DESC) ON ROWS
from [Impressoes]
```

3.8.3.5 Quais os meios de pagamento mais usados em 2009?

Voltando a atenção para o cubo 'Carregamentos' do *data warehouse*, introduz-se também uma outra funcionalidade do Mondrian, isto é, a realização de análises gráficas, mais úteis para informação agregada.

Assim, nesta análise, podemos ver o cruzamento das dimensões 'Data' e 'Tipo de Pagamento', que é visível na figura 57, onde se podem ver os meios de pagamento usados em 2009 ordenados por ordem crescente. A *query* MDX daí resultante é a que se pode ver abaixo.

```
select NON EMPTY {[Measures].[Valor]} ON
COLUMNS,
    NON EMPTY Order(Crossjoin({[Data].[2009]},
{[TipoPagamento.Tipo].[Carregamento Manual],
[TipoPagamento.Tipo].[PayPal],
[TipoPagamento.Tipo].[Promocode],
[TipoPagamento.Tipo].[Wallet]}), [Measures].
[Valor], ASC) ON ROWS
from [Carregamentos]
```

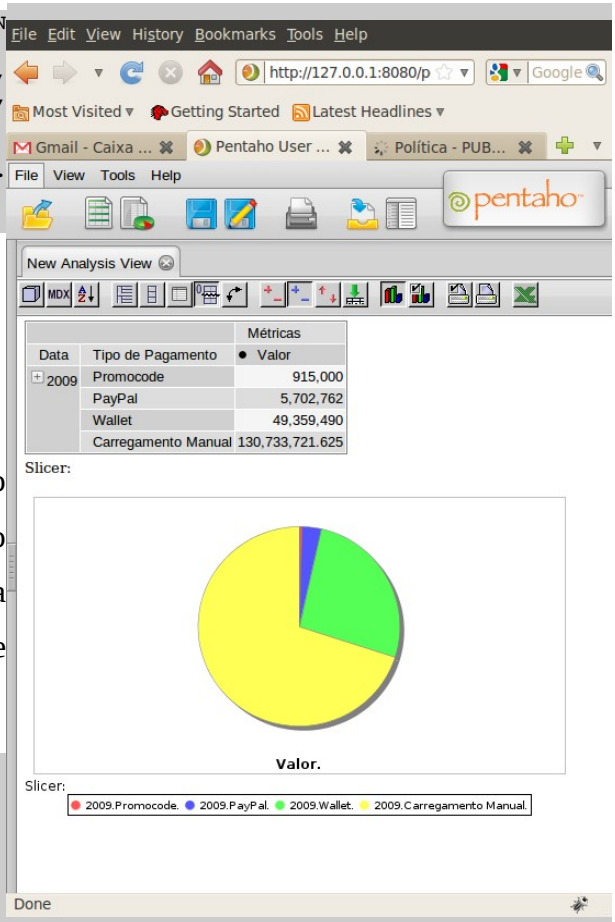


Figura 57: Meios de pagamento em 2009

3.8.3.6 Como evoluiu a utilização PayPal ao longo do tempo na rede SAPO?

Focando agora um meio de pagamento específico, analisamos a sua evolução ao longo do tempo, produzindo-se o gráfico da figura 58. Abaixo pode ver-se a *query* MDX que produz este resultado.

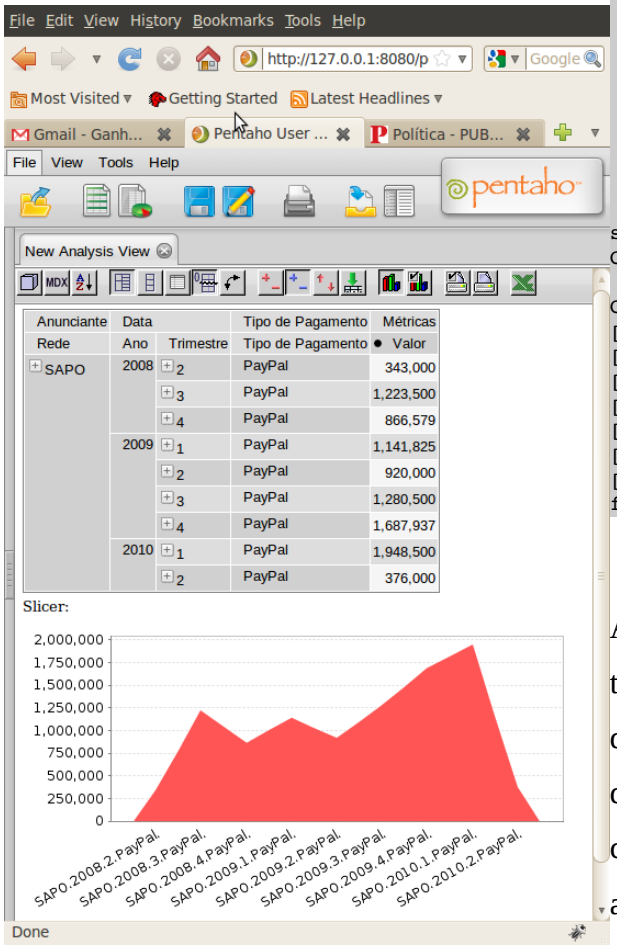


Figura 58: Evolução do PayPal na rede SAPO

```
select NON EMPTY {[Measures].[Valor]} ON
COLUMNS,
  NON EMPTY Crossjoin({[Anunciante].[SAPO]},
Crossjoin({[Data].[2007].[1], [Data].[2007].
[2], [Data].[2007].[3], [Data].[2007].[4],
[Data].[2008].[1], [Data].[2008].[2], [Data].
[2008].[3], [Data].[2008].[4], [Data].[2009].
[1], [Data].[2009].[2], [Data].[2009].[3],
[Data].[2009].[4], [Data].[2010].[1], [Data].
[2010].[2], [Data].[2010].[3], [Data].[2010].
[4]}, {[TipoPagamento.Tipo].[PayPal]})) ON ROWS
from [Carregamentos]
```

3.9 Produção

A fase de entrada em produção inclui, tipicamente, a realização de testes integrados de todos os componentes da arquitetura e ações de formação dos utilizadores, antes da concessão do acesso do *data warehouse* e aplicações associadas àqueles.

Tratando-se de uma prova de conceito que não se destina a entrar em produção, as únicas ações relacionadas estas atividades foram a realização de alguns testes para garantir a estabilidade da solução, em que os pequenos problemas detetados foram corrigidos, e uma apresentação a alguns utilizadores chave, com perfil técnico e comercial, em que se conseguiu a sua sensibilização para as vantagens das soluções de BI em geral, uma vez que as vantagens das soluções *open source* são por demais conhecidas na unidade orgânica Anúncios SAPO da PT Comunicações, com aplicações críticas de negócio apoiadas em *software* aberto em utilização diária.

4 Conclusão

Neste capítulo, que pretende fazer uma síntese do trabalho desenvolvido, interessa, em primeiro lugar, recuperar as questões e objetivos de investigação, apresentados na Tabela 1: Questões e Objetivos de Investigação , que aqui se reproduz novamente:

	Questões de Investigação	*Objetivos de Investigação*
1.	Porque motivos poderão as organizações adotar uma solução BI open source em detrimento das soluções proprietárias?	Determinar as vantagens das soluções BI *open source* em relação às soluções tradicionais.
2.	Porque motivos não são estas soluções adaptados com maior frequência?	Determinar os riscos acrescidos e objeções, se existirem, inerentes às soluções de BI *open source*.
3.	Como se organizam, trabalham e obtêm receitas as empresas produtoras de soluções de BI open source?	Determinar o modelo de negócio das empresas produtoras de soluções de BI *open source*.
4.	Como justificar a implementação de uma solução deste tipo numa grande empresa portuguesa?	Implementar de um protótipo com base em ferramentas de BI *open source* numa grande empresa portuguesa no sentido de demonstrar ou refutar a sua viabilidade.

Tabela 16: Questões e Objetivos de Investigação Recuperados

Assim, respondendo à primeira questão de investigação, as organizações poderão adotar uma solução BI *open source* em detrimento de soluções proprietárias essencialmente pelos seguintes motivos, analisados em detalhe em 2.3 - Vantagens Open Source, que se prendem com as vantagens do *open source* em geral:

- Não existem barreiras à adoção[7]: qualquer pessoa pode descarregar a solução da Internet, experimentá-la e começar a implementá-la nos seus projetos – tudo sem o controlo dos vendedores;
- Os custos iniciais e totais são baixos;
- O pagamento, quando existe, é baseado na utilização real, sem *shelfware;*
- Não existem custos de *runtime;*
- Grande oferta de especialistas no mercado, sem necessidade de recorrer aos serviços de consultoria dos vendedores.

Relativamente à questão 2, importa em primeiro lugar desmistificar alguns mitos relativos às soluções *open source*, que podem constituir autênticas barreiras à adoção, analisados em

7 Em termos de custos e disponibilidade das soluções. Poderão existir barreiras de outros tipos, como se discutirá mais tarde.

detalhe em 2.4 - Mitos. Ali, verificou-se que o *software open source* não é necessariamente grátis; que não se trata apenas de uma moda passageira, sendo utilizado de uma forma generalizada; que o seu desenvolvimento é rigorosamente controlado; que existem opções de suporte adequadas a estas estas soluções, nomeadamente através do chamado *open source* comercial; e que existem aplicações críticas de negócio suportadas exclusivamente em *software open source*, constituindo soluções seguras.

Existem, no entanto, alguns tipos de licenças *open source* que podem não ser adequadas para uso comercial, e há sempre que considerar custos indiretos, como sejam formação e outros serviços.

Em 2.5 - Barreiras à Adoção identificaram-se algumas barreiras à adoção que podem, de facto, ser difíceis de ultrapassar. Algumas ferramentas *open source* apresentam problemas de instalação, configuração, fiabilidade e escalabilidade, que revelam que algumas soluções ainda não atingiram a maturidade necessária à sua utilização em grandes projetos Pode, no entanto, dizer-se o mesmo em relação a alguns produtos proprietários, pelo que este tipo de problemas não pode considerar-se exclusivo de aplicações *open source*.

Face ao exposto, dado o carácter grátis ou de custos comparativamente muito mais baixos destas soluções, coloca-se a questão de investigação 3, relacionada com a rentabilidade das empresas que propõem estas soluções. Concluíu-se, em 2.8 - Modelos de Negócio, que as estratégias mais comuns em termos de obtenção de receitas usadas pelas empresas que atuam na área do *open source*, são a venda de licenças de versões comerciais dos seus produtos, bem como a prestação de serviços de formação, suporte e consultoria.

Estas empresas, de acordo com a informação disponível, parecem ser sólidas e fiáveis, embora seja difícil avaliar a sua sustentabilidade financeira, dado tratarem-se, na sua maioria, de empresas não cotadas, que não divulgam os seus dados financeiros. A JasperSoft e a Pentaho, as empresas que distribuem as mais populares soluções de BI *open source* possuem uma extensa rede de parceiros tecnológicos, revendedores e integradores de sistemas, composta por grandes empresas com presença mundial, e também por distribuidores locais, alguns presentes em Portugal.

Com a prova de conceito realizada, descrita em 3 - Prova de Conceito, procurou-se responder à quarta questão de investigação, demonstrando-se a viabilidade de uma destas soluções num

ambiente empresarial, usando o Linux como sistema operativo, MySQL como sistema de gestão de bases de dados e as versões *open source* das ferramentas Pentaho BI Suite para os restantes componentes.

A experiência de utilização deste pacote de ferramentas revelou ser bastante semelhante à de outros pacotes comerciais existentes no mercado, existindo, no entanto, alguns problemas de compatibilidade com uma das ferramentas do pacote (o Pentaho Aggregation Designer, que gera tabelas de agregação de dados cujos nomes não seguem as convenções estabelecidas para o servidor OLAP da *suite*). A especificação dos cubos e hierarquias também não é tão intuitiva como noutras ferramentas, mas em todos os passos existe ampla documentação de suporte, assim como fóruns de utilizadores, com recurso aos quais foi possível resolver ou contornar todos os problemas que foram surgindo na construção da solução.

De assinalar que a própria cultura empresarial da unidade de negócio Anúncios SAPO da PT Comunicações em muito contribuiu para o sucesso da iniciativa, dada a sua forte abertura a soluções baseadas em *software open source*, utilizado em larga escala na organização. Em organizações mais "tradicionais", poderiam existir dificuldades acrescidas em termos de credibilização e aceitação de uma solução baseada em ferramentas abertas, seja pela questão cultural referida, seja por investimentos anteriores efetuados noutras tecnologias.

4.1 Limitações

Apesar de ter sido possível, *grosso modo*, concretizar os objetivos de investigação propostos inicialmente, as conclusões apresentadas devem ser lidas à luz de algumas limitações impostas essencialmente ao nível do desenvolvimento da prova de conceito. Nomeadamente:

- Não foi possível avaliar adequadamente o desempenho da solução em termos de tempo de execução das *queries* MDX devido às limitações do *hardware* utilizado, descrito em 3.5.1 - Hardware;
- Tratando-se de um protótipo, o âmbito da iniciativa foi reduzido, consistindo em apenas dois hipercubos e algumas dimensões;
- A granularidade dos registos nas tabelas de factos do *data warehouse* é elevada, sendo os dados agregados por dia, o que é insuficiente nesta área de negócio;

- A análise à qualidade dos dados dos sistemas fonte foi algo superficial, tendo sido identificados erros que inviabilizariam a utilização de alguma informação num projeto real;
- No desenvolvimento da prova de conceito foi usada uma abordagem simplificada da metodologia “Business Dimensional Lifecycle”, apresentada por Ralph Kimball em [KIM08], em que algumas fases e/ou tarefas foram omitidas ou abordadas apenas superficialmente.

4.2 Recomendações

A metodologia de Kimball prevê um conjunto de passos adicionais relacionados com a manutenção e a evolução da solução, na sua fase 3 (ver figura 1), que não são aplicáveis a este caso concreto.

No entanto, é possível referir que, no caso de evolução para um projeto concreto, a solução terá obrigatoriamente que ser executada em *hardware* adequado, os processos de ETL deverão ser redesenhados com a inclusão de pontos de controlo e de recuperação em caso de falha, a criação de uma área de *staging* com uma tabela de equivalências de chaves substitutas e naturais, a inclusão de novas fontes de dados (os *logs* JSON) e o carregamento das tabelas de agregação, entre outros aspetos.

O próprio desenho do *data warehouse* e dos cubos precisará de ser revisto, no sentido de incluir mais informação capaz de suportar melhor os processos de tomada de decisão.

Relativamente às aplicações analíticas, será necessário construir relatórios estáticos com recurso ao Pentaho Report Designer, novas análises com o Pentaho Design Studio e soluções de *data mining*, utilizando o Pentaho Weka, que poderão trazer um grande valor acrescentado à solução.

Será também necessária a implementação de mecanismos de segurança e a introdução de níveis de acesso para os utilizadores do sistema, no sentido de prevenir acessos não autorizados.

Como resultado da análise à qualidade dos dados do sistema fonte, recomenda-se ainda a melhoria do procedimento de registo dos utilizadores da plataforma Anúncios SAPO,

nomeadamente, a inclusão de validações para garantir que as moradas e nomes de utilizador introduzidos são válidos, com particular importância para os códigos postais. Recomenda-se também a alteração da base de dados operacional no sentido de distinguir as campanhas e utilizadores internos do grupo PT dos restantes utilizadores.

Bibliografia

HEN08: Henry, Andrew, Economic Crisis Will Boost Open Source Adoption, IT World Canada, Outubro 2008, *http://www.itworldcanada.com/news/economic-crisis-will-boost-open-source-adoption-/04156*

SMI08: Smith, J.T., Recession Worries? Open Source Software Is a Great Way to Cut Costs, CIO, Outubro 2008, *http://www.cio.com/article/454766/Recession_Worries_Open_Source_Software_Is_a_Great_Way_to_Cut_Costs*

IMH08: Imhof, Claudia Phd., Open Sesame: Why Open Source BI and Data Warehousing Solutions are Gaining in Acceptance, Intelligent Solutions, Outubro, 2008

NIC08: Nickolett, Chip, DeliverIng Business Intelligence with Open Source Software, Ingres Corporation, Redwood City, 2008

PEN07: A New Business Model to Drive Business Intelligence Acceptance and Adoption, Pentaho, 2007

VEN08: Making Open Source BI Viable for the Enterprise, Ventana Research, Pleasanton, 2008

BIT08: Bitterer, Andreas, Who's Who in Open-Source Business Intelligence, Gartner Group, Stamford, Abril, 2008

SAU09: Saunders, Mark et al, Research Methods for Business Students, Financial Times / Prentice Hall, Harlow, Abril, 2009

KIM08: Kimball, Ralph et al, The Data Warehouse Lifecycle Toolkit, Wiley, Indianapolis, Janeiro, 2008

NEG04: Negash, Solomon, Business Intelligence, Association for Information Systems, Atlanta, 2004

ROS04: Rosen, Lawrence, Open Source Licensing, Prentice Hall, Indianapolis, Agosto, 2004

FSF04: Free Software, Free Software Foundation, Outubro, 2004

STL04: St. Laurent, Andrew M., Understanding Open Source and Free Software Licensing, O'Reily, Sebastopol, Agosto, 2004

KAV04: Kavanaugh, Paul, Open Source Software, Elsevier Digital Press, Burlington, 2004

MEE08: Meeker, Heather J., The Open Source Alternative, Wiley, Hoboken, Fevereiro, 2008

GUL05: Gautam Guliani, Dan Woods, Open Source for the Enterprise, O'Reily, Sebastopol, Julho, 2005

SIN10: Relatório e Contas 09, SINFIC, Alfragide, 2010

SIF10: Ofertas de Emprego, SINFIC, 2010, *http://www.sinfic.pt/SinficWeb/displayconteudo.do2?numero=23836*

INM07: W. H. Inmon, The Evolution of Integration, Inmon Consulting Services, 2007

MAD09: Madsen, Mark, Open Source Solutions: Managing, Analyzing and Delivering Business Information, BeyerNETWORK, Boulder, 2009

VES07: Vesset, Dan; McDonough, Brian, Competitive Analysis - Worldwide Business Intelligence Tools 2006 Vendor Shares, IDC, Framingham, 2007

GHO06: Ghosh, Rishab Aiyer, Study on the Economic impact of open source software on innovation and the competitiveness of the Information and Communication Technologies (ICT) sector in the EU, UNU-MERIT, Holanda, Setembro, 2006

CEN10: Cenatic, Report on the International Status of Open Source Software 2010, National Open Source Software Observatory (ONSFA), Almendralejo (Badajoz), 2010

DRI01: Driver, Mark; Weiss, George, The Future of Linux and Open Source Softwares, Gartner Group, Junho, 2001
LEY00: Leyden, Jon, Gartner debunks open source myths, Incisive Media, Londres, Maio, 2000
MCG08: Mcgrattan, Emma, 6 Myths About open source, Ingres Corporation, 2008
FOR07: Open Source Software's Expanding Role in the Enterprise, Forrester Research, 2007
WIK09: Linux Caixa Mágica, Wikipedia, 2009, *pt.wikipedia.org/wiki/Linux_Caixa_Mágica*
LEH09: Lehey, Greg, Explaining BSD, FreeBSD, 2009, *www.freebsd.org/doc/en/articles/explaining-bsd*
LEE02: Lee, James; Ware, Brent, Open Source Web Development with LAMP, Addison-Wesley, Boston, Dezembro, 2002
OPT07: Open Source Catalogue 2007, Optaros, 2007
BUR04: Burgelman, Robert A., MySQL Open Source Database in 2004, Stanford School of Business, 2004
MYS09: CIO Guide: The Strategic Value of MySQL, MySQ, 2009
MYS07: Enterprise Data Warehousing with MySQL, MySQL, 2007
MYS08: Disruptive Data Warehousing for Communications Service Providers, MySQL, Maio, 2008
POS08: PostgreSQL 8.3.6 Documentation, The PostgreSQL Global Development Group, 2008
ING08: Ingres Products, Ingres, 2008
GAR05: Garry, Charlie, The End of Database Licensing?, IT World Canada, 2005, *http://www.itworldcanada.com/news/the-end-of-database-licensing/112062*
INF08: Customer Relationship Management: SugarCRM, InfoTech Research Group, Julho, 2007
OPE09: Openbravo ERP: Why?, Openbravo, 2009
XTU09: Solutions, XTuple, 2009
COM08: Compiere Professional Edition Capabilities Overview, Compiere, Redwood Shores, 2008
COM09: COMPIERE PLATFORM, Compiere, Redwood Shores, 2009
HAK08: Hakala, David, The Top 10 Open-Source CRM Solutions, Focus, Maio, 2009, *http://www.focus.com/briefs/crm/top-10-open-source-crm-solutions/*
WIS10: Wise, Lyndsay, The Influence Of Open Source On Mainstream BI, Dashboard Insight, Toronto, Setembro, 2010, *http://www.dashboardinsight.com/articles/business-verticals/the-influence-of-open-source-on-mainstream-bi.aspx*
HOW09: Howard, Philip, Talend Open Studio v3, Bloor Research, Londres, Janeiro, 2009,
45108: Open Source is not a Business Model – Commercial Acceptance Of Open Source, The 451 Group, Nova Iorque, 2008
PAN09: Panettieri, Joe, The VAR Guy's Open Source 50, The VAR Guy and Nine Lives Media, Nova Iorque, 2009
ORA10: MySQL 5.1 Reference Manual, Oracle, Outubro, 2010
RIC09: Richardson, James et al, Magic Quadrant for Business Intelligence Platforms, Gartner, Stamford, Janeiro, 2009,
GOL09: Golfarelli, Matteo, Open Source BI Platforms: a Functional and Architectural Comparison, Universidade de Bolonha, Bolonha, 2009
KIM04: Kimball, Ralph; Caserta, Joe, Tihe Data Warehouse ETL Toolkit, Wiley, Indianapolis,

Setembro, 2004
CAS10: Casters, Matt, Dimension Lookup-Update, Pentaho, 2010, *http://wiki.pentaho.com/display/EAI/Dimension+Lookup-Update*
PEN09: Introducing the Pentaho BI Suite 3.5 Community Edition, Pentaho, Orlando, 2009
WIK10: MultiDimensional eXpressions, Wikipedia, 2009, *http://en.wikipedia.org/wiki/Multidimensional_Expressions*
WOO07: Wood, Sherman, Mondrian Schema Workbench, Pentaho, 2007, *http://mondrian.pentaho.com/documentation/workbench.php*
BAK10: Baker, Marina, Publishing an Analysis Schema Using Schema Workbench, Pentaho, 2010, *http://wiki.pentaho.com/display/analysis/Publishing+an+Analysis+Schema+Using+Schema+Workbench*
EMB08: Emberson, Richard, Aggregate Tables, Pentaho, 2008, *http://mondrian.pentaho.com/documentation/aggregate_tables.php*
WOH07: Wood, Sherman; Hyde, Julian, Optimizing Mondrian Performance, Pentaho, 2007, *http://mondrian.pentaho.com/documentation/performance.php*
PEN08: Pentaho, Pentaho Aggregation Designer 1.0 User Guide, Pentaho, Orlando, 2008

Anexo A – *Open Source* Definition

Adaptado de [COA06].

Introduction

Open source doesn't just mean access to the source code. The distribution terms of open-source software must comply with the following criteria:

Free Redistribution

The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

Source Code

The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

Derived Works

The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

Integrity of The Author's Source Code

The license may restrict source-code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of

software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

No Discrimination Against Persons or Groups

The license must not discriminate against any person or group of persons.

No Discrimination Against Fields of Endeavor

The license must not restrict anyone from making use of the program in a specific field of endeavor.

For example, it may not restrict the program from being used in a business, or from being used for genetic research.

Distribution of License

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

License Must Not Be Specific to a Product

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

License Must Not Restrict Other software

The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.

License Must Be Technology-Neutral

No provision of the license may be predicated on any individual technology or style of interface.

Anexo B – Licenças *open source* Mais Comuns

Licença	Hereditariedade	Autor/Steward	Projetos	Observações
Academic Free License	Não	Lawrence Rosen	N/A	
Adaptive Public License	Ver Observações	N/A	N/A	Permite ao autor a adaptação da licença a diferentes requisitos.
Affero GPL (1.0)	Sim	Affero, Inc./FSF	Projecto Affero	Geralmente considerada a mais "viral" das licenças.
Apache software License (1.1)	Não	Fundação Apache	Apache (revogado pela versão 2.0)	A Apache 1.0 quase já não é usada. A versão 1.1 retirou a cláusula relativa a publicidade.
Apache License (2.0)	Não	Fundação Apache	Ant, Tomcat, Xerces	Licença permissiva, mas muito mais detalhada que a MIT ou BSD, ou a Apache 1.0 ou 1.1.
Apple Public Source Licence	N/A	Apple	Darwin	Licença de estilo corporativo, com muitas cláusulas específicas à Apple.
Artistic License	Não	Larry Wall	PERL, CPAN	Não hereditária, mas mais restritiva que a maioria das licenças permissivas. Muitos dos projetos sob esta licença são dualmente licenciados sob a GPL.
Attribution Assurance Licenses	Não	Edwin A. Suominen	N/A	Baseada na BSD, com adição de algumas cláusulas.
New BSD License	Não	FreeBSD	N/A	Licença modelo – muitas variantes em utilização.
Computer Associates Trusted *open source* License 1.1	Sim	Computer Associates	Ingres	Licença de estilo corporativo, derivada da CPL.
Common Development and Distribution License	Sim	Sun	OpenSolaris NetBeans	Sucessora da Sun Public License. Adição recente à lista da OSI. Baseada na MPL.
Common Public License 1.0	Sim	IBM	N/A	Sucessora da IBM Public License. Ver também a Eclipse license.
Eclipse Public License	Sim	Fundação Eclipse	Projecto Eclipse	Variante da CPL
Educational Community License	Não	N/A	N/A	N/A
Frameworx Open License	Sim	Frameworx Company	N/A	N/A
GNU General Public License (GPL)	Sim	FSF	Linux, GNU	Licença mais comum.
''GPL plus exception''	Sim	FSF	GNU Classpath	Variante da GPL que permite a "linkagem" com código proprietário.
GNU Library ou Lesser General Public License (LGPL)	Sim	FSF	N/A	Variante da GPL que permite a "linkagem" com código proprietário.
Historical Permission Notice and Disclaimer	Não	N/A	Kernel BSD	Variante da BSD. Voluntariamente repudiada pelo seu autor.
IBM Public License	Sim	IBM	N/A	Genericamente semelhante à GPL, mas escrita de forma mais convencional. Em grande medida, tornada obsoleta pela CPL e Eclipse.
Intel *open source* License	Não	Intel	N/A	BSD com cláusula adicional. A Intel desaprovou o seu uso.
MIT license	Não	N/A	Sistema X Windows PuTTY	Licença permissiva mais popular a seguir à BSD. Também conhecida como licença X ou X11.
Mozilla Public License 1.0 (MPL)	Sim	Fundação Mozilla	N/A	Licença hereditária baseada na Netscape Public License.
MySQL (GPL + FLOSS Exception)	Sim	MySQL	MySQL	Semelhante à GPL, mas permite a linkagem a código *open source*.
Netscape Public License	N/A	Netscape	Browser Netscape	Base original para a Mozilla Public License, não aprovada como open source. Contém direitos específicos para a Netscape.
Open software License	Sim	Lawrence Rosen	N/A	Requisitos de copyleft são acionados através da utilização online. Contém garantias limitadas respeitantes à originalidade do código.
OpenSSL	Não	N/A	OpenSSL	N/A
PHP License	Não	PHP Group	Linguagem de programação PHP	N/A
Python software Foundation License	Não	Python Foundation	Linguagem de programação Python	Licença permissiva desenhada de forma mais convencional que a BSD ou a MIT. Ver http://www.python.org/ download/releases/2.4.2/ license.

Licença	Hereditariedade	Autor/Steward	Projetos	Observações
Qt Public License (QPL)	Sim	Trolltech	QT Toolkit	Variante da GPL.
Reciprocal Public License	Sim	Technical Pursuit	N/A	Maiores restrições que na GPL.
Ricoh Source Code Public License	Sim	Ricoh	N/A	Baseada na Netscape Public License.
Sleepycat License	Sim	Sleepycat	Berkeley DB	Similar na forma e no estilo com a BSD, mas contendo maiores restrições à distribuição. Os produtos da Sleepycat têm uma licença dual em termos comerciais.

Tabela 17: Licenças open source mais comuns

Anexo C – Requisitos de Negócio Definidos pela PT Comunicações

Pesquisa

Métricas de Utilização Pesquisa	Nº de utilizadores únicos
	Origem das Pesquisas efectuadas
	Tipo de Pesquisa (barra)
	Pesquisa web por tipo (Portugal, Português, toda a web)
	Host referral
	Top de Queries Pesquisadas
	Número de pesquisas avançadas
	Utilização do [+] (cliques em abrir nova página, pesquisar em, enviar página a amigo, guardar em pdf, adicionar ao Sapo Tags)
	Utilização do preview
	Utilização do dicionário
	% activação de módulos (por tipo de módulos)
	% cliques nos módulos
	Número de cliques nos primeiros três resultados
	Número de cliques na primeira página
	Número de cliques em seguinte/anterior
	Número de pesquisas únicas (pesquisas por termos únicos - ou seja, número de termos únicos pesquisados)
	Número de utilizadores com cada pesquisa default (Portugal, Português, toda a web)
	Nº termos por pesquisa (distribuição)
	Tempo por pesquisa (distribuição)
	Total de Pageviews
	Total de Pesquisas
	Pesquisas por visitante único
	Tempo médio de sessão (genérico, não tem a ver só com a pesquisa, mas suponho que já o estejam a medir)
	Número de pesquisas
	Nº anúncios por bloco de pesquisa
	CTR no top3 resultados por diferente opção pesquisa (pesquisa mundial, nacional, notícias, ...) - – identificar isto com a distribuição dos top e bottom searchers (ex: o top 10% de maiores pesquisadores, apresentam uma fraca utilização do seguinte/anterior – isto só deve dar para clientes adsl, suponho?)
	Palavras com maior CTR no top3 resultados
	Tempo médio por pesquisa
	Nº páginas por pesquisa (utilização do seguinte/anterior) – identificar isto com a distribuição dos top e bottom searchers (ex: o top 10% de maiores pesquisadores, apresentam uma fraca utilização do seguinte/anterior – isto só deve dar para clientes adsl, suponho?)
	Nº médio de palavras por pesquisa
	Split de pesquisas com um termo vs. >1 termo
	Correlação entre número de pesquisas e tempo médio por pesquisa
	Visitantes Únicos
Dimensões de Análise	Geografia
	Provider (SAPO, Telepac, Free, Outros)
	ADSL ou Não ADSL
	Período (Dia/Semana/Mês/X Dias/Acumulado)
	Dia da semana (2ª/3ª/4ª/5ª/6ª/Sábado/Domingo)
	Faixa horária

Tabela 18: Requisitos Preliminares de Negócio - Pesquisas

Anúncios

Métricas de Utilização Anúncios	Distribuição das palavras com número de Anúncios SAPO (ideia de podermos cruzar com as palavras que apresentam um ctr maior e verificar se há uma correlação entre número de anúncios e ctr)
	ARPU (usando um CPC médio, conseguimos identificar quais os segmentos de utilizador que geram maior revenue)
	Valor médio dos carregamentos
	Tempo médio de carregamento
	CTR nos Anúncios SAPO por diferente opção de pesquisa (e por tipo de anúncio – no topo ou laterais) - – identificar isto com a distribuição dos top e bottom searchers (ex: o top 10% de maiores pesquisadores, apresentam uma fraca utilização do seguinte/anterior – isto só deve dar para clientes adsl, suponho?)
	Palavras com maior CTR nos Anúncios SAPO
	Cobertura Anúncios Sapo
	Cliques em anúncios SAPO (total/top/bottom/lateral)
	Nº Anunciantes
	Nº Anunciantes Activos a 30 dias
Dimensões de Análise	Geografia
	Provider (SAPO, Telepac, Free, Outros)
	ADSL ou Não ADSL
	Período (Dia/Semana/Mês/X Dias/Acumulado)
	Dia da semana (2ª/3ª/4ª/5ª/6ª/Sábado/Domingo)
	Sexo
	Idade
	Faixa horária

Tabela 19: Requisitos Preliminares de Negócio - Anúncios

Anexo D - Plano de Projeto

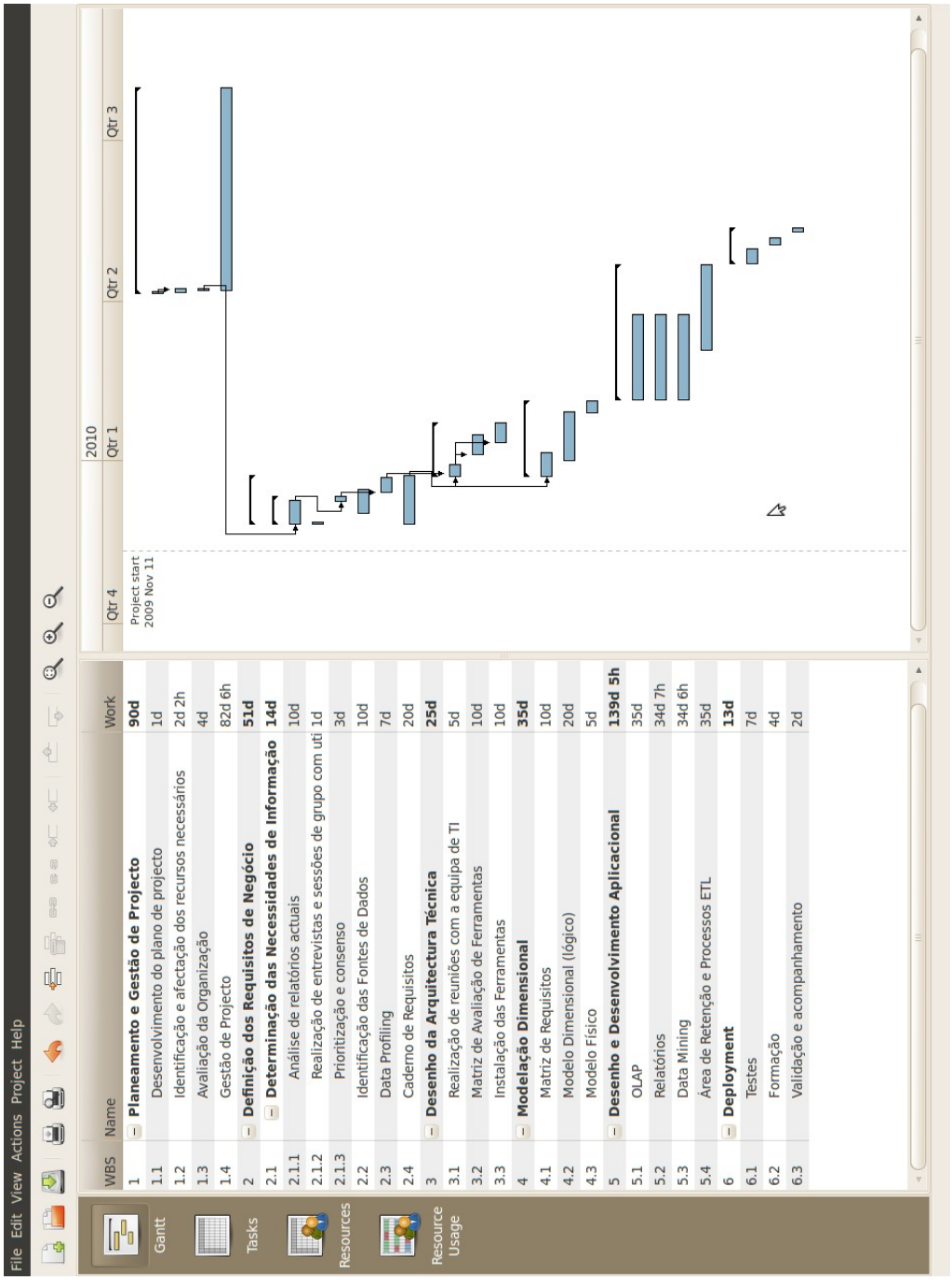


Figura 59: Plano de Projecto da Prova de Conceito

Anexo E – Informação do *Backoffice* Anúncios SAPO

- Utilizadores ativos por intervalo de datas;
- Utilizadores por nível de saldo;
- Histórico de carregamentos por utilizadores e afiliados;
- Repartição de receitas por parceiro;
- Estatísticas de cliques, incluindo cliques por canal;
- Impressões por palavra;
- Métricas:
  - Número de cliques;
  - CPC;
  - Valor;
- Estatísticas de faturação (pedidos de faturas);
- Estatísticas para anunciantes por período de tempo, incluindo os seguintes dados:
  - Data;
  - Campanha;
  - Grupo de anúncios;
  - Palavra;
  - CPC;
  - Impressões;
  - Cliques;
  - CTR;
  - Custo;

    - Posição Média;
    - Conversões;
- Relatórios para afiliados;
    - Blocos de anúncios;
    - Evolução de impressões, cliques, receitas, CTR, CPC, RPM, ECPM;
    - Estatísticas sobre pesquisas, sites e blocos de anúncios;
- Estatísticas sobre impressões do hi5;
- Los de acesso (cliques e impressões).

Anexo F – Matriz do Data Warehouse

	Impressões	Cliques	Impressões	CPC Médio	RPM	Custo Total	CTR	Conversões	Posição Média	Carregamentos	Valor
Data	X	X	X	X	X	X	X	X	X	X	X
Data	X	X	X	X	X	X	X	X	X	X	X
Dia	X	X	X	X	X	X	X	X	X	X	X
DiaSemana	X	X	X	X	X	X	X	X	X	X	X
Feriado	X	X	X	X	X	X	X	X	X	X	X
Evento	X	X	X	X	X	X	X	X	X	X	X
Semana	X	X	X	X	X	X	X	X	X	X	X
Mes	X	X	X	X	X	X	X	X	X	X	X
Trimestre	X	X	X	X	X	X	X	X	X	X	X
Ano	X	X	X	X	X	X	X	X	X	X	X
Geografia	X	X	X	X	X	X	X	X	X	X	X
codLocalidade	X	X	X	X	X	X	X	X	X	X	X
Pais	X	X	X	X	X	X	X	X	X	X	X
Distrito	X	X	X	X	X	X	X	X	X	X	X
Concelho	X	X	X	X	X	X	X	X	X	X	X
Localidade	X	X	X	X	X	X	X	X	X	X	X
CodPostal	X	X	X	X	X	X	X	X	X	X	X
LocalCodPostal	X	X	X	X	X	X	X	X	X	X	X
Palavras	X	X	X	X	X	X	X	X	X		
word	X	X	X	X	X	X	X	X	X		
Tamanho	X	X	X	X	X	X	X	X	X		
Palavra	X	X	X	X	X	X	X	X	X		
Anunciantes	X	X	X	X	X	X	X	X	X	X	X
user_id	X	X	X	X	X	X	X	X	X	X	X
netbi	X	X	X	X	X	X	X	X	X	X	X
Nome	X	X	X	X	X	X	X	X	X	X	X
Morada1	X	X	X	X	X	X	X	X	X	X	X
Morada2	X	X	X	X	X	X	X	X	X	X	X
CodPostal	X	X	X	X	X	X	X	X	X	X	X
email	X	X	X	X	X	X	X	X	X	X	X
Rede	X	X	X	X	X	X	X	X	X	X	X
Geografia	X	X	X	X	X	X	X	X	X	X	X
codLocalidade	X	X	X	X	X	X	X	X	X	X	X
Pais	X	X	X	X	X	X	X	X	X	X	X
Distrito	X	X	X	X	X	X	X	X	X	X	X
Concelho	X	X	X	X	X	X	X	X	X	X	X
Localidade	X	X	X	X	X	X	X	X	X	X	X
CodPostal	X	X	X	X	X	X	X	X	X	X	X
LocalCodPostal	X	X	X	X	X	X	X	X	X	X	X

	Impressões	*Cliques*	*Impressões*	*CPC Médio*	*RPM*	*Custo Total*	*CTR*	*Conversões*	*Posição Média*	***Carregamentos***	*Valor*
Anuncios	X	X	X	X	X	X	X	X	X		
ads_id	X	X	X	X	X	X	X	X	X		
cam_id	X	X	X	X	X	X	X	X	X		
CampanhaDescricao	X	X	X	X	X	X	X	X	X		
CampanhaDataInicio	X	X	X	X	X	X	X	X	X		
CampanhaDataFim	X	X	X	X	X	X	X	X	X		
CampanhaDataCriacao	X	X	X	X	X	X	X	X	X		
CampanhaApagada	X	X	X	X	X	X	X	X	X		
CampanhaSegmentacao	X	X	X	X	X	X	X	X	X		
CPCMaximo	X	X	X	X	X	X	X	X	X		
adg_id	X	X	X	X	X	X	X	X	X		
GrupoDescricao	X	X	X	X	X	X	X	X	X		
GrupoApagado	X	X	X	X	X	X	X	X	X		
GrupoDataCriacao	X	X	X	X	X	X	X	X	X		
Titulo	X	X	X	X	X	X	X	X	X		
Url	X	X	X	X	X	X	X	X	X		
AnuncioApagado	X	X	X	X	X	X	X	X	X		
Anunciantes	X	X	X	X	X	X	X	X	X		
user_id	X	X	X	X	X	X	X	X	X		
netbi	X	X	X	X	X	X	X	X	X		
Nome	X	X	X	X	X	X	X	X	X		
Morada1	X	X	X	X	X	X	X	X	X		
Morada2	X	X	X	X	X	X	X	X	X		
CodPostal	X	X	X	X	X	X	X	X	X		
email	X	X	X	X	X	X	X	X	X		
Rede	X	X	X	X	X	X	X	X	X		
Geografia	X	X	X	X	X	X	X	X	X		
codLocalidade	X	X	X	X	X	X	X	X	X		
Pais	X	X	X	X	X	X	X	X	X		
Distrito	X	X	X	X	X	X	X	X	X		
Concelho	X	X	X	X	X	X	X	X	X		
Localidade	X	X	X	X	X	X	X	X	X		
CodPostal	X	X	X	X	X	X	X	X	X		
LocalCodPostal	X	X	X	X	X	X	X	X	X		
Tipos de Pagamento										X	X
payment_type										X	X
Tipo										X	X

Tabela 20: Matriz do Data Warehouse

Anexo G – Mapa Lógico do Data Warehouse

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimData	IdTempo	INTEGER	Dimensão	I					Chave Substituta
DimData	Data	DATE	Dimensão	I	Data (Excel)	Tempo	Data	DATE	Chave Natural
DimData	Dia	TINYINT	Dimensão	I	Data (Excel)	Tempo	Dia	INTEGER	SELECT Dia FROM Tempo
DimData	DiaSemana	VARCHAR(10)	Dimensão	I	Data (Excel)	Tempo	DiaSemana	INTEGER	1. SELECT DiaSemana From Tempo; 2. Converter DiaSemana para VARCHAR: 1 = 'Domingo'; 2 = 'Segunda'; 3 = 'Terça'; 4 = 'Quarta'; 5 = 'Quinta'; 6 = 'Sexta'; 7 = 'Sábado'.
DimData	Feriado	VARCHAR(50)	Dimensão	I	Data (Excel)	Tempo	Feriado	STRING	SELECT Feriado FROM Tempo
DimData	Evento	VARCHAR(255)	Dimensão	I	Data (Excel)	Tempo	Evento	STRING	SELECT Evento FROM Tempo
DimData	Semana	TINYINT	Dimensão	I	Data (Excel)	Tempo	Semana	INTEGER	SELECT Semana FROM Tempo

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimData	Mes	VARCHAR(20)	Dimensão	I	Data (Excel)	Tempo	Mes	INTEGER	1. SELECT Mes From Tempo; 2. Converter Mes para VARCHAR: 1 = 'Janeiro'; 2 = 'Fevereiro'; 3 = 'Março'; 4 = 'Abril'; 5 = 'Maio'; 6 = 'Junho'; 7 = 'Julho'; 8 = 'Agosto'; 9 = 'Setembro'; 10 = 'Outubro'; 11 = 'Novembro'; 12 = 'Dezembro'.
DimData	Trimestre	TINYINT	Dimensão	I	Data (Excel)	Tempo	Trimestre	INTEGER	SELECT Trimestre FROM Tempo
DimData	Ano	TINYINT	Dimensão	I	Data (Excel)	Tempo	Ano	INTEGER	SELECT Ano FROM Tempo
DimPalavras	IdPalavra	INTEGER	Dimensão	I					Chave Substituta
DimPalavras	word_id	VARCHAR(255)	Dimensão	I	Adwords2010	wrd_words	wrd_word	VARCHAR(255)	Chave Natural
DimPalavras	word	VARCHAR(255)	Dimensão	I	Adwords2010	wrd_words	word	VARCHAR(255)	SELECT word FROM wrd_words
DimPalavras	tamanho	INTEGER	Dimensão	I	Adwords2010	wrd_words	word_size	INT(4)	SELECT word_size FROM wrd_words
DimGeografia	IdGeografia	INTEGER	Dimensão	I					Chave Substituta
DimGeografia	CodLocalidade	VARCHAR(20)	Dimensão	I	CTT	todos_cp	DD, CC, LLLL	STRING	Chave Natural SELECT CONCAT(DD, CC, LLLL) FROM todos_cp
DimGeografia	Pais	VARCHAR(255)	Dimensão	I					*Hardcoded ('Portugal')*
DimGeografia	Distrito	VARCHAR(255)	Dimensão	I	CTT	distritos	DescDistrito	STRING	SELECT distritos.DescDistrito FROM distritos, todos_cp WHERE distritos.CodDistrito LIKE todos_cp.DD

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimGeografia	Concelho	VARCHAR(255)	Dimensão	I	CTT	concelhos	DescConcelho	STRING	SELECT concelhos.DescConcelho FROM concelhos, todos_cp WHERE concelhos.CodConcelho LIKE todos_cp.CC
DimGeografia	Localidade	VARCHAR(255)	Dimensão	I	CTT	todos_cp	LOCALIDADE	STRING	SELECT LOCALIDADE FROM todos_cp
DimGeografia	CodPostal	VARCHAR(10)	Dimensão	I	CTT	todos_cp	CP4, CP3	STRING	SELECT CONCAT(CP4, '-', CP3) FROM todos_cp
DimGeografia	LocalCodPostal	VARCHAR(255)	Dimensão	I	CTT	todos_cp	CPALF	STRING	SELECT CPALF FROM todos_cp
DimAnunciantes	IdAnunciante	INTEGER	Dimensão	I					Chave Substituta
DimAnunciantes	user_id	INTEGER	Dimensão	I	Adwords2010	usr_users	usr_id	INTEGER	Chave Natural
DimAnunciantes	netbi	VARCHAR(50)	Dimensão	I	Adwords2010	usr_users	usr_netbi	VARCHAR(50)	SELECT usr_netbi FROM usr_users
DimAnunciantes	nome	VARCHAR(255)	Dimensão	I	Adwords2010	usr_users	usr_name	VARCHAR(255)	SELECT usr_name FROM usr_users
DimAnunciantes	morada1	VARCHAR(255)	Dimensão	II	Adwords2010	usr_users	usr_addr1	VARCHAR(255)	SELECT usr_addr1 FROM usr_users
DimAnunciantes	morada2	VARCHAR(255)	Dimensão	II	Adwords2010	usr_users	usr_addr2	VARCHAR(255)	SELECT usr_addr2 FROM usr_users
DimAnunciantes	CodPostal	VARCHAR(8)	Dimensão	II	Adwords2010	usr_users	usr_zip	VARCHAR(255)	SELECT LEFT(TRIM(usr_zip), 8) FROM usr_users
DimAnunciantes	Localidade	VARCHAR(32)	Dimensão	II	Adwords2010	usr_users	usr_city	VARCHAR(100)	SELECT usr_city FROM usr_users
DimAnunciantes	email	VARCHAR(255)	Dimensão	I	Adwords2010	usr_users	usr_email	VARCHAR(100)	SELECT usr_email FROM usr_users
DimAnunciantes	Tipo	VARCHAR(20)	Dimensão	I	Adwords2010	usr_users	is_corporate	TINYINT(1)	1. SELECT is_corporate FROM usr_users 2. Converter is_corporate para VARCHAR 0 = 'Particular' 1 = 'Empresarial'

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimAnunciantes	Rede	VARCHAR(255)	Dimensão	II	Adwords2010	ads_portals	name	VARCHAR(255)	SELECT name FROM ads_portals, usr_users WHERE ads_portals.portal_id = usr_users.portal
DimAnunciantes	IdGeografia	INTEGER	Dimensão	II	DW	DimGeografia	IdGeografia	INTEGER	SELECT IdGeografia FROM ALVO.DimGeografia, Adwords2010.usr_users WHERE DimGeografia.CodPostal LIKE LEFT(TRIM(usr_users.usr_zip), 8)
DimAnuncios	IdAnuncio	INTEGER	Dimensão	I					Chave Substituta
DimAnuncios	ads_id	INTEGER	Dimensão	I	Adwords2010	ads_ads	ads_id	INTEGER	Chave Natural
DimAnuncios	adg_id	INTEGER	Dimensão	I	Adwords2010	adg_adgroups	adg_id	INTEGER	SELECT adg.adgroups.adg_id FROM adg_adgroups, ads_ads WHERE ads_ads.adg_id = ads_ads.adg_id
DimAnuncios	cam_id	INTEGER	Dimensão	I	Adwords2010	cam_campaigns	cam_id	INTEGER	SELECT cam_campaigns.cam_id FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	CampanhaDescricao	VARCHAR(120)	Dimensão	I	Adwords2010	cam_campaigns	cam_description	VARCHAR(120)	SELECT cam_campaigns.cam_description FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimAnuncios	CampanhaDataInicio	DATE	Dimensão	I	Adwords2010	cam_campaigns	cam_start_date	DATETIME	SELECT cam_campaigns.cam_start_date FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	CampanhaDataFim	DATE	Dimensão	I	Adwords2010	cam_campaigns	cam_end_date	DATETIME	SELECT cam_campaigns.cam_end_date FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	CampanhaDataCriacao	DATE	Dimensão	I	Adwords2010	cam_campaigns	cam_created	DATETIME	SELECT cam_campaigns.cam_created FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	CampanhaApagada	TINYINT	Dimensão	I	Adwords2010	cam_campaigns	cam_deleted	TINYINT(1)	SELECT cam_campaigns.cam_deleted FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
DimAnuncios	CampanhaSegmentacao	VARCHAR(10)	Dimensão	I	Adwords2010	cam_campaigns	site_targeted	TINYINT(1)	1. SELECT cam_campaigns.site_targeted FROM cam_campaigns, adg_adgroups, ads_ads WHERE cam_campaigns.cam_id = adg_groups.cam_id AND adg_adgroups.adg_id = ads_ads.adg_id 2. CONVERTER site_targeted para varchar: 0 = 'Palavras'; 1 = 'Sites'.
DimAnuncios	CPCMaximo	FLOAT	Dimensão	I	Adwords2010	adg_adgroups	adg_max_cpc	FLOAT(9,2)	SELECT adg_adgroups.adg_max_cpc WHERE adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	GrupoDescricao	VARCHAR(120)	Dimensão	I	Adwords2010	adg_adgroups	adg_description	VARCHAR(120)	SELECT adg_adgroups.adg_description WHERE adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	GrupoApagado	TINYINT	Dimensão	I	Adwords2010	adg_adgroups	adg_deleted	TINYINT(1)	SELECT adg_adgroups.adg_description WHERE adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	GrupoDataCriacao	DATE	Dimensão	I	Adwords2010	adg_adgroups	adg_created	DATETIME	SELECT adg_adgroups.adg_created WHERE adg_adgroups.adg_id = ads_ads.adg_id
DimAnuncios	Titulo	VARCHAR(255)	Dimensão	I	Adwords2010	ads_ads	ads_title	VARCHAR(255)	SELECT ads_title FROM ads_ads
DimAnuncios	Tipo	VARCHAR(6)	Dimensão	I	Adwords2010	ads_ads	ads_type	TINYINT(4)	1. SELECT ads_type FROM ads_ads 2. CONVERTER ads_type para VARCHAR: 0 = 'Texto';

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
									2 = 'Imagem'.
DimAnuncios	Url	VARCHAR(255)	Dimensão	I	Adwords2010	ads_ads	ads_url_display	VARCHAR(255)	SELECT ads_url_display FROM ads_ads
DimAnuncios	AnuncioDataCriadao	DATE	Dimensão	I	Adwords2010	ads_ads	ads_created	DATETIME	SELECT ads_created FROM ads_ads
DimAnuncios	AnuncioApagado	TINYINT	Dimensão	I	Adwords2010	ads_ads	ads_deleted	TINYINT(1)	SELECT ads_deleted FROM ads_ads
DimAnuncios	IdAnunciante	INTEGER	Dimensão	I	DW	DimAnunciantes	IdAnunciante	INTEGER	SELECT IdAnunciante FROM DimAnunciantes, ads_ads WHERE DimAnunciantes.user_id = ads_ads.usr_id
DimTipo	IdTipo	INTEGER	Dimensão	I					Chave Substituta
DimTipo	payment_type	INTEGER	Dimensão	I	DATA GRID	DATA GRID	CodTipo	INTEGER	Chave Natural
DimTipo	Tipo	VARCHAR(255)	Dimensão	I	DATA GRID	DATA GRID	Tipo	STRING	CodTipo = 0 ► Tipo = 'Wallet'; CodTipo = 1 ► Tipo = 'PayPal'; CodTipo = 2 ► Tipo = 'Promocode'; CodTipo = 3 ► Tipo = 'Carregamento Manual'.
FactCarregamentos	IdAnunciante	INTEGER	Facto	NA	DW	DimAnunciantes	IdAnunciante	INTEGER	SELECT IdAnunciante FROM DimAnunciantes, usr_payments WHERE DimAnunciantes.user_id = usr_payments.usr_id
FactCarregamentos	IdGeografia	INTEGER	Facto	NA	DW	DimAnunciantes	IdGeografia	INTEGER	SELECT IdGeografia FROM DimAnunciantes, usr_payments WHERE DimAnunciantes.user_id = usr_payments.usr_id
FactCarregamentos	IdData	INTEGER	Facto	NA	DW	DimData	IdData	INTEGER	SELECT IdData FROM DimData, usr_payments

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
									WHERE DimData.Data = usr_payments.pay_init_date
FactCarregamentos	IdTipo	INTEGER	Facto	NA	DW	DimTipo	IdTipo	INTEGER	SELECT IdTipo FROM DimTipo, usr_payments WHERE DimTipo.IdTipo = usr_payments.pay_type
FactCarregamentos	Valor	FLOAT	Facto	NA	Adwords2010	usr_payments	pay_value	FLOAT	SELECT SUM(pay_value) FROM usr_payments GROUP BY usr_id, pay_init_date, pay_type
FactImpressoes	IdAnunciante	INTEGER	Facto	NA	DW	DimAnunciantes	IdAnunciante	INTEGER	SELECT IdAnunciante FROM DimAnunciantes, usr_impressions_report WHERE DimAnunciantes.user_id = usr_impressions_report.usr_id
FactImpressoes	IdGeografia	INTEGER	Facto	NA	DW	DimAnunciantes	IdGeografia	INTEGER	SELECT IdGeografia FROM DimAnunciantes, usr_impressions_report WHERE DimAnunciantes.user_id = usr_impressions_report.usr_id
FactImpressoes	IdData	INTEGER	Facto	NA	DW	DimData	IdData	INTEGER	SELECT IdData FROM DimData, usr_impressions_report WHERE DimData.Data = usr_impressions_report.data
FactImpressoes	IdAnuncio	INTEGER	Facto	NA	DW	DimAnuncios	idAnuncio	INTEGER	SELECT IdAnuncio FROM DimAnuncios, usr_impressions_report WHERE DimAnuncios.ads_id = usr_impressions_report.ads_id
FactImpressoes	IdPalavra	INTEGER	Facto	NA	DW	DimPalavras	IdPalavra	INTEGER	SELECT IdPalavra FROM DimPalavras, usr_impressions_report WHERE DimPalavras.word LIKE

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
									usr_impressions_report.palavra
FactImpressoes	Impressoes	INTEGER	Facto	NA	Adwords2010	usr_impressions_report	impressoes	INTEGER	SELECT SUM(impressoes) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	Cliques	INTEGER	Facto	NA	Adwords2010	usr_impressions_report	clicks	INTEGER	SELECT SUM(clicks) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	CPCMedio	FLOAT	Facto	NA	Adwords2010	usr_impressions_report	custo, clicks	FLOAT, INTEGER	SELECT AVG(IF(clicks = 0, 0, custo/clicks)) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	RPM	FLOAT	Facto	NA	Adwords2010	usr_impressions_report	custo, impressoes	FLOAT, INTEGER	SELECT AVG(IF(impressoes = 0, 0, custo/impressoes)) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	CustoTotal	FLOAT	Facto	NA	Adwords2010	usr_impressions_report	custo	FLOAT	SELECT SUM(custo) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	CTR	FLOAT	Facto	NA	Adwords2010	usr_impressions_report	custo, impressoes	INTEGER	SELECT AVG(IF(impressoes = 0, 0, clicks/impressoes)) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra
FactImpressoes	Conversoes	INTEGER	Facto	NA	Adwords2010	usr_impressions_report	conversions	INTEGER	SELECT SUM(conversions) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra

Data Warehouse					Fontes				
Tabela	**Coluna**	**Tipo Dados**	**Tipo Tabela**	**SCD**	**Base de dados**	**Tabela**	**Coluna**	**Tipo**	**Transformação**
FactImpressoes	PosicaoMedia	FLOAT	Facto	NA	Adwords2010	usr_impressions_report	posicao_media	FLOAT	SELECT AVG(posicao_media) FROM usr_impressions_report GROUP BY usr_id, data, ads_id, palavra

Tabela 21: Mapa Lógico do Data Warehouse

Anexo H – Relatórios de *Data Profiling*

Tabela 'wrd_words'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
word	2010-11-02 05:26:42	1221633	12	3	0%	1181077	97%	1	233	19.7	!	‘sid:584	
wrd_clicks	2010-11-02 05:26:42	1221633	4	0	0%	3432	0%	1	7	1	0	2192593	35.3
wrd_id	2010-11-02 05:26:42	1221633	4	0	0%	1221633	100%	1	7	6.1	1	1222900	612,035.2
wrd_impressions	2010-11-02 05:26:42	1221633	-5	0	0%	26392	2%	1	11	1.3	0	14313948538	126,836.4
wrd_size	2010-11-02 05:26:42	1221633	4	0	0%	43	0%	1	2	1	0	46	2.5
wrd_word	2010-11-02 05:26:42	1221633	12	0	0%	1221633	100%	4	255	27.7	+m5pY28=	ZzZp	

Tabela 22: Data Profiling *da tabela 'wrd_words'*

Tabela 'usr_users'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
activation_code	2010-11-02 05:38:30	15507	12	0	0%	45	0%	0	40	0.1		faec133acdf8f028ba63561e0f1d3aadc1fdb2fe	
is_corporate	2010-11-02 05:38:30	15507	-7	0	0%	2	0%	1	1	1			

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
recover_pwd_code	2010-11-02 05:38:30	15507	12	0	0%	43	0%	0	40	0.1		fce789d6f979956ce94afe94d6f053157a9e1f4f	
recover_pwd_date	2010-11-02 05:38:30	15507	93	0	0%	44	0%	19	19	19			55,708,575,688.1
usr_addr1	2010-11-02 05:38:30	15507	12	0	0%	8692	56%	0	211	18.2	PRIOR VELHO	zooexotico1	
usr_addr2	2010-11-02 05:38:30	15507	12	30	0%	1766	11%	0	86	2.2		zz<X>ZXZ>Xz<X	
usr_balance	2010-11-02 05:38:30	15507	7	0	0%	1996	13%	4	14	4.4	-98.52	1.0E10	647,682.3
usr_cellphone	2010-11-02 05:38:30	15507	12	0	0%	5791	37%	0	20	4.1		paulo.ferreira.alves	
usr_city	2010-11-02 05:38:30	15507	12	0	0%	2131	14%	0	45	6.7		~	
usr_created	2010-11-02 05:38:30	15507	93	0	0%	10987	71%	19	19	19			20,071,473,497,605.2
usr_credit	2010-11-02 05:38:30	15507	-7	0	0%	2	0%	1	1	1			
usr_daily_costs	2010-11-02 05:38:30	15507	7	0	0%	2790	18%	4	9	4.4	0.0	392047.47	277.5
usr_deleted	2010-11-02 05:38:30	15507	-7	0	0%	2	0%	1	1	1			
usr_email	2010-11-02 05:38:30	15507	12	1	0%	11934	77%	0	100	18.4		zzy@sapo.pt	
usr_gender	2010-11-02	15507	1	10282	66%	3	0%	0	1	0.7			

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
	05:38:30												
usr_id	2010-11-02 05:38:30	15507	4	0	0%	15507	100%	1	5	4.3	1	15536	7,762.6
usr_login	2010-11-02 05:38:30	15507	12	0	0%	12803	83%	0	50	8.9		zulfatex	
usr_modified	2010-11-02 05:38:30	15507	93	0	0%	1260	8%	19	19	19			7,106,068,734,211.6
usr_name	2010-11-02 05:38:30	15507	12	0	0%	12035	78%	0	112	16.4		zulfatex	
usr_netbi	2010-11-02 05:38:30	15507	12	0	0%	15424	99%	0	50	12.7		zzy	
usr_phone	2010-11-02 05:38:30	15507	12	0	0%	6885	44%	0	15	4.9		xpto	
usr_portal	2010-11-02 05:38:30	15507	4	0	0%	7	0%	1	1	1	0	7	0.1
usr_primarymail	2010-11-02 05:38:30	15507	12	0	0%	14579	94%	0	44	17		zonab@zonab.net	
usr_pwd	2010-11-02 05:38:30	15507	12	0	0%	202	1%	0	40	0.5		ff957de7eb721feed55968576c08e941c0aafa16	
usr_status	2010-11-02 05:38:30	15507	-6	0	0%	2	0%	1	1	1	0	1	1
usr_zip	2010-11-02 05:38:30	15507	12	0	0%	7975	51%	0	30	6		ºk87099	

Tabela 23: Data Profiling *da tabela 'usr_users'*

Tabela 'ads_portals'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
adsense	2010-10-04 05:03:20	9	-7	0	0%	1	11%	1	1	1			
adwords	2010-10-04 05:03:20	9	-7	0	0%	1	11%	1	1	1			
deal_adsense	2010-10-04 05:03:20	9	7	0	0%	3	33%	6	6	6	0.0	4.0	0.55555556
deal_adwords	2010-10-04 05:03:20	9	7	0	0%	2	22%	6	6	6	0.0	1.0	0.11111111
name	2010-10-04 05:03:20	9	12	0	0%	9	100%	3	14	6.1	AEIOU	SAPO	
portal_id	2010-10-04 05:03:20	9	4	0	0%	9	100%	1	2	1.1	-1	7	3

Tabela 24: Data Profiling *da tabela 'ads_portals'*

Tabela 'cam_campaigns'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
cam_created	2010-11-04 04:53:15	32239	93	0	0%	27009	84%	19	19	19			20,077,972,783,020
cam_daily_costs	2010-11-04 04:53:15	32239	7	0	0%	122	0%	4	5	4	0.0	17.91	0
cam_deleted	2010-11-04 04:53:15	32239	-7	0	0%	2	0%	1	1	1			
cam_description	2010-11-04	32239	12	0	0%	21320	66%	2	109	18.6	bi	__test_KWD|	

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
	04:53:15										juterias	M3088-Artecer	
cam_end_date	2010-11-04 04:53:15	32239	93	0	0%	13217	41%	19	19	19			20,138,898,629,006.4
cam_id	2010-11-04 04:53:15	32239	4	0	0%	32239	100%	1	5	4.7	1	32258	16,131.1
cam_max_budget	2010-11-04 04:53:15	32239	7	0	0%	506	2%	4	11	5.2	-0.79	1.0E7	9,873.7
cam_modified	2010-11-04 04:53:15	32239	93	10807	34%	17444	54%	19	19	19			20,076,905,304,630.5
cam_start_date	2010-11-04 04:53:15	32239	93	0	0%	12488	39%	19	19	19			20,068,536,162,467.8
cam_status	2010-11-04 04:53:15	32239	-6	0	0%	2	0%	1	1	1	1	9	6
content_network	2010-11-04 04:53:15	32239	-7	0	0%	3	0%	1	1	1			
geo_id	2010-11-04 04:53:15	32239	4	0	0%	19	0%	1	2	1	0	18	0.4
site_targeted	2010-11-04 04:53:15	32239	-7	0	0%	2	0%	1	1	1			
target_site	2010-11-04 04:53:15	32239	-7	0	0%	1	0%	1	1	1			
usr_id	2010-11-04 04:53:15	32239	4	0	0%	11962	37%	1	5	4.4	0	15536	8,801.4

Tabela 25: Data Profiling *da tabela 'cam_campaigns'*

Tabela 'adg_adgroups'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
adg_created	2010-11-04 05:01:54	137894	93	0	0%	56920	41%	19	19	19	2006-04-19 18:56:11.0	2010-04-30 04:07:34.0	20,083,375,464,060.2
adg_default_url	2010-11-04 05:01:54	137894	-1	0	0%	518	0%	0	137	0.4		https://www.hosdat.com	0.1
adg_deleted	2010-11-04 05:01:54	137894	-7	0	0%	1	0%	1	1	1			
adg_description	2010-11-04 05:01:54	137894	12	0	0%	88281	64%	0	120	17		~Creditos ?	
adg_id	2010-11-04 05:01:54	137894	4	0	0%	137894	100%	1	6	5.2	1	138111	69,062.7
adg_max_cpc	2010-11-04 05:01:54	137894	7	0	0%	232	0%	4	11	4	0.0	1.0E7	219.4
adg_medium	2010-11-04 05:01:54	137894	-7	0	0%	2	0%	1	1	1			
adg_modified	2010-11-04 05:01:54	137894	93	117006	85%	15278	11%	19	19	19	2006-04-19 18:56:11.0	2010-04-30 04:07:35.0	20,086,834,292,852
adg_status	2010-11-04 05:01:54	137894	-6	0	0%	2	0%	1	1	1	0	1	0.9
cam_id	2010-11-04 05:01:54	137894	4	0	0%	32170	23%	1	5	4.9	0	32258	18,080
usr_id	2010-11-04 05:01:54	137894	4	0	0%	11953	9%	1	5	4.3	0	15536	9,878.1

Tabela 26: Data Profiling *da tabela 'adg_adgroups'*

Tabela 'ads_ads'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
adg_id	2010-11-04 05:10:34	425165	4	0	0%	126985	30%	1	6	5.5	0	138110	99,882.2
ads_clicks	2010-11-04 05:10:34	425165	4	0	0%	4461	1%	1	6	1.1	0	526599	100.3
ads_created	2010-11-04 05:10:34	425165	93	119	0%	68593	16%	19	19	19	2004-05-21 16:39:57.0	2010-04-30 04:07:33.0	20,091,366,360,576.4

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
ads_deleted	2010-11-04 05:10:34	425165	-7	0	0%	2	0%	1	1	1			
ads_id	2010-11-04 05:10:34	425165	4	0	0%	425165	100%	1	6	5.7	1	425720	212,853.9
ads_impressions	2010-11-04 05:10:34	425165	-5	0	0%	34072	8%	1	10	1.6	0	3016465421	386,099.3
ads_line1	2010-11-04 05:10:34	425165	12	0	0%	32387	8%	0	89	26.6	Divirta-se e Jogue na sua melhor	€10.000/mes em torneios gratutitos	
ads_line2	2010-11-04 05:10:34	425165	12	0	0%	29316	7%	0	255	24.6	à gestao de trafego rodoviario, iluminacão publica, sistemas de controle do Meio-Ambiente entre outros.	¬ 169.9 em vez de ¬ 189.9.Compre JÃ¡!	
ads_modified	2010-11-04 05:10:34	425165	93	389701	92%	24328	6%	19	19	19	2004-05-07 10:01:21.0	2010-04-29 19:25:21.0	20,085,862,300,583
ads_title	2010-11-04 05:10:34	425165	12	0	0%	35912	8%	0	139	16	Software de GestÃ£o PMEs	€3000 num mês?	
ads_type	2010-11-04 05:10:34	425165	-6	0	0%	2	0%	1	1	1	0	2	0
ads_url_display	2010-11-04 05:10:34	425165	12	0	0%	19776	5%	0	161	15.5		zumex.aguiaportugal.pt	
ads_url_redirect	2010-11-04 05:10:34	425165	-1	0	0%	60480	14%	0	426	68.9		https://www.virtualworlddirect.com/xxangai	0
img_path	2010-11-04 05:10:34	425165	12	233432	55%	128	0%	0	87	0		0	
img_title	2010-11-04 05:10:34	425165	12	422990	99%	1	0%	0	0	0			

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
usr_id	2010-11-04 05:10:34	425165	4	0	0%	8887	2%	1	5	4.4	0	15536	10,844.8

Tabela 27: Data Profiling *da tabela 'ads_ads'*

Tabela 'usr_payments'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
pay_end_date	2010-11-04 05:26:43	48509	93	0	0%	16589	34%	0	0	0			
pay_expected_md5	2010-11-04 05:26:43	48509	12	0	0%	1342	3%	0	54	2.9		teste	
pay_fact_emission_date	2010-11-04 05:26:43	48509	93	0	0%	4538	9%	0	0	0			
pay_fact_request_date	2010-11-04 05:26:43	48509	93	0	0%	9178	19%	0	0	0			
pay_fact_requested	2010-11-04 05:26:43	48509	-7	0	0%	3	0%	1	1	1			
pay_id	2010-11-04 05:26:43	48509	4	0	0%	48509	100%	1	5	4.8	1	48563	24,275.3
pay_init_date	2010-11-	48509	93	0	0%	48369	100%	19	19	19	2006-04-19	2010-04-30 00:36:22.0	20,077,836

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
	04 05:26:43										19:39:52.0		,424,414.6
pay_payref	2010-11-04 05:26:43	48509	12	1	0%	20599	42%	0	59	5.9		wallet:wservir:779c0482-c68e-4fb3-93dc-88379d9a2a0a:6	
pay_status	2010-11-04 05:26:43	48509	4	0	0%	4	0%	1	1	1	0	5	1
pay_type	2010-11-04 05:26:43	48509	-6	0	0%	3	0%	1	1	1	0	2	0.1
pay_value	2010-11-04 05:26:43	48509	7	0	0%	434	1%	1	12	4.2	-2.14748006E9	5000000.0	-30,568.9
usr_id	2010-11-04 05:26:43	48509	4	0	0%	7011	14%	1	5	4.2	0	15535	7,246.9

Tabela 28: Data Profiling da tabela 'usr_payments'

Tabela 'usr_impressions_report'

COLUMN	RUNDATE	RECORD COUNT	DATA TYPE	NULL COUNT	PERCENT NULL	UNIQUE COUNT	PERCENT UNIQUE	MIN LENGTH	MAX LENGTH	AVERAGE LENGTH	MIN VALUE	MAX VALUE	AVERAGE VALUE
adg_id	2010-11-09 02:52:40	61882923	4	0	0%	40995	0%	3	6	5.1	105	138018	71,851.9
adg_wrd_id	2010-11-09 02:52:40	61882923	4	0	0%	1298645	2%	3	7	6.8	829	4240110	1,797,221
ads_id	2010-11-09 02:52:40	61882923	4	0	0%	84998	0%	1	6	5.5	0	425471	114,211.5
cam_id	2010-11-09 02:52:40	61882923	4	0	0%	14192	0%	3	5	4.9	105	32211	18,311.8
campanha	2010-11-09 02:52:40	61882923	12	0	0%	12289	0%	2	109	15.5	MICROGERAÃâ¯¡ÃÆ’O	_BOM_TRABALHOST1000	
clicks	2010-11-09 02:52:40	61882923	4	0	0%	2275	0%	1	4	1	0	9363	0.5
conversions	2010-11-09 02:52:40	61882923	4	0	0%	46	0%	1	3	1	0	113	0
custo	2010-11-09 02:52:40	61882923	7	0	0%	9360	0%	4	7	4	0.0	5560.68	0.1
data	2010-11-09 02:52:40	61882923	91	0	0%	851	0%	10	10	10	2008-01-01	2010-04-30	
grupo	2010-11-09 02:52:40	61882923	12	0	0%	33163	0%	2	120	16.4	MICROGERAÃâ¯¡ÃÆ’O	~Creditos.pt?	
impressoes	2010-11-09 02:52:40	61882923	4	0	0%	187669	0%	1	8	1.8	0	68856222	2,563.9
palavra	2010-11-09 02:52:40	61882923	12	0	0,00%	257386	0%	2	96	13.7	+m5pY29z	â€šid:464	
posicao_media	2010-11-09 02:52:40	61882923	7	0	0%	1113	0%	4	5	4	0.0	20.0	2.2
url	2010-11-09 02:52:40	61882923	12	0	0%	54240	0%	0	255	9.2		ZON.pt/Zonbox	
usr_id	2010-11-09 02:52:40	61882923	4	0	0%	4306	0%	1	5	4.5	0	15535	10,036.9
wrd_id	2010-11-09 02:52:40	61882923	4	0	0%	257356	0%	1	7	5.3	6	1214766	151,071.3

Tabela 29: Data Profiling da tabela 'usr_impressions_report'

Anexo I – Esquema do Mondrian

```
<Schema name="DataMart" measuresCaption="M&#233;tricas">
  <Dimension type="StandardDimension" name="Anunciante" caption="Anunciante">
    <Hierarchy name="Anunciante" hasAll="true" allMemberName="Todos"
allMemberCaption="(Todos)" allLevelName="(Todos)" primaryKey="idAnunciante"
defaultMember="Todos">
      <Table name="dimAnunciante">
      </Table>
      <Level name="Rede" column="Rede" type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never" caption="Rede" captionColumn="Rede">
      </Level>
      <Level name="Anunciante" column="idAnunciante" nameColumn="Nome" type="Integer"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Anunciante"
captionColumn="Nome">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" name="Geografia" caption="Geografia">
    <Hierarchy name="Geografia" hasAll="true" allMemberName="Todos" allMemberCaption="(Todos)"
allLevelName="(Todos)" primaryKey="idGeografia" defaultMember="Todos" caption="Geografia">
      <Table name="dimGeografia">
      </Table>
      <Level name="Pais" column="Pais" nameColumn="Pais" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never" caption="Pa&#237;s" captionColumn="Pais">
      </Level>
      <Level name="Distrito" column="Distrito" nameColumn="Distrito" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Distrito"
captionColumn="Distrito">
      </Level>
      <Level name="Concelho" column="Concelho" nameColumn="Concelho" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Concelho"
captionColumn="Concelho">
      </Level>
      <Level name="Localidade" column="CodLocalidade" nameColumn="LocalCodPostal"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never"
caption="Localidade" captionColumn="LocalCodPostal">
      </Level>
      <Level name="CodPostal" column="CodPostal" nameColumn="CodPostal" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="C&#243;digo Postal"
captionColumn="CodPostal">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="TimeDimension" name="Data" caption="Data">
    <Hierarchy name="Data" hasAll="true" allMemberName="Todas" allMemberCaption="(Todas)"
allLevelName="(Todas)" primaryKey="idData" defaultMember="Todas">
      <Table name="dimData">
      </Table>
      <Level name="Ano" column="Ano" type="Integer" uniqueMembers="false"
levelType="TimeYears" hideMemberIf="Never" caption="Ano">
      </Level>
      <Level name="Trimestre" column="Trimestre" type="Integer" uniqueMembers="false"
levelType="TimeQuarters" hideMemberIf="Never" caption="Trimestre">
      </Level>
      <Level name="Mes" column="Mes" nameColumn="Mes" type="String" uniqueMembers="false"
levelType="TimeMonths" hideMemberIf="Never" captionColumn="Mes">
      </Level>
      <Level name="Dia" column="Dia" type="Integer" uniqueMembers="false" levelType="TimeDays"
hideMemberIf="Never" caption="Dia">
      </Level>
    </Hierarchy>
    <Hierarchy name="Semana" hasAll="true" allMemberName="Todas" allMemberCaption="(Todas)"
allLevelName="(Todas)" primaryKey="idData" defaultMember="Todas" caption="Semana">
      <Table name="dimData">
      </Table>
      <Level name="Semana" column="Semana" type="Integer" uniqueMembers="false"
levelType="TimeWeeks" hideMemberIf="Never" caption="Semana">
      </Level>
      <Level name="DiaSemana" column="DiaSemana" nameColumn="DiaSemana" type="String"
```

```
uniqueMembers="false" levelType="TimeDays" hideMemberIf="Never" caption="Dia da Semana"
captionColumn="DiaSemana">
      </Level>
      <Level name="Dia" column="Dia" nameColumn="Dia" type="Integer" uniqueMembers="false"
levelType="TimeDays" hideMemberIf="Never" caption="Dia" captionColumn="Dia">
      </Level>
    </Hierarchy>
  </Dimension>
  <Cube name="Impressoes" caption="Impress&#245;es" cache="false" enabled="true">
    <Table name="FactImpressoes">
    </Table>
    <Dimension type="StandardDimension" foreignKey="idAnuncio" name="Anuncio"
caption="An&#250;ncio">
      <Hierarchy name="Campanha" hasAll="true" allMemberName="Todas"
allMemberCaption="(Todas)" allLevelName="(Todas)" primaryKey="idAnuncio" defaultMember="Todas"
caption="Campanha">
        <Table name="dimAnuncio">
        </Table>
        <Level name="TipoCampanha" column="CampanhaSegmentacao"
nameColumn="CampanhaSegmentacao" type="String" uniqueMembers="false" levelType="Regular"
hideMemberIf="Never" caption="Tipo de Campanha" captionColumn="CampanhaSegmentacao">
        </Level>
        <Level name="Campanha" column="cam_id" nameColumn="CampanhaDescricao" type="Integer"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Campanha"
captionColumn="CampanhaDescricao">
        </Level>
        <Level name="Grupo" column="adg_id" nameColumn="GrupoDescricao" type="Integer"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="Grupo"
captionColumn="GrupoDescricao">
        </Level>
        <Level name="Anuncio" column="ads_id" nameColumn="Titulo" type="Integer"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="An&#250;ncio">
        </Level>
      </Hierarchy>
      <Hierarchy name="Tipo" hasAll="true" allMemberName="Todos" allMemberCaption="(Todos)"
allLevelName="(Todos)" primaryKey="idAnuncio" defaultMember="Todos" caption="Tipo de
An&#250;ncio">
        <Table name="dimAnuncio">
        </Table>
        <Level name="Tipo" column="Tipo" nameColumn="Tipo" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never" caption="Tipo de An&#250;ncio" captionColumn="Tipo">
        </Level>
        <Level name="Anuncio" column="ads_id" nameColumn="Titulo" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never" caption="An&#250;ncio"
captionColumn="Titulo">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="idPalavra" name="Palavras">
      <Hierarchy name="Palavras" hasAll="true" allMemberName="Todas"
allMemberCaption="(Todas)" allLevelName="(Todas)" primaryKey="idPalavra"
defaultMember="Todas">
        <Table name="dimPalavras">
        </Table>
        <Level name="Palavras" column="word" nameColumn="Palavra" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <DimensionUsage source="Anunciante" name="Anunciante" caption="Anunciante"
foreignKey="idAnunciante">
    </DimensionUsage>
    <DimensionUsage source="Geografia" name="Geografia" caption="Geografia"
foreignKey="idGeografia">
    </DimensionUsage>
    <DimensionUsage source="Data" name="Data" caption="Data" foreignKey="idData">
    </DimensionUsage>
    <Measure name="Impressoes" column="impressoes" datatype="Integer" aggregator="sum"
caption="Impress&#245;es" visible="true">
    </Measure>
    <Measure name="Clicks" column="clicks" datatype="Integer" aggregator="sum"
caption="Cliques" visible="true">
```

```
    </Measure>
    <Measure name="CPCMedio" column="CPCMedio" datatype="Numeric" aggregator="avg"
caption="CPC M&#233;dio" visible="true">
    </Measure>
    <Measure name="CTR" column="CTR" datatype="Numeric" aggregator="avg"
caption="ClickThroughRate" visible="true">
    </Measure>
    <Measure name="RPM" column="RPM" datatype="Numeric" aggregator="avg" caption="Revenue Per
Mil" visible="true">
    </Measure>
    <Measure name="Conversoes" column="conversoes" datatype="Integer" aggregator="sum"
caption="Convers&#245;es" visible="true">
    </Measure>
    <Measure name="Custo" column="custo" datatype="Numeric" aggregator="sum" caption="Custo"
visible="true">
    </Measure>
    <Measure name="PosicaoMedia" column="posicao_media" datatype="Numeric" aggregator="avg"
visible="true">
    </Measure>
  </Cube>
  <Cube name="Carregamentos" caption="Carregamentos" cache="true" enabled="true">
    <Table name="FactCarregamentos" alias="Carregamentos">
    </Table>
    <DimensionUsage source="Anunciante" name="Anunciante" foreignKey="idAnunciante">
    </DimensionUsage>
    <DimensionUsage source="Geografia" name="Geografia" foreignKey="idGeografia">
    </DimensionUsage>
    <DimensionUsage source="Data" name="Data" foreignKey="idData">
    </DimensionUsage>
    <Dimension type="StandardDimension" foreignKey="idTipo" name="TipoPagamento" caption="Tipo
de Pagamento">
      <Hierarchy name="Tipo" hasAll="true" allMemberName="Todos" allMemberCaption="(Todos)"
allLevelName="(Todos)" primaryKey="idTipo" defaultMember="Todos" caption="Tipo de Pagamento">
        <Table name="dimTipoPagamento">
        </Table>
        <Level name="Tipo" column="Tipo" nameColumn="Tipo" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never" caption="Tipo de Pagamento" captionColumn="Tipo">
        </Level>
      </Hierarchy>
    </Dimension>
    <Measure name="Valor" column="Valor" datatype="Numeric" aggregator="sum" visible="true">
    </Measure>
  </Cube>
</Schema>
```

Informação pessoal

Apelido(s) / Nome(s) próprio(s)	**Ferreirinho Nunes, Filipe Manuel**
Morada(s)	Vale de Milhaços 2855-424 CORROIOS
Telefone(s)	Telemóvel
Data de nascimento	24/04/1975

Emprego pretendido / Área funcional

Business Intelligence

Experiência profissional

Datas	01 de Setembro 2009 →
Função ou cargo ocupado	Professor de Informática
Principais atividades e responsabilidades	- Lecionação de várias disciplinas de Informática no âmbito do ensino profissional e do terceiro ciclo do ensino básico, nomeadamente Sistemas Digitais e Arquiteturas de Computadores, Comunicação de Dados, Área de Projeto e Formação Cívica; - Direção de turma do terceiro ciclo do ensino básico.
Nome e morada do empregador	Escola Secundária João de Barros Rua Dr. Manuel de Arriaga, Corroios
Tipo de empresa ou sector	Educação

Datas	01 de Setembro 2006 - 31 de Agosto 2009
Função ou cargo ocupado	Professor de Informática
Principais atividades e responsabilidades	- Lecionação de várias disciplinas de Informática no âmbito do ensino profissional e secundário, nomeadamente Tecnologias da Informação e da Comunicação e Organização de Empresas e Aplicações de Gestão; - Direção de turma do ensino profissional; - Em paralelo, realizou diversas intervenções no sistema de gestão de leasing do Millennium BCP, na plataforma Lotus Notes/Domino, como consultor independente.
Nome e morada do empregador	Escola Secundária de Santo André Av. Escola Fuzileiros Navais, Barreiro
Tipo de empresa ou sector	Educação

Datas	01 de Setembro 2005 - 31 de Agosto 2006
Função ou cargo ocupado	Professor de Informática
Principais atividades e responsabilidades	- Lecionação de várias disciplinas de Informática no âmbito do terceiro ciclo do ensino básico e secundário, nomeadamente Tecnologias da Informação e da Comunicação e Redes e Internet.
Nome e morada do empregador	Escola Secundária da Baixa da Banheira Praceta Maria Helena Vieira da Silva, Vale da Amoreira, Moita
Tipo de empresa ou sector	Educação

Datas	01 de Outubro 2003 - 31 de Agosto 2005
Função ou cargo ocupado	Consultor de Tecnologias de Informação
Principais atividades e responsabilidades	- Projeto de evolução do sistema de gestão documental do Banco de Portugal, construído sobre a plataforma Lotus Notes/Domino e Domino.Doc; - Intervenção no sistema de gestão de leasing do Millennium BCP, na plataforma Lotus Notes/Domino.

Nome e morada do empregador	Novabase Av. Dom João II, Lote 1.03.2.3 Parque das Nações, Lisboa
Tipo de empresa ou sector	Atividades de Consultoria, Científicas, Técnicas e Similares
Datas	08 de Março 1998 - 30 de Setembro 2003
Função ou cargo ocupado	Consultor de Tecnologias de Informação
Principais atividades e responsabilidades	- Coordenação e participação em equipas de projeto de tecnologias de informação na área da Banca, Seguros e Serviços Profissionais; - (de Março de 1998 a Agosto de 1999) – Membro português da equipa do Centro de Competência de Data Warehousing da região EMEIA (Europa, Médio Oriente, Índia e África), organismo responsável pela partilha de conhecimento e definição de metodologias no âmbito da disciplina referida na Arthur Andersen. - Projetos mais relevantes: - Sistema de Workflow para o Barclays Bank Portugal em Lotus Notes e diversas intervenções evolutivas no mesmo sistema; - Desenvolvimento de materiais para ações de formação em equipas multinacionais; - Conversão para o Euro do sistema de gestão de cobranças e tesouraria na seguradora Mundial-Confiança, atual Fidelidade-Mundial (COBOL/CICS/DB2); - Coordenação do desenvolvimento das aplicações de gestão de propostas e Quality Assurance da Deloitte & Touche, em Lotus Notes, e sua posterior conversão para a plataforma .NET da Microsoft.
Nome e morada do empregador	Deloitte & Touche / Arthur Andersen Praça Duque de Saldanha, 1 - 6º, Lisboa (Portugal)
Tipo de empresa ou sector	Atividades de Consultoria, Científicas, Técnicas e Similares

Educação e formação

Datas	01/09/2007 - 31/08/2008
Designação da qualificação atribuída	Sistemas Integrados de Apoio à Decisão (Business Intelligence)
Principais disciplinas/competências profissionais	Business Intelligence, ETL, Data Warehousing, OLAP, Reporting, Data Mining.
Nome e tipo da organização de ensino ou formação	Instituto Universitário de Lisboa - ISCTE Av. das Forças Armadas, Lisboa
Nível segundo a classificação nacional ou internacional	Pós Graduação
Datas	01/09/2005 - 31/07/2007
Designação da qualificação atribuída	Habilitação Profissional para a Docência – grupo 550 (Profissionalização em Serviço)
Principais disciplinas/competências profissionais	Psicologia e Sociologia da Educação, Desenvolvimento Curricular, Didática Específica
Nome e tipo da organização de ensino ou formação	Instituto Politécnico de Setúbal (Escola Superior de Educação) Campus do Instituto Politécnico de Setúbal, Setúbal
Nível segundo a classificação nacional ou internacional	Pós Graduação
Datas	01/10/1993 - 31/07/1998
Designação da qualificação atribuída	Informática e Gestão de Empresas
Nome e tipo da organização de ensino ou formação	Instituto Superior de Ciências do Trabalho e da Empresa Av. das Forças Armadas, Lisboa
Nível segundo a classificação nacional ou internacional	Licenciatura

Aptidões e competências pessoais

Primeira língua

Português

Outra(s) língua(s)

Auto-avaliação	Compreensão				Conversação				Escrita	
Nível europeu (*)	Compreensão oral		Leitura		Interação oral		Produção oral			
Inglês	C1	Utilizador avançado	C1	Utilizador avançado	C1	Utilizador avançado	C1	Utilizador avançado	C1	Utilizador avançado
Espanhol / Castelhano	B1	Utilizador independente	B1	Utilizador independente	A2	Utilizador básico	A2	Utilizador básico	A1	Utilizador básico
Francês	A1	Utilizador básico	A1	Utilizador básico	A1	Utilizador básico	A1	Utilizador básico	A1	Utilizador básico

(*) Nível do Quadro Europeu Comum de Referência (CECR)

Aptidões e competências sociais

- Espírito de equipa;
- Capacidade de adaptação a ambientes multi-culturais, adquirida com experiência de trabalho no estrangeiro;
- Boa capacidade de comunicação adquirida através da experiência como professor.

Aptidões e competências de organização

Contabilidade e Finanças, adquirida através de formação de base e experiência profissional na banca.

Aptidões e competências informáticas

Sistemas de informação, SQL, UML, programação em várias linguagens (C, Java, .NET, entre outras), Business Intelligence (Pentaho), Linux, open source.

Aptidões e competências artísticas

Música

Carta de condução

A1, A, B

Informação adicional

Membro da Direção da Associação de Motociclistas Cristãos de Portugal, onde desempenha o cargo de Tesoureiro.