OLAP em âmbito hospitalar: Transformação de dados de enfermagem para análise multidimensional

João Silva and José Saias

m5672@alunos.uevora.pt, jsaias@di.uevora.pt
Mestrado em Engenharia Informática, Universidade de Évora

Resumo O desenvolvimento a que assistimos nos dias que correm leva-nos a ter que tomar decisões cada vez mais sustentadas e correctas. Tal encaminha-nos para uma procura incessante de mais e melhor informação.
O uso de bases de dados e a constante angariação de dados subjacente para salvaguardar e optimizar o funcionamento das organizações de diferentes áreas vieram trazer um grande avanço na perspectiva da colecta de dados que posteriormente se tornam numa grande fonte de conhecimentos.
Os Data Warehouses (DW), aliados a sistemas de Online Analytical Processing (OLAP), estão contidos no rol de soluções disponíveis para análise de dados e são uma grande mais valia para os analistas que vêem o seu trabalho bastante facilitado.
Este artigo visa descrever um pouco, estas duas tecnologias e mostrar como elas se complementam, levando ao desenvolvimento de uma ferramenta incluída na temática de Business Intelligence.

1 Introdução

O facto de estarmos bem informados sobre uma determinada área do nosso interesse sempre foi vital para o desenrolar do nosso comportamento e para qualquer tomada de decisão da nossa parte.

O problema que se põe é o facto de nem sempre a informação que procuramos estar disponível de forma clara. Isto leva-nos a procurar soluções sofisticadas de forma a conseguir descobrir essa informação, contida no meio de uma quantidade infindável de dados armazenados.

Quando falamos ao nível organizacional essas informações tomam proporções bastante maiores, uma vez que nas organizações qualquer mudança com vista a melhorar, pode trazer bastantes vantagens tanto ao nível funcional como principalmente ao nível lucrativo.

Durante muito tempo estas tomadas de decisão foram orientadas exclusivamente pelo saber administrativo, sem necessidade de se recorrer a bases de dados organizadas.

Nos dias que correm o uso das bases de dados e a constante angariação de dados durante o funcionamento das organizações trouxe-nos a possibilidade

de armazenar uma quantidade enorme de informação. Desta feita existe um conceito designado por *Business Intelligence* que podemos traduzir facilmente por Inteligência nos Negócios.

Existem várias técnicas ou metodologias englobadas no conceito de BI, onde todas elas têm o intuito comum de fornecer conhecimentos importantes.

Podemos então definir o conceito de BI um conjunto de componentes e processos, que juntos permitem a angariação dos dados provenientes de diversas fontes, organizando-os, processando-os e armazenando-os de forma correcta para que sejam apresentados ao utilizador final, de modo a facilitar o processo de tomada de decisões[9].

Neste contexto os sistemas de OLAP constituem uma das soluções mais utilizadas por parte das organizações, pois estas fornecem ao utilizador final a possibilidade de navegação pelos dados de forma bastante intuitiva, sendo possível a criação de tabelas, gráficos, entre outras possibilidades.

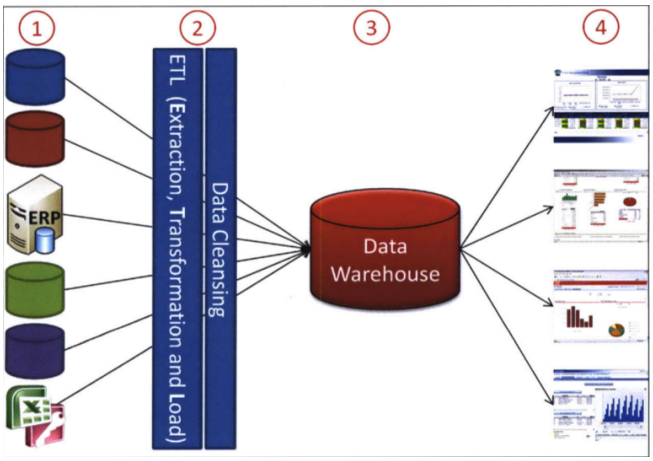

Figura 1. Diversos passos para o desenvolvimento de uma solução de BI[9]

O bom funcionamento destas ferramentas depende muito de uma boa implementação do repositório onde os dados analisados estão armazenados. Estes repositórios designam-se por Data Warehouses e, como podemos observar na figura 1, estes constituem a parte central e uma das partes mais importantes para o bom funcionamento deste tipo de ferramentas.

2 Data Warehouse

Ao longo dos anos, têm surgido diversas formas de definir um DW. Em termos teóricos e de acordo com W. H. Inmon, o seu principal arquitecto, “a data

warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection os data in support of management's decision making process"[4]. Por meio destes quatro termos é possível descrever as características de um DW, isto é:

- Subject-oriented: um DW é desenvolvido e organizado, de modo a satisfazer as necessidades de análise de uma organização, relativamente a um ou mais aspectos chave. Relativamente ao trabalho realizado e descrito neste artigo, um dos aspectos chave foi a Taxa de prevalência relativa aos diversos diagnósticos, presentes nos registos de enfermagem.
- Integrated: um DW é por norma desenvolvido, utilizando diversas fontes de dados externas, como bases de dados relacionais, folhas de cálculo, entre outras. Como tal, alguns problemas de consistência dos dados necessitam de resolução.
- Nonvolatile: um DW apenas permite o carregamento dos e a leitura dos mesmos. Operações de modificação e de remoção não são permitidas.
- Time-variant: os dados são armazenados de modo a fornecerem uma perspectiva histórica dos dados.

A necessidade do uso deste tipo especial de armazenamento surge, devido ao facto de as bases de dados convencionais estarem mais optimizadas para gerir um grande número de transacções e um constante fluxo de dados com a preocupação de manter a consistência dos dados. Como tal, não estão preparadas para em tempo útil processar consultas complexas, que são efectuadas por sistemas de análise como é o caso do OLAP[3].

Em suma estes são descritos como uma arquitectura que obedece a determinadas especificações técnicas, sendo formado através da integração de diversas fontes de dados, para o suporte de sistemas de análise com o intuito de optimizar e apoiar as organizações no processo de tomada de decisões.

2.1 Modelo Multidimentional

Para proporcionarem uma boa fluidez os DW baseiam-se no modelo multidimensional, também conhecido por *Cubo de dados*.

Através da observação da figura 2 podemos constatar a existência de **dimensões** que fornecem o contexto ao utilizador sobre o qual se deseja analisar as **medidas** existentes. Estas medidas correspondem aos valores quantificáveis que são usados para analisar as relações entre as dimensões[3]. Um exemplo possível e referente ao trabalho realizado, é a análise da medida *Taxa de Prevalência* num determinado contexto, fornecido pelas dimensões *Tempo*, *Geográfica* e *Diagnóstico*.

De referir que dentro de cada dimensão, existem *hierarquias*, estruturadas de modo a definirem vários níveis de granularidade, que pode ser maior ou menor conforme se sobe ou desce, respectivamente na hierarquia[5]. Dentro da dimensão tempo, uma hierarquia possível é a sequência constituída por anos, seguida de meses e por último dias. Sendo que o nível anos é o que possui menor granularidade e por sua vez o nível dias o que possui maior. Por último, cada nível de

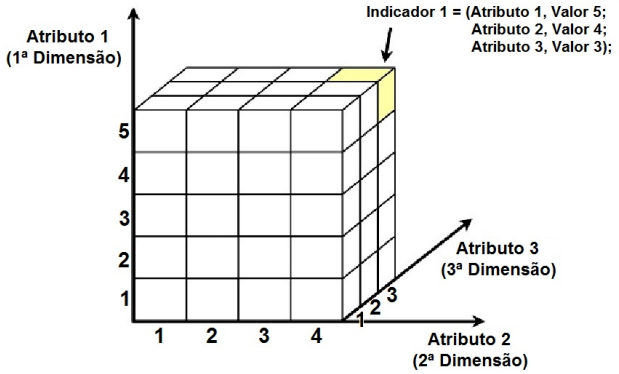

Figura 2. Cubo de dados

uma hierarquia possui os seus *membros*, o que permite filtrar os dados dentro de uma dimensão, caso seja necessário encontrar uma situação mais concreta.

2.2 Esquema em Estrela e Esquema em Floco de Neve

Estes dois esquema representam duas maneiras de implementar o modelo multidimensional em bases de dados relacionais.

O esquema em estrela é o principal esquema utilizado e, ao mesmo tempo, o mais comum[1]. Contém uma tabela central normalmente designada por *tabela de factos*, ou seja, é a tabela que contém os factos que podem corresponder às medidas ou que possibilitam o seu cálculo, e por sua vez é a que possui maior quantidade de dados. À volta desta estão tabelas mais pequenas que representam as tabelas das dimensões.

A diferença entre o anterior e o esquema em floco de neve é a normalização das tabelas das dimensões, pois em vez de cada dimensão ser constituída por uma só tabela, estas estão divididas em várias.

Com esta normalização, algumas hierarquias passam a ser explícitas[10]. Esta vem trazer benefícios ao nível do espaço utilizado pelas tabelas das dimensões, sendo necessário menos espaço de armazenamento, diminuindo também a redundância destas tabelas. Contudo, dado que as tabelas estão separadas, são precisos mais *joins* aquando da execução de consultas, o que pode afectar bastante a performance.

2.3 Extracção, Transformação e Carregamento (ETL)

Posteriormente à estruturação do **DW**, inicia-se o processo de **ETL**. Este processo é por norma o mais difícil e moroso quando se trata do desenvolvimento deste tipo de soluções[9]. O principal objectivo é a angariação e a transferência dos dados de diversas fontes de uma organização para o **DW**, ou seja, juntar os

dados, por norma heterogéneos, para uma representação homogénea que permita processos de análise eficazes e eficientes.

Por vezes, deve ser realizado o processo de *Limpeza dos Dados* antes que estes sejam extraídos, transformados e finalmente carregados para o DW. Tal deve-se ao facto de os dados nem sempre se apresentarem de forma correcta ou completa, criando problemas de consistência.

3 Servidores OLAP

Entre o DW e as ferramentas de análise, existe ainda uma camada bastante importante, designada por *Servidor OLAP*, dado que OLAP funciona através do conceito multi-utilizador cliente/servidor[6].

Os dois servidores OLAP mais utilizados são o **M**ultidimensional **OLAP** e o **R**elational **OLAP** e a grande diferença entre eles é a forma como os dados são armazenados.

Multidimentional OLAP Como o nome indica no MOLAP os dados são armazenados de forma multidimensional, isto é, em estruturas multidimensionais do tipo array. O seu funcionamento baseia-se no calculo prévio das agregações de diversas combinações das dimensões existentes, sendo estas armazenadas nas estruturas mencionadas anteriormente[11].

Este tipo de armazenamento não é muito apropriado aquando da existência de um grande número de dimensões dado que ao efectuar o pré-calculo de todas as combinações possíveis, o tempo necessário para efectuar esta operação pode tornar-se bastante elevado.

Relational OLAP Ao contrário do anterior, este tipo de servidor utiliza a própria base de dados relacional como forma de armazenamento. Como tal, o seu papel é servir de intermediário entre o servidor relacional, onde estão guardados os dados, e o cliente, estendendo as capacidades destes servidores. Deste modo estes passam a suportar as consultas multidimensionais, características das ferramentas OLAP[2].

Por norma quando se trata de um conjunto baixo de dimensões a diferença de performance para o MOLAP não é significativa e no caso de haver um aumento no número destas ou no volume de dados, então o ROLAP ganha vantagem pois como não necessita de efectuar o cálculo prévio das agregações, o tempo de resposta a este cenário é bem mais rápido[11].

4 Operações OLAP

Recorrendo à organização dos dados de forma multidimensional e aliado ao facto das dimensões serem hierárquicas, as ferramentas de OLAP permitem uma grande flexibilidade de navegação por entre os dados.

As operações mais importantes fornecidas por estas ferramentas são as seguintes:

- **Drill-Up** ou **Roll-Up** - permite subir na hierarquia de uma dimensão ou mesmo de remove-la. Por exemplo, passar de uma vista por cidades para a vista por distritos, subindo na hierarquia da dimensão localização;
- **Drill-Down** - permite efectuar oposto da operação anterior;
- **Slice e Dice** - permitem efectuar cortes na visualização dos dados. Por exemplo, podemos querer visualizar os dados relativos apenas ao primeiro trimestre do ano 2010, o que corresponde a efectuar um *slice* na dimensão tempo;
- **Drill-Through** - permite observar a fonte dos dados que deram origem a uma determinada agregação.

Embora estas sejam as principais funcionalidades normalmente presentes neste tipo de ferramentas, existem outras, como a capacidade criar gráficos, exportar os dados observados para outros formatos como PDF ou Excel, etc.

5 Trabalho Realizado

Com base nos conhecimentos adquiridos que foram, de uma forma resumida, descritos nas secções anteriores deste artigo, falta então descrever um pouco o trabalho realizado. Através deste foi desenvolvida uma solução de OLAP de modo a permitir a análise de medidas, relativas a registos de enfermagem. Como tal, foi necessário efectuar todo o processo de construção de um DW, criando o repositório multidimensional, efectuando o processo de ETL necessário.

Para a realização do mesmo foram utilizadas ferramentas open source disponíveis para a comunidade que, à excepção do sistema de gestão de bases de dados MySQL[7], são pertencentes à empresa Pentaho[8], responsável pelo desenvolvimento de tecnologias direccionadas para a temática de BI.

5.1 Repositório Multidimentional

O primeiro passo foi desenvolver o repositório multidimensional, ou seja, a base de dados que constitui o DW.

Na fase inicial utilizou-se um esquema em estrela para desenvolver a base de dados, tal como mostra a figura 3. Através da observação deste podemos perceber quais as dimensões utilizadas, dentro destas quais os campos que permitem formar as hierarquias de cada uma e por último quais os factos que permitem o cálculo da medida referente a esta tabela de factos. A medida *taxa de prevalência* é posteriormente calculada através do campo *codigo_pacientes* que constituem os factos desta tabela.

Dado que este trabalho não se referia a apenas uma medida, ao esquema da figura 3, foram adicionadas mais duas tabelas de factos em que estas possuem dimensões exclusivas de cada uma e outras partilhadas entre si.

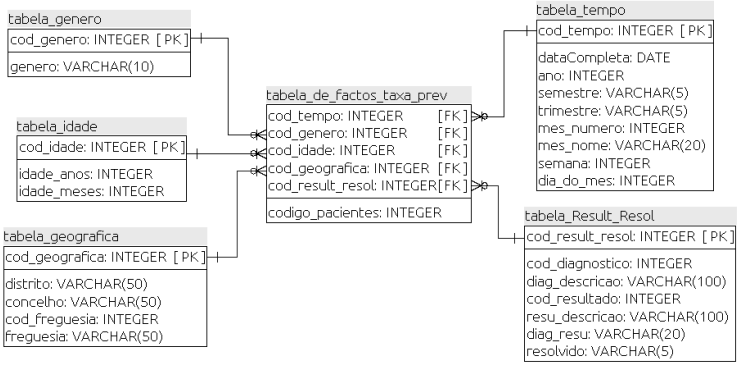

Figura 3. Esquema em estrela

5.2 Processo de ETL

Depois de construído o repositório, o passo seguinte foi extrair os dados da única fonte existente, uma base de dados relacional pertencente a um sistema proprietário de suporte ao processo de registos de enfermagem. De seguida efectuar as transformações necessárias e por último carrega-los para o DW. Para tal foi utilizada uma ferramenta designada por Pentaho Data Integration, a qual permitiu efectuar todo este processo.

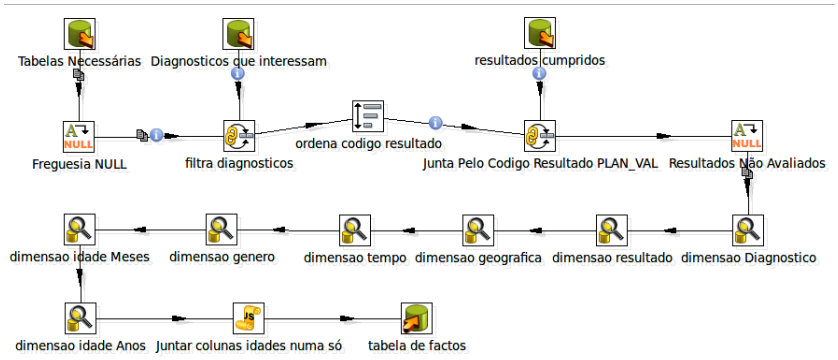

Figura 4. Transformação para a tabela de factos referente à medida Taxa de Prevalência

Como podemos observar pela figura 4, à medida que vão sendo feitas as consultas à base de dados fonte, de modo a extrair a informação necessária, vão sendo efectuadas as transformações nos dados extraídos através dos passos seguintes.

Como as tabelas de factos possuem as chaves primárias de cada dimensão, foi necessário encontrar as diferentes chaves correspondentes para cada facto antes de ser efectuado o carregamento, Como tal, foi necessário efectuar um processo semelhante de carregamento para todas as dimensões, de modo a ser possível fornecer o contexto a cada facto.

5.3 OLAP

Para desenvolver a fase final do trabalho, foi utilizado um servidor OLAP baseado na arquitectura ROLAP, designado por Pentaho BI Server. Aliado a este foi utilizado também uma ferramenta de visualização designado por STPivot, responsável pelo fornecimento das operações OLAP anteriormente referidas.

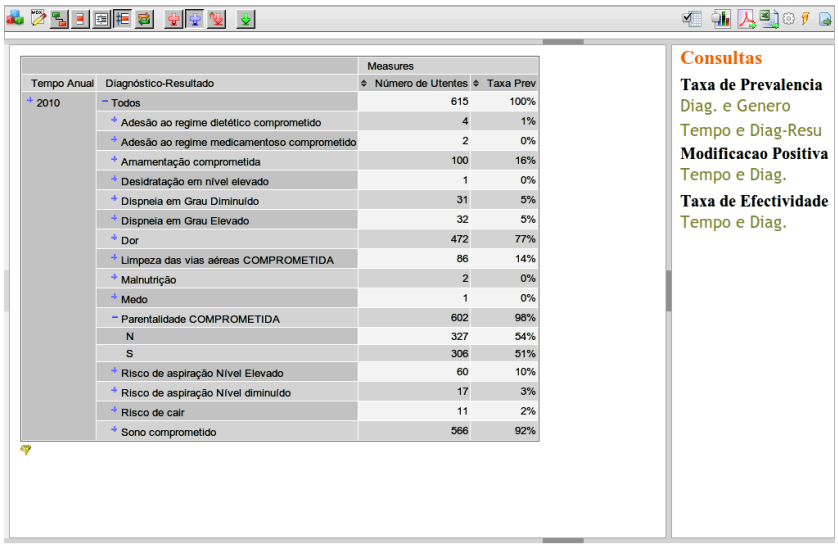

Figura 5. Consulta efectuada através do STPivot

A figura 5 mostra o aspecto final da ferramenta de OLAP e representa a consulta da taxa de prevalência por tempo, neste caso referente ao ano de 2010 e por diagnóstico. Nesta dimensão foi efectuado o *drill-down*, mostrando todos os diagnósticos existentes no ano de 2010 e por sua vez um segundo drill-down, mostrando os resultados resolvidos e não resolvidos para o diagnóstico *Parentalidade Comprometida* no mesmo ano.

De referir que as consultas são efectuadas numa linguagem multidimensional desenvolvida exclusivamente para os sistemas OLAP, designada por MDX. Como tal foi necessário criar um esquema através da ferramenta Schema Workbench, responsável por fornecer ao servidor ROLAP a informação necessária de modo a que este consiga efectuar a tradução das consultas MDX para código SQL referente à base de dados que constitui o DW.

6 Conclusão

Todas estas tecnologias tanto de armazenamento, como de análise, entre outras, têm vindo a evoluir bastante e de modo síncrono, dado que se trata de tecnologias complementares. Embora sejam os sistemas OLAP que permitem a interacção com os dados, não nos podemos esquecer que sem os Data Warehouses, estas se tornariam inúteis, pois uma a boa estruturação dos dados é fundamental para o bom funcionamento destas ferramentas de análise.

OLAP tem vindo a ganhar um papel cada vez mais importante na vida das organizações, não só na área dos negócios mas também noutras, como a da saúde, onde a tomada de decisões é bastante importante de modo a ganhar uma maior eficiência e rapidez nos serviços.

Como exemplo disso está o trabalho realizado onde o objectivo principal foi desenvolver um sistema de OLAP para a área da saúde mais propriamente para análise de registos de enfermagem.

Referências

1. Gauree Bhole. Building a data mart using star schema, 2010.
2. S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. 1997.
3. J. Han and M. Kamber. Data mining: Concepts and techniques, 2000.
4. W. H. Inmon. *Building the Data Warehouse*. QED Technical Publishing Group, Wellesley, Massachusetts, 1992.
5. E. Malinowski and E. Zimányi. Hierarchies in a multidimensional model: From conceptual modeling to logical representation. *Data & Knowledge Engineering*, 2006.
6. Fred Curtis Moulton. Olap and olap server definitions. `http://www.moulton.com/olap/olap.glossary.html`. Acedido em Outubro de 2011.
7. Mysql. Mysql: The world's most popular open source database. `http://www.mysql.com/`.
8. Pentaho. Pentaho open source business intelligence. `http://www.pentaho.com`.
9. Eumir P. Reyes. A systems thinking approach to business intelligence solutions based on cloud computing, 2010.
10. Yin Jenny Tam. Datacube: Its implementation and application in olap mining, 1998.
11. Per Westerlund. Business intelligence: Multidimensional data analysis, 2008.