Data Warehousing na Saúde: Melhorando a Tomada de Decisão Médico-Analítica

Cristina Dutra de Aguiar Ciferri
Universidade de São Paulo
Departamento de Ciências de Computação
São Carlos, SP, Brasil, CEP 13.560-970
cdac@icmc.usp.br

Ricardo Rodrigues Ciferri
Universidade Federal de São Carlos
Departamento de Computação
São Carlos, SP, Brasil, CEP 13.565-905
ricardo@dc.ufscar.br

Reinaldo Jiunji Ogata
André Moraes Paula Lima
Universidade Estadual de Maringá
Departamento de Informática
Maringá, PR, Brasil, CEP 87.020-900
amplima@din.uem.br

Agma Juci Machado Traina
Universidade de São Paulo
Departamento de Ciências de Computação
São Carlos, SP, Brasil, CEP 13.560-970
agma@icmc.usp.br

Abstract

In this paper we investigate the use and the importance of the data warehousing technology in the healthcare area. We propose five multidimensional schemas focusing on the most interesting business processes of three medical institutions: monthly costs of products by sector, occupation of facilities, disease treatment, costs of medical procedures and pregnancy control. The proposed schemas consider data granularity regarding different dimensions and are aimed at the three levels of information of any healthcare institution. Besides, these schemas can be used as a basis for creating new healthcare data warehousing applications. In this paper we also describe aspects related to the implementation of the monthly costs of products by sector multidimensional schema.

Keywords: Database in healthcare area, Data warehousing, Multidimensional modelling.

Resumo

Este artigo enfoca a aplicação e a importância da tecnologia de *data warehousing* na área de saúde. Sua principal contribuição é a proposta de cinco esquemas multidimensionais referentes aos assuntos de maior interesse para três instituições médicas: custo mensal de produtos por setor, ocupação de instalações, tratamento por doença, custo dos procedimentos e controle de gestantes. Os esquemas propostos tratam a granularidade dos dados considerando-se diferentes dimensões e enfocam os três níveis de informação de uma instituição de saúde. Ademais, estes esquemas servem de base para a criação de novas aplicações de *data warehousing* para a área de saúde. O artigo também descreve aspectos de implementação do esquema multidimensional custo mensal de produtos por setor.

Palavras chaves: Banco de dados em saúde, *Data warehousing*, Modelagem multidimensional.

1. INTRODUÇÃO

A obtenção de informações estratégicas, relativas ao contexto de tomada de decisão, é de suma importância para o sucesso de qualquer instituição. Isto se verifica inclusive na área de saúde, na qual a importância do negócio tem se mostrado cada vez mais evidente. A análise de informações de saúde pode subsidiar processos decisórios possibilitando o planejamento de ações de saúde de forma mais eficiente. Por exemplo, instituições de saúde governamentais podem definir estratégias mais eficazes contra surtos de doenças, incluindo campanhas de vacinação e campanhas educativas.

Nesse sentido, o *data warehousing* (DWing) surge como um ambiente altamente eficaz, que atende às necessidades de obtenção de informação estratégica sobre o negócio, por meio da recuperação sumarizada de informação. Nesse ambiente, o acesso à informação de provedores autônomos, heterogêneos e distribuídos é realizado, geralmente, em duas

etapas: (i) uma vez que estes provedores podem possuir uma variedade de formatos e modelos e podem incluir desde sistemas de banco de dados relacionais até bases de conhecimento, sistemas legados e documentos não estruturados, a informação de cada provedor é previamente extraída, traduzida, filtrada, integrada à informação relevante de outros provedores e finalmente armazenada no *data warehouse* (DW); e (ii) as consultas, quando realizadas, são executadas diretamente no DW, sem acessar os provedores originais [4, 5, 12, 20, 28].

Desta forma, a informação integrada torna-se disponível para consulta ou análise imediata de usuários de sistemas de suporte à decisão (SSD), tais como o superintendente de um hospital, membros da diretoria e líderes de equipes médicas. Uma aplicação de DWing para a área de saúde, por exemplo, poderia integrar informações relativas a determinadas doenças ao longo dos anos. Usuários de SSD desta aplicação poderiam realizar, utilizando estas informações integradas, análises de tendências simples (e.g., "Qual a incidência de dengue em julho de 2005?"), análises comparativas (e.g. "Qual a incidência de dengue nos últimos três anos?") e análises de tendência múltiplas (e.g. "Qual a incidência de dengue nos últimos três anos, nas diferentes estações do ano?").

Como o DW armazena informações integradas, cujas diferenças semânticas e de modelo já foram eliminadas, ambas as consultas e as análises podem ser realizadas rápida e eficientemente. Além disto, uma vez que as consultas e as análises são executadas diretamente no DW sem acessar os provedores originais, o DW encontra-se disponível mesmo quando os provedores não estiverem disponíveis. Outra vantagem refere-se ao fato de que o processamento local nos provedores originais não é afetado por causa da participação destes no ambiente de DWing.

Este artigo descreve uma experiência do uso da tecnologia de DWing na área médica. Suas principais contribuições para a área de banco de dados em saúde são:

- a proposta de esquemas multidimensionais referentes aos assuntos de maior interesse para três instituições reais da área de saúde da cidade de Maringá, Brasil. Esses esquemas enfocam cinco diferentes assuntos: (i) custo mensal de produtos por setor; (ii) ocupação de instalações; (iii) tratamento por doença; (iv) custo dos procedimentos; e (iv) controle de gestantes; e
- a validação de um dos esquemas propostos, por meio da discussão de sua implementação utilizando o SGBD (sistema gerenciador de banco de dados) Oracle9i® e ferramentas de DWing da Oracle® Corporation.

O objetivo dos esquemas multidimensionais propostos é, além de atender às necessidades das instituições para as quais foram projetados, servir de base para a criação de novas aplicações de DWing para a área de saúde. Desta forma, os esquemas propostos apresentam grande aplicabilidade prática.

O artigo está estruturado da seguinte forma. A próxima seção resume trabalhos correlatos. A fundamentação teórica é apresentada nas seções 3 e 4. Mais detalhadamente, a seção 3 destaca motivações para a utilização da tecnologia de DWing na área de saúde em termos dos principais conceitos desta tecnologia, enquanto que a seção 4 descreve aspectos relacionados à multidimensionalidade dos dados. A seção 5 apresenta os esquemas multidimensionais propostos. Já a seção 6 destaca aspectos de implementação de um destes esquemas. O artigo é concluído na seção 7.

2. TRABALHOS CORRELATOS

Na literatura, outros trabalhos também aplicam a tecnologia de DWing na área de saúde. Miranda [19] utiliza a tecnologia de DWing com o objetivo de tornar mais eficazes os processos informacionais relacionados à análise dos diagnósticos de saúde realizados anualmente pela Secretaria Municipal de Saúde da cidade de Belo Horizonte, Brasil. Em Miranda, são propostos esquemas multidimensionais para os assuntos nascimento, mortalidade pessoa, mortalidade fetal, notificações, morbidade ambulatorial, produção ambulatorial e internações hospitalares. Murphy *et al*. [21] propõem um conjunto de esquemas projetado para um DW de clínicas médicas pertencentes ao *Partners HealthCare System*. O enfoque destes esquemas é a descoberta de pacientes que possam participar de programas de pesquisa em saúde. Assim, o DW armazena as características clínicas dos pacientes e seus dados demográficos, além de informações sobre diagnósticos, procedimentos médicos e instituições pertencentes ao *Partners HealthCare System*.

Lai & Nicollet [17] descrevem uma iniciativa similar de um DW de clínicas médicas, o qual foi implementado no Oracle 8.0.5. Em [17], são propostos dois esquemas multidimensionais, voltados respectivamente para os assuntos consultas médicas e registros de laboratório. Ambos os esquemas mantêm informações sobre pacientes. Além das tabelas de fatos, foram definidas 27 tabelas de dimensão, das quais 6 tabelas representam informações básicas (e.g., dados demográficos dos pacientes) e 21 tabelas representam informações clínicas (e.g., serviços e testes de laboratório). Outro exemplo de DW aplicado à área de saúde é descrito em [29]. Neste exemplo, o DW de Columbia, Estados Unidos, é utilizado tanto para investigar doenças hereditárias quanto para reconhecer padrões de doenças em situações nas quais o histórico de saúde das famílias não se encontra disponível. O esquema multidimensional proposto em [29] contém não somente o número de registro médico de cada paciente, mas também o primeiro nome, o último nome e o endereço de cada paciente, sendo que estes últimos dados permitem a extração de informações hereditárias.

No entanto, estes trabalhos correlatos apresentam como restrição o fato de não enfocarem a proposta de esquemas multidimensionais com base nos três níveis de informação de uma instituição de saúde [11], como é o caso dos esquemas propostos neste artigo. Outras restrições referem-se ao fato de que estes trabalhos abordam assuntos

predominantemente diferentes quando comparado com os assuntos enfocados neste artigo e, com exceção do trabalho de Miranda, também não atendem aos requisitos de tomada de decisão requeridos por instituições brasileiras de saúde.

3. Tecnologia de Data Warehousing Aplicada à Área de Saúde

Existem diversas motivações para a utilização de um ambiente de DWing na área de saúde. O DW, principal componente do ambiente, armazena dados integrados a partir de provedores operacionais que contêm informações importantes relativas ao negócio. No processo de extração, somente um subconjunto dos dados operacionais necessários à análise estratégica é copiado, sendo requerido neste processo a correção de possíveis inconsistências devido à heterogeneidade dos dados. Como exemplo, devem ser resolvidos problemas de homônimos e de sinônimos, conflitos de chave e de domínio, além da forma de representação dos dados. Na área de saúde, dados valiosos (e.g, de pacientes e doenças) encontram-se dispersos em vários bancos de dados, os quais foram projetados de forma independente e são manipulados pelos mais variados sistemas. Esses dados podem conter erros e, em geral, possuem diferenças semânticas e estruturais. Maiores dificuldades são enfrentadas quando a tomada de decisão gera a necessidade de cruzamento das informações dos diversos sistemas para se obter a informação desejada.

Os dados do DW também são orientados a assunto, ou seja, relativos aos temas de negócio de maior interesse na instituição. Assuntos na área de saúde incluem controle de natalidade, internação hospitalar e tratamento de doenças. Por exemplo, por meio de uma avaliação de diagnósticos de saúde, é possível estudar o comportamento epidemiológico das doenças, fornecendo orientação técnica permanente para o controle das doenças, identificando os problemas prioritários, investigando o fator de hereditariedade e determinando as áreas geográficas ou os grupos populacionais de risco. A abordagem de orientação a assunto é centrada em entidades de alto nível, com ênfase na modelagem de dados, contendo somente dados relevantes ao contexto de tomada de decisão e estruturados, em geral, de forma desnormalizada.

Além de integrados e orientados a assunto, os dados do DW são também históricos [13], ou seja, relevantes a algum período de tempo, em contraste com o ambiente operacional, no qual os dados são válidos somente para o momento de acesso. Para cada mudança relevante no ambiente operacional é criada uma nova entrada no DW, a qual contém um componente de tempo associado implícita ou explicitamente. Na área de saúde, informações valiosas muitas vezes somente são obtidas ao se armazenar dados por vários meses/anos consecutivos. Por exemplo, muitas doenças são sazonais [19], tais como determinados tipos de gripe. A característica temporal dos dados possibilita a identificação de padrões dessas doenças, conforme exemplificado na introdução. Em geral, os dados históricos são relativos a um grande espectro de tempo, como exemplo de 5 a 10 anos. Isto influencia diretamente o volume do DW.

Finalmente, a característica de não-volatilidade está relacionada ao fato de que o conteúdo do DW permanece estável por longos períodos de tempo. O ambiente de DWing é caracterizado pela carga volumosa de dados e pelo acesso a estes dados através de consultas efetuadas por usuários de SSD (ambiente do tipo carregamento de dados e acesso). Assim, apenas dois tipos de operações, ambas caracterizadas por serem de longa duração e complexas, são geralmente efetuadas: (i) transação de manutenção para a carga dos dados, visando a manutenção da consistência do conteúdo do DW com relação ao dados dos provedores; e (ii) consultas dos usuários de SSD do tipo somente para leitura. Em sistemas de DWing comerciais atuais, a transação de manutenção é tipicamente a única transação a atualizar o conteúdo dos dados. Essa transação é realizada durante a "janela noturna", a qual representa o período no qual o DW permanece indisponível para consultas. Ou seja, remoções, inserções e atualizações não são efetuadas durante o período de utilização do DW. Com isto, pode-se efetuar simplificações no gerenciamento dos dados, tais como nos mecanismos de controle de *deadlock* e recuperação de falhas. Em especial, atualizações e remoções nos dados ocorrem somente em caso de carga incorreta, ao passo que inserções são caracterizadas por apenas anexarem dados aos já existentes.

Devido às características dos dados do DW, ambas as consultas e as análises podem ser realizadas rápida e eficientemente pelos usuários de SSD. Na área médica, é comum a necessidade de pesquisa de históricos individuais de pacientes, bem como de estudos de agrupamentos para análises estatísticas [11]. Neste contexto, um ambiente de DWing pode melhorar o desempenho do processamento de consultas OLAP (*on-line analytical processing*), por meio do pré-armazenamento de agregações, reduzindo-se assim o tempo de resposta para tais consultas. Ademais, usuários de SSD podem interagir com o ambiente através de ferramentas dedicadas à análise e à consulta dos dados, as quais oferecem funcionalidades de navegação e de visualização, permitindo que informações relevantes ao contexto de tomada de decisão sejam derivadas a partir de análises de tendências, da monitoração de problemas e de análises comparativas.

Os esquemas multidimensionais propostos neste artigo enfocam os assuntos de interesse de maior importância para três instituições da área de saúde. Cada esquema contém um componente de tempo associado explicitamente. Já a implementação descrita no artigo destaca a utilização de ferramentas da Oracle® Corporation voltadas à integração dos dados e ao oferecimento de funcionalidades de análise e consulta.

4. Característica Multidimensional dos Dados do Data Warehouse

Os dados do DW são usualmente modelados multidimensionalmente, em função das análises efetuadas pelos usuários de SSD, as quais têm por objetivo a visualização dos dados segundo diferentes perspectivas (i.e., dimensões). Uma visão

multidimensional inclui um conjunto de medidas numéricas, que são os objetos de análise relevantes ao negócio, e um conjunto de dimensões, as quais determinam o contexto para a medida [4]. Uma medida numérica pode ser definida como uma função de suas dimensões correspondentes, representando, desta forma, um valor no espaço multidimensional. Como exemplo, a medida numérica contagem de utilização pode ser determinada pelas dimensões data, instalação, status de utilização, médico e setor (e.g., Figura 2). Desta forma, pode-se examinar os dados sobre esta medida numérica segundo diferentes perspectivas, tais como contagem de utilização por data, contagem de utilização por setor, contagem de utilização por instalação por médico, contagem de utilização por médico, por setor, etc. Quanto às dimensões, cada uma delas pode ser descrita por um conjunto de atributos. Exemplos de atributos para a dimensão setor são nome e descrição do setor. Em geral, dimensões possuem uma grande quantidade de atributos descritivos.

Medidas numéricas podem ser classificadas em aditivas, semi-aditivas e não aditivas. Uma medida numérica é aditiva quando pode ser somada por todas as suas dimensões. A medida numérica contagem de utilização da Figura 2 é aditiva, uma vez que por meio da combinação de suas dimensões ela pode ser aritmeticamente somada. Assim, pode-se determinar a contagem de utilização mensal por instalação, por status de utilização, por médico por setor somando-se os valores da contagem de utilização para cada um dos dias que formam o mês. Medidas numéricas semi-aditivas, por outro lado, podem ser somadas somente por intermédio de algumas de suas dimensões, enquanto que para outras dimensões o processo aditivo não tem significado algum. Já medidas numéricas não aditivas simplesmente não podem ser somadas, e devem ser calculadas quando necessário utilizando-se média aritmética ou outro cálculo mais complexo.

Em sistemas relacionais, os dados do DW são armazenados em relações organizadas de forma a refletir a visão multidimensional. Um esquema que tem sido amplamente utilizado para este fim é o esquema estrela [14, 15], o qual possui uma tabela de fatos dominante localizada visualmente no centro da estrela e um conjunto de tabelas de dimensão nas extremidades (e.g., Figura 1). A tabela de fatos armazena as medidas numéricas relevantes ao negócio (i.e., os fatos), além dos valores das dimensões descritivas para cada instância, os quais são responsáveis pela ligação dos fatos às diversas dimensões. A chave primária da tabela de fatos é, portanto, uma combinação das chaves primárias das tabelas de dimensão. Em geral, tabelas de fatos são longas e finas, ou seja, possuem um grande número de tuplas (bilhões de tuplas) e uma quantidade reduzida de colunas (chave primária e medidas numéricas). Por outro lado, cada dimensão é descrita por sua própria tabela, a qual armazena os atributos alfanuméricos da dimensão e possui uma chave primária para cada uma de suas instâncias. Em geral, tabelas de dimensão são curtas e largas, ou seja, possuem um número pequeno de tuplas e uma grande quantidade de colunas (freqüentemente superior a 100 colunas). Em alguns casos, estruturas mais complexas, chamadas de constelações de fatos, podem ser empregadas, nas quais várias tabelas de fato compartilham tabelas de dimensão comuns.

Em ambientes de DWing, a granularidade dos dados armazenados no DW é uma questão de projeto muito importante, uma vez que determina a dimensionalidade do banco de dados e afeta os tipos de consulta que podem ser respondidas. A granularidade refere-se ao nível de detalhamento das informações armazenadas. Por exemplo, para a dimensão tempo, a granularidade pode ser diária, mensal ou anual. Quanto maior o nível de detalhe que se deseja obter do DW, menor a granularidade dos dados. Analogamente, o tamanho do grão é inversamente proporcional ao volume de dados armazenados no DW, ou seja, quanto menor o grão, maior o espaço para armazenar os dados. Quando se tem um nível de granularidade muito pequeno, o tamanho do DW é muito grande, porém praticamente qualquer consulta pode ser respondida. Por outro lado, quando o nível de granularidade é muito alto, ambos o tamanho do DW e o número de consultas que podem ser respondidas são reduzidos.

Neste artigo o esquema estrela é adotado para representar os esquemas multidimensionais propostos, facilitando o entendimento do usuário e a posterior implementação destes esquemas em SGBD relacionais e objeto-relacionais. A granularidade dos dados em cada um dos esquemas é a menor dentro das necessidades de cada assunto, permitindo assim a realização de uma maior gama de consultas.

5. Esquemas Multidimensionais para a Área de Saúde

Segundo Garcia *et al*. [11], um DW deve ser projetado baseando-se nos três níveis de informação de uma instituição de saúde e, assim, atender aos diferentes interesses de análise em cada um dos níveis. Esses três níveis de informação são relativos:

- ao paciente: no qual a disfunção do individuo é colocada em evidência para que lhe seja aplicado um tratamento adequado;
- à comunidade: no qual se podem avaliar por meio de pesquisas casos ocorridos na instituição com o objetivo de melhorar a qualidade dos tratamentos e avaliar os índices de confiabilidade; e
- à instituição: no qual se podem investigar as despesas da instituição visando-se diminuir os gastos.

Esta seção apresenta a proposta de esquemas multidimensionais para três instituições da área de saúde da cidade de Maringá, a saber: Hospital e Maternidade São Marcos (seção 5.1), Hospital Universitário de Maringá (seção 5.2) e Secretaria Municipal da Saúde de Maringá (seção 5.3). Considerações adicionais sobre os esquemas propostos são discutidas na seção 5.4.

Para cada esquema proposto, inicialmente é apresentada a sua descrição e a sua representação no esquema estrela. Na seqüência, para validar o esquema proposto, são listadas as principais consultas de interesse dos usuários de SSD da instituição. Os esquemas propostos neste artigo abrangem os três níveis de informação acima listados. Nos esquemas, somente alguns atributos das dimensões são representados, embora essas dimensões possuam muitos atributos.

5.1. Hospital e Maternidade São Marcos

O Hospital e Maternidade São Marcos é uma entidade particular, composta durante o desenvolvimento deste trabalho por 65 leitos, 6 salas de cirurgia e 5 UTI, além de salas de curativo e serviço de pronto atendimento. Um dos desafios de qualquer administrador é diminuir os custos, porém oferecer serviços de qualidade. Assim, o profissional deve conhecer o funcionamento da instituição e saber onde os recursos estão sendo investidos e quais gastos estão sendo realizados.

Com base nos assuntos de interesse do Hospital, são propostos neste artigo dois esquemas multidimensionais, possibilitando a obtenção de informações sobre os gastos de cada setor e sobre a ocupação das instalações do Hospital. A Figura 1 ilustra o esquema estrela para os custos do Hospital, sendo o grão custo mensal de cada produto por setor. A partir da granularidade escolhida, foram identificadas as tabelas de dimensão Data, Produto e Setor, e os fatos quantidade de produto consumido, custo individual do produto e custo total do produto. O fato quantidade de produto é aditivo em todas as dimensões. Já o custo do produto em Reais é um fato não aditivo, uma vez que o cálculo do custo anual a partir do custo mensal não é realizado somando-se todos os custos correspondentes aos meses que formam um ano. Para este caso, o custo anual do produto em Reais pode ser calculado utilizando-se média aritmética.

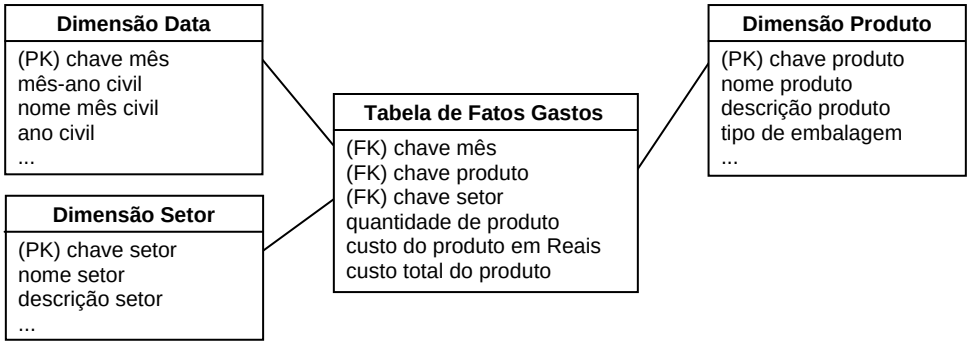

Figura 1. Esquema “Gastos”.

A partir do esquema “Gastos”, pode-se obter informações tais como: (i) Qual a despesa total com o setor de alimentação em 2005? ; (ii) Em qual mês o Hospital mais gastou com o setor de lavanderia? ; (iii) Quais setores apresentaram os maiores custos para o Hospital nos últimos três anos? ; (iv) Quais produtos geraram maior custo em fevereiro de 2006? ; e (v) Qual a quantidade de sabão em pó consumida pelo setor de lavanderia em julho de 2005?

Já a Figura 2 enfoca o esquema multidimensional de ocupação de instalações do Hospital, sendo o grão escolhido a ocupação diária de instalações por setor. Por meio do grão selecionado, foram determinadas as dimensões Data, Instalação (i.e., inclui os leitos), Setor, Status da Utilização e Médico. A dimensão Status da Utilização inclui um descritor textual que determina se a instalação encontra-se “disponível” ou “utilizada”.

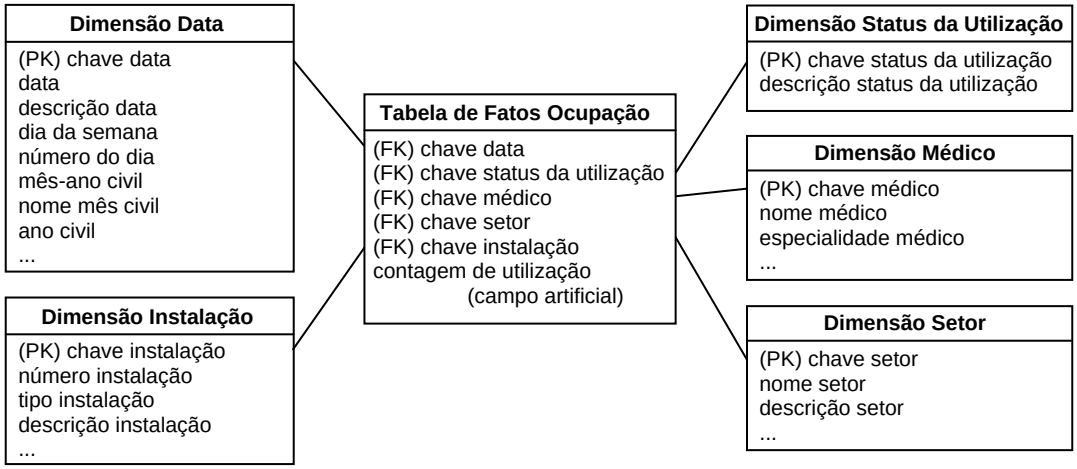

Figura 2. Esquema “Ocupação de Instalações”.

A tabela de fatos “Ocupação” possui um nível pequeno de granularidade. Isto significa que as suas instâncias registram a ocupação (ou não) de uma instalação de um certo setor por um determinado médico em um dia específico. Este tipo de tabela de fatos é chamada de tabela de fatos sem fato (i.e., *factless fact table*). Em geral, um campo artificial é criado para esse tipo de tabela de fatos, o qual é povoado sempre com o valor 1. O uso deste campo facilita a formulação de consultas SQL (*Structured Query Language*) envolvendo soma e agrupamento. Na Figura 2, o fato contagem de utilização representa este campo artificial.

A partir do esquema “Ocupação de Instalações”, pode-se obter informações tais como: (i) Durante quantos dias os leitos da UTI permaneceram ocupados no mês de janeiro de 2005? ; (ii) Em qual setor a não ocupação dos leitos é maior? ; (iii) Qual a taxa de ocupação média dos quartos do hospital por setor? ; e (iv) Quais instalações do setor de pediatria o médico “João Silva” utilizou no mês de fevereiro de 2006?

5.2. Hospital Universitário de Maringá

O Hospital Universitário de Maringá era composto durante o desenvolvimento deste trabalho por 93 leitos, 2 salas de cirurgia e 2 UTI, além de diversos setores, tais como clínicas, farmácia e pronto atendimento. Esta instituição, além de prestar serviço à comunidade de Maringá e de toda região, tem um papel fundamental que é a formação de profissionais da área de saúde.

Pelo fato do Hospital Universitário de Maringá ser uma instituição de ensino, a capacidade de reter e de gerar dados de pacientes torna-se fundamental em uma área em que a informação é uma grande aliada na busca por rapidez e melhores tratamentos. Surge, neste sentido, a necessidade de se analisar diferentes tipos de tratamento para uma determinada doença. A Figura 3 ilustra o esquema projetado para este fim, sendo o grão utilizado doença por paciente.

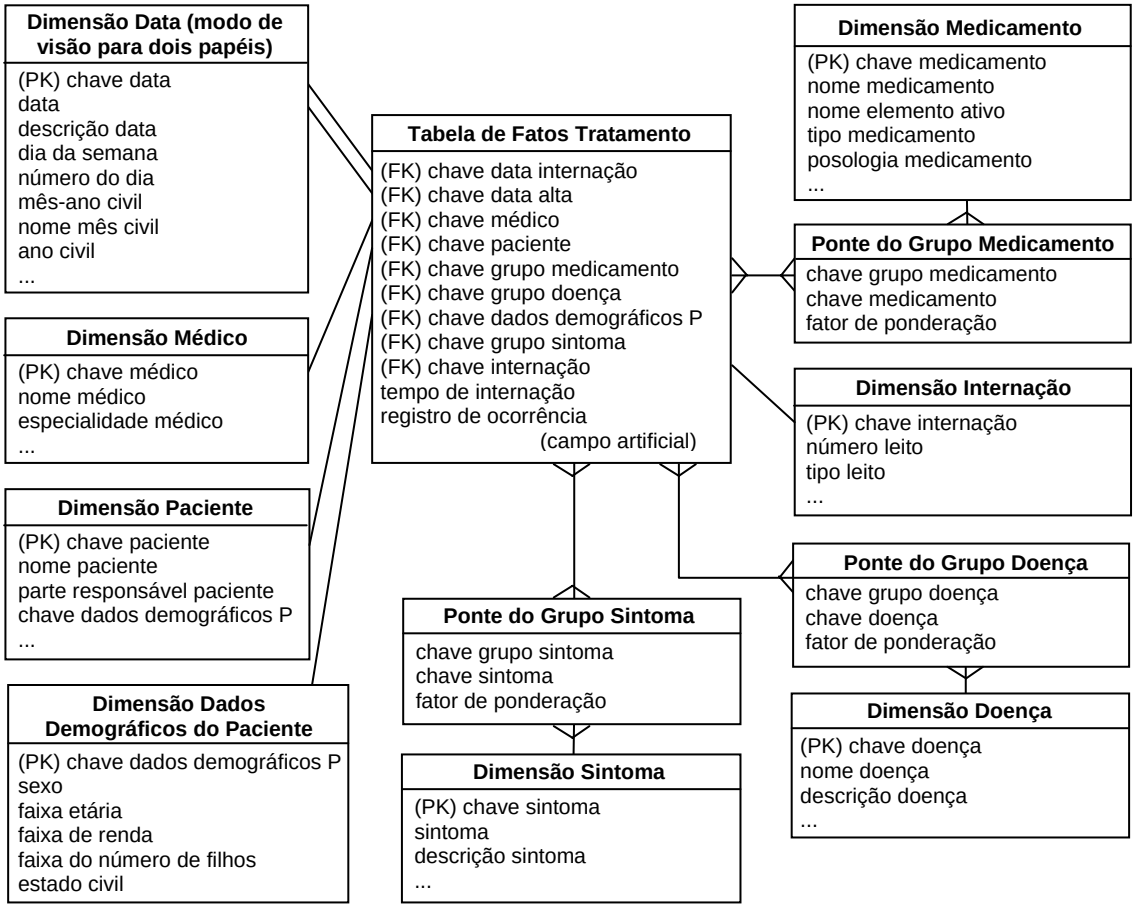

Figura 3. Esquema “Tratamento”.

A dimensão Data possui dois modos de visão: data da internação e data da alta. Cada uma dessas datas é uma chave estrangeira na tabela de fatos. No entanto, no esquema estrela apenas uma tabela de dimensão Data é modelada. Neste caso, são criadas visões da tabela de dimensão Data, que podem ser usadas como se fossem tabelas independentes. Pode-se dizer, portanto, que a tabela de dimensão Data possui visão para dois papéis [14, 15].

Já a tabela de dimensão Dados Demográficos do Paciente é uma minidimensão da dimensão Paciente. A criação desta tabela justifica-se pelo fato de que os seus dados são constantemente analisados ou alterados. Quando se cria uma minidimensão, atributos que variam com freqüência tais como idade e renda, por exemplo, devem ser convertidos em faixas. Normalmente os usuários de SSD não estão interessados em uma idade específica ou em um determinado valor de renda. Estes usuários, por outro lado, realizam suas análises sobre faixas etárias, por exemplo, pacientes de 35 a 50 anos, ou então sobre faixas salariais, tais como renda familiar inferior a R$ 350,00.

Outra característica do esquema da Figura 3 é a existência de tabelas de dimensão multivalor, para as dimensões Doença, Sintoma e Medicamento. Uma tabela de dimensão é multivalor quando é conectada à tabela de fatos por meio de uma outra tabela (i.e., ponte). A ponte funciona como um relacionamento de muitos para muitos. Segundo Song *et al.* [26], o fator de ponderação é um atributo adicional que deve ser modelado sempre que existir uma ponte, pois permite a construção de somatórios corretos.

A partir do esquema "Tratamento", pode-se obter informações tais como: (i) Qual o tempo médio de internamento de pacientes com dengue que foram tratados com um certo conjunto de medicamentos? ; (ii) Quais os sintomas mais comuns da desnutrição, em pacientes com idade inferior a um ano? ; e (iii) Qual o tempo de internação de um paciente com meningite que foi tratado por um determinado médico?

Além da necessidade de se analisar diferentes tipos de tratamento para uma determinada doença, outro aspecto importante a ser considerado é que, por ser uma instituição pública, o Hospital Universitário de Maringá também enfrenta dificuldades relacionadas ao remanejamento e à distribuição de recursos financeiros destinados mensalmente à área de saúde pelo SUS (Sistema Único de Saúde). Muitas vezes esses recursos são insuficientes quando comparados com os custos dos procedimentos.

Na Figura 4, a granularidade dos dados é uma instância para cada procedimento realizado. A partir deste grão, identificam-se as dimensões Data, Procedimento, Paciente, Médico Responsável, Material Utilizado, Médico Assistente e Medicamento. Em especial, as três últimas dimensões são multivalor. Já os fatos modelados são: valor recebido pelo SUS, custo dos medicamentos, custo do médico responsável, custo dos médicos assistentes, custo do material utilizado, custo total e lucro bruto. Destes fatos, apenas o lucro bruto e o custo total são aditivos em todas as dimensões. Os demais fatos referem-se a valores determinados por procedimento.

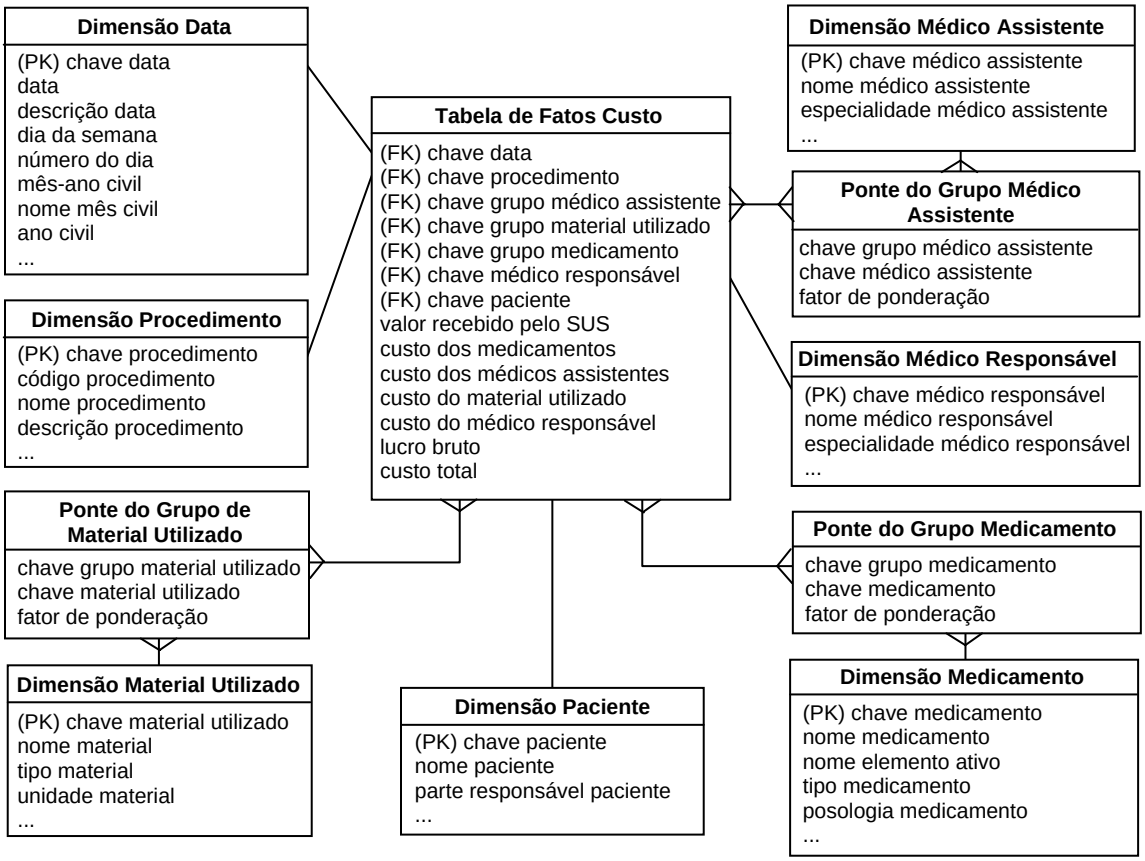

Figura 4. Esquema "Custo dos Procedimentos".

A partir do esquema “Custo dos Procedimentos”, pode-se obter informações tais como: (i) Qual o custo total de um determinado procedimento realizado em um certo paciente? ; (ii) Qual foi a diferença (i.e., lucro bruto) entre o valor recebido pelo SUS e o valor total gasto por procedimento por mês? ; (iii) Qual o procedimento que gerou mais prejuízo no ano passado? ; e (iii) Qual o custo total dos procedimentos realizados por uma certa equipe médica no ano de 2005?

5.3. Secretaria Municipal da Saúde de Maringá

A Secretaria Municipal da Saúde de Maringá tem um papel muito importante no controle de doenças, epidemias e geração de dados estatísticos que são utilizados pelos administradores públicos para processos de tomada de decisão. A Secretaria também coordena diversos programas que propiciam uma melhor qualidade de vida à população.

Apesar das inúmeras funcionalidades desempenhadas pela Secretaria, este artigo enfoca o controle de gestantes. No esquema da Figura 5, os fatos total de gestantes e porcentagem de gestantes são contextualizados pelas dimensões Data, Região, Exame Pré-natal, Dados Demográficos das Gestantes e Dados Estatísticos do Pré-natal. O grão selecionado é uma instância para cada região do município (e.g., zona norte).

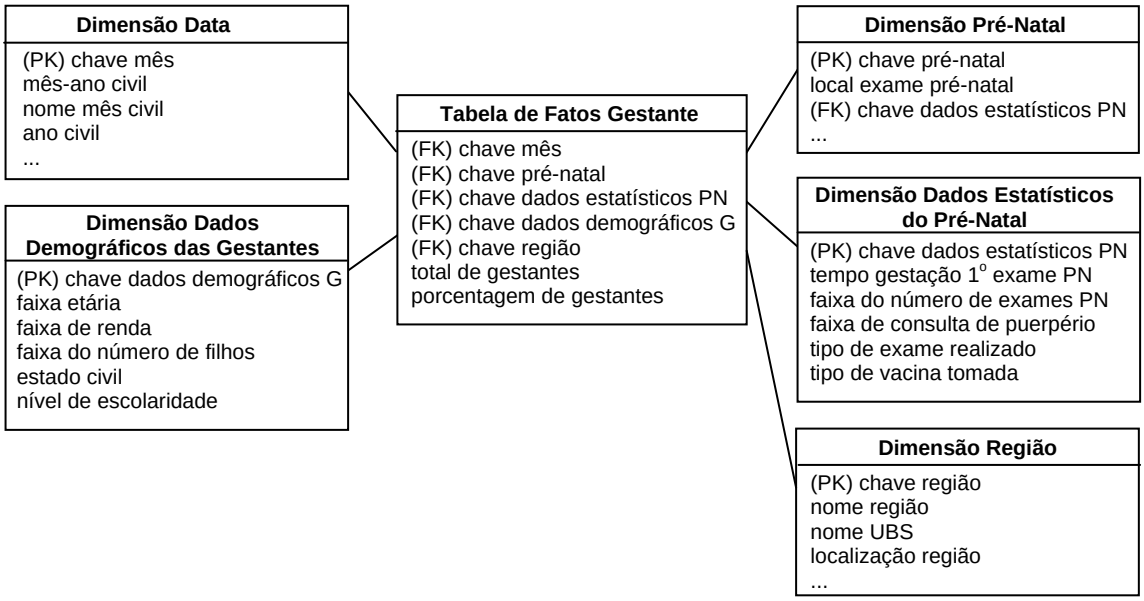

Figura 5. Esquema “Controle de Gestantes”.

A partir do esquema “Controle de Gestantes”, pode-se obter informações tais como: (i) Qual o percentual de gestantes por região que realizaram mais do que seis consultas de pré-natal e que possuem faixa de renda não superior ao salário mínimo? ; (ii) Qual o percentual de gestantes que realizaram de quatro a seis consultas de pré-natal, a consulta de puerpério, todos os exames básicos, o teste anti-HIV e alguma dose da vacina antitetânica? ; e (iii) Qual o total de gestantes por faixa etária que realizaram todos os exames nos meses de janeiro a março de 2006 ?

5.4. Considerações Adicionais

Uma questão importante dos esquemas propostos diz respeito à granularidade utilizada. Muitas vezes existe a necessidade de se realizar um acompanhamento individual. Esta necessidade é modelada pela dimensão Paciente no esquema “Tratamento”. Outras vezes não existe a necessidade de se realizar um acompanhamento tão detalhado. Neste caso, dados particulares de cada indivíduo não constituem o enfoque de interesse, mas sim dados mais gerais referentes a grupos de indivíduos. Esta necessidade é modelada pela dimensão Dados Demográficos das Gestantes no esquema “Controle de Gestantes”. Para este último esquema, em particular, o enfoque de interesse da Secretaria Municipal da Saúde de Maringá é coletar dados que são periodicamente encaminhados para outros órgãos de saúde pública para planejamento de ações de saúde em nível mais macro.

Outra questão refere-se ao compartilhamento de dimensões em comum pelos esquemas em constelações de fatos. Isto permite a realização de consultas *drill-across*, as quais comparam medidas numéricas distintas que são relacionadas entre si por pelo menos uma dimensão em comum. Mesmo que as dimensões comuns difiram entre si pelo nível de granularidade, a realização desta operação ainda é plausível. Desde que os atributos da dimensão de maior granularidade sejam corretamente construídos a partir de agregações da dimensão de menor granularidade, a operação *drill-across* pode ser realizada baseada somente em atributos que existam em ambas as versões das dimensões. Por exemplo, os

esquemas das Figuras 1 e 2 compartilham as dimensões Data e Setor. Assim, a seguinte consulta pode ser realizada: Qual a ocupação e quais os gastos de produtos no setor de pediatria em todos os meses do ano de 2004?

Nos esquemas propostos, também é possível a realização de consultas *drill-down* e *roll-up*. *Drill-down* consiste no processo de analisar os dados em níveis progressivos de detalhamento, ou de menor granularidade. Por outro lado, *roll-up* representa o processo inverso, possibilitando que a análise dos dados seja realizada em níveis progressivamente menos detalhados, ou de maior granularidade. Como exemplo de consultas *roll-up* para o esquema da Figura 5, um usuário de SSD pode iniciar a análise da dimensão região em um baixo nível de granularidade (i.e., total de gestantes por região da cidade) e sucessivamente estender a sua análise utilizando outros atributos desta dimensão (i.e., total de gestantes por cidade, a seguir total de gestante por estado, a seguir total de gestantes por região do Brasil, e a seguir total de gestantes do Brasil).

6. PROTÓTIPO "GASTOS"

Esta seção descreve aspectos de implementação do esquema "Gastos" utilizando o SGBD Oracle9i® e outras ferramentas de DWing da Oracle® Corporation. O protótipo "Gastos" teve como objetivo permitir que os usuários de SSD do Hospital e Maternidade São Marcos vislumbrassem os benefícios e as facilidades da utilização de um DW projetado de forma a atender às suas necessidades de tomada de decisão.

O desenvolvimento do protótipo teve início com a criação das tabelas Dim_Data, Dim_Setor, Dim_Produto e F_CustoSetor, correspondentes respectivamente às dimensões Data, Setor e Produto e à tabela de fatos Gastos da Figura 1. Para cada uma destas tabelas, foram especificadas restrições de chave primária e estrangeira de acordo com a definição de esquema estrela realizada na seção 4. Cada uma das tabelas foi então povoada com dados sintéticos.

Em especial, foi desenvolvido um aplicativo em Delphi6® para facilitar o povoamento das tabelas Dim_Data e F_CustoSetor. Para a tabela Dim_Data, o gerador de dados sintéticos percorreu a tabela, que até o momento possuía apenas seus campos chave mês e mês-ano civil preenchidos, e inseriu automaticamente valores nos demais campos de acordo com o conteúdo do campo mês-ano civil. Já a tabela F_CustoSetor foi preenchida selecionando-se aleatoriamente uma instância de cada uma das três tabelas de dimensão, inserindo-se suas respectivas chaves na tabela de fatos, determinando-se valores aleatórios para o campo quantidade de produto e atribuindo-se valores adequados para o campo custo do produto em Reais e custo total do produto. O preenchimento dos campos das tabelas Dim_Data e F_CustoSetor foi auxiliado por vetores adicionais especialmente criados para este fim. Por exemplo, para armazenar os preços, foi criado um vetor contendo o preço médio de mercado para cada produto inserido na tabela Dim_Produto. Outro exemplo refere-se ao vetor utilizado para armazenar os nomes dos meses do ano. A interface do aplicativo desenvolvido pode ser visualizada na Figura 6.

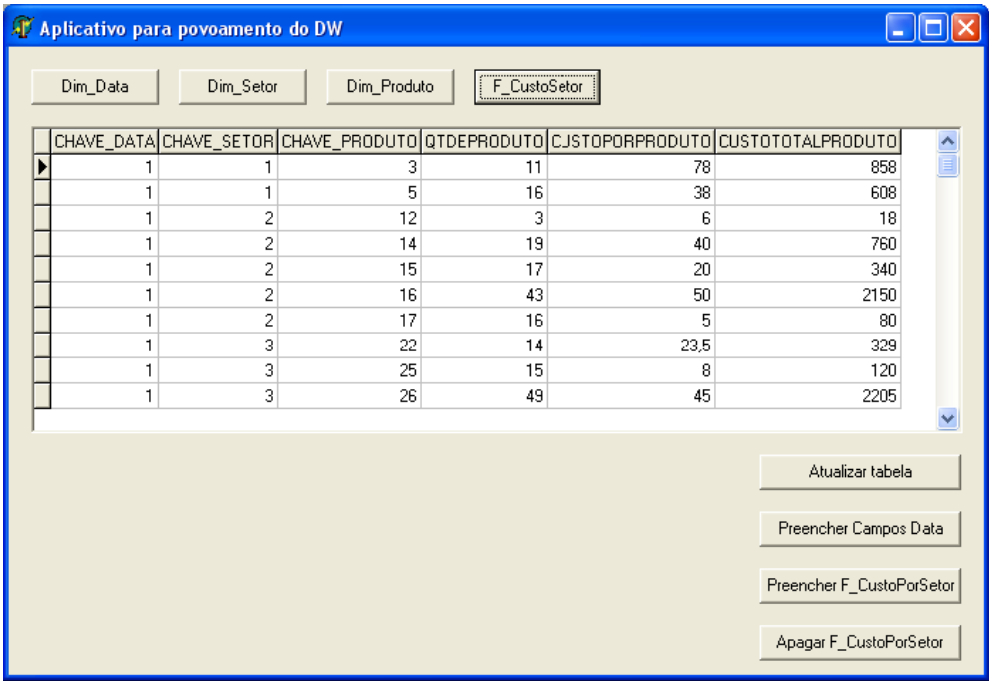

Figura 6. Interface do aplicativo gerador de dados sintéticos.

Utilizando o DW criado e povoado como base, o protótipo “Gastos” explorou os recursos oferecidos pelo SGBD Oracle9i® destinados a oferecer suporte à construção de agregações e à produção de relatórios. Dentre os recursos investigados, pode-se citar os modos de visualização de agregados e de redirecionamento automático de resumos, além do uso das operações de agregação estendidas *rollup* e *cube* e das funções de classificação *rank* e *dense rank* [10, 22, 23, 24, 25].

O modo de visualização de agregados está relacionado à criação de resumos, que são tipos especiais de visões agregadas que reduzem os tempos de resposta para as consultas, pré-calculando junções e operações de agregação complexas e armazenado estes resultados em uma tabela específica no banco de dados. Utilizando como base o esquema “Gastos”, pode-se, por exemplo, criar e armazenar uma tabela que contenha a soma das despesas trimestrais por setor. Já o redirecionamento automático de resumos reescreve automaticamente as consultas dos usuários de SSD definidas sobre as tabelas básicas do DW em novas consultas que são processadas nas tabelas de resumo mais apropriadas. Em particular, a materialização de agregações e a reescrita de consultas tem sido um tópico de pesquisa bastante investigado por parte da área acadêmica [1, 2, 3, 6, 16, 27].

As operações de agregação *rollup* e *cube* estendem a cláusula *group by* da linguagem SQL, por meio da geração de agregações dos dados, ou seja, por meio da criação de subtotais que envolvem desde o nível mais detalhado até um total geral. Por exemplo, a Tabela 1 mostra a diferença entre estas duas operações, considerando uma questão bastante requisitada pelos administradores do Hospital e Maternidade São Marcos: Quais as despesas do Hospital nos anos de 2004 e 2005, separadas pelos diferentes setores? Já com relação às funções de classificação *rank* e *dense rank*, elas têm como objetivo calcular a classificação de uma tupla em relação a outras tuplas presentes no conjunto de dados. Estas funções são extremamente importantes para responder consultas tais como: (i) Quais os 10 medicamentos mais utilizados em cada mês? ; e (ii) Quais os medicamentos que contribuem com os maiores gastos do Hospital?

Uso de rollup		
ano	**setor**	**despesa total**
2004	Limpeza	83.618,0
2004	Lavanderia	92.473,5
2004	Secretaria	74.631,0
2004	Alimentação	135.529,8
2004	Medicamentos	297.800,0
2004		684.052,3
2005	Limpeza	80.571,0
2005	Lavanderia	126.022,5
2005	Secretaria	87.622,5
2005	Alimentação	141.860,8
2005	Medicamentos	663.600,0
2005		1.099.676,8
		1.783.729,1

Uso de cube		
ano	**setor**	**despesa total**
		1.783.729,1
	Limpeza	164.189,0
	Lavanderia	218.496,0
	Secretaria	162.253,5
	Alimentação	277.390,6
	Medicamentos	961.400,0
2004		684.052,3
2004	Limpeza	83.618,0
2004	Lavanderia	92.473,5
2004	Secretaria	74.631,0
2004	Alimentação	135.529,8
2004	Medicamentos	297.800,0
2005		1.099.676,8
2005	Limpeza	80.571,0
2005	Lavanderia	126.022,5
2005	Secretaria	87.622,5
2005	Alimentação	141.860,8
2005	Medicamentos	663.600,0

Tabela 1. Diferença entre as tuplas retornadas pelas operações de agregação *rollup* e *cube*.

O protótipo também explorou a utilização das ferramentas Oracle9i Warehouse Builder® e Oracle9i Discoverer®. A primeira delas oferece como algumas de suas funcionalidades extrair os dados relevantes dos provedores, transformá-los/integrá-los em um formato homogêneo e então carregá-los no DW. Estas funcionalidades são realizadas por meio de uma interface gráfica que permite que os usuários de SSD desenhem figuras representando os fluxos de dados dos provedores em direção ao DW. Estas figuras são então interpretadas e o código referente ao mapeamento dos dados é automaticamente gerado.

Já a segunda ferramenta provê recursos voltados à análise e à exploração dos dados pelos usuários de SSD para a tomada de decisão. Dentre as funcionalidades específicas desta ferramenta pode-se citar o gerenciamento da camada do usuário final, a criação de cadernos de trabalho, a análise dos dados e a geração de gráficos. Além de isolar os usuários de SSD da estrutura física do DW, a camada do usuário final também permite a criação de uma ou mais áreas de negócio direcionadas a grupos de usuários particulares. Cada área de negócio contém somente os dados relacionados a determinados assuntos, de acordo com o interesse de um grupo de usuários. Um caderno de trabalho, por sua vez, contém várias folhas de trabalho mostrando dados voltados a tarefas específicas. Por exemplo, para um caderno de trabalho do esquema “Gastos”, uma de suas folhas de trabalho pode conter uma tabela listando as despesas mensais em cada setor para todos os meses do ano de 2005, enquanto que outra folha de trabalho pode mostrar quais produtos estão gerando as maiores despesas para o Hospital dentro de cada setor.

Como atividades adicionais realizadas no decorrer do trabalho, foram desenvolvidos quatro manuais de utilização das ferramentas exploradas, utilizando como base o esquema "Gastos" implementado no protótipo [18]. Estes manuais demonstram, passo-a-passo, como:

- criar e povoar agregações materializadas;
- mapear (i.e., mover e transformar) os dados de provedores operacionais nos dados a serem armazenados no DW;
- criar e personalizar uma área de negócios; e
- utilizar os recursos relacionados à visualização dos dados desejados do DW e à geração de gráficos e relatórios a partir desses dados, por meio da criação de cadernos de trabalho.

7. CONCLUSÃO

Este artigo descreveu uma experiência do uso da tecnologia de DWing na área de saúde. Desta forma, contribuiu para a área de banco de dados em saúde:

- propondo cinco esquemas multidimensionais voltados às necessidades essenciais de três instituições reais da área de saúde da cidade de Maringá, Brasil, a saber: custo mensal de produtos por setor, ocupação de instalações, tratamento por doença, custo dos procedimentos, e controle de gestantes;
- empregando, nos esquemas propostos, conceitos de modelagem que refletem a natureza complexa de aplicações da área de saúde;
- discutindo questões relacionadas à granularidade dos dados e à realização de consultas *drill-across*, *drill-down* e *roll-up* no contexto dos esquemas propostos; e
- descrevendo diversas atividades desenvolvidas na implementação do protótipo "Gastos", referente ao assunto custo mensal de produtos por setor.

Os esquemas propostos neste artigo são inovadores por tratarem a granularidade dos dados considerando-se outras dimensões além da data, e pelo fato de enfocarem os três níveis de informação de uma instituição de saúde. Ademais, além de atenderem às necessidades de tomada de decisão das instituições para as quais foram projetados, os esquemas propostos também servem de base para a criação de novas aplicações de DWing para a área de saúde, desde que modelam problemas geralmente enfrentados por qualquer instituição desta área. Estes esquemas podem ser adaptados para serem utilizados por exemplo em clínicas médicas, centros hospitalares e secretarias de saúde, dentre outros. Desta forma, os esquemas propostos apresentam grande aplicabilidade prática.

No desenvolvimento deste trabalho, verificou-se que todas as instituições sob análise estavam informatizadas ou em fase de informatização. Entretanto, apesar deste avanço tecnológico, constatou-se que todas elas utilizavam vários sistemas diferentes, que não eram integrados entre si, causando muitas vezes o recadastramento dos dados. Além disto, em geral havia a necessidade de se cruzar dados de diversos relatórios para se obter as informações necessárias. Nesse sentido, pôde-se concluir que o DWing é um ambiente adequado para solucionar os problemas observados.

Em particular, verificou-se que os esquemas multidimensionais propostos atenderam às necessidades de análise dos usuários de SSD dessas instituições. A validação do esquema "Gastos" foi demonstrada a partir de um DW sintético, isto é, a partir de dados gerados artificialmente. O uso de um DW sintético para o esquema "Gastos" permitiu rapidez na avaliação dos resultados e em particular garantiu a correção do esquema de forma interativa com os usuários de SSD. Como resultado, por meio do protótipo os usuários puderam constatar que os dados já preparados (e muitas vezes armazenados) na forma que serão recuperados contribui para diminuir o tempo de resposta de consultas complexas, possibilitando a tomada de decisão médico-analítica mais eficiente.

Como extensão a este trabalho está sendo explorada a fragmentação (vertical e horizontal) dos dados de DW volumosos, que é o caso de aplicações de DWing na área de saúde. A fragmentação dos dados visa não somente melhorar o desempenho do processamento de consultas dessas aplicações, mas também refletir a natureza organizacional distribuída de muitas instituições de saúde [7, 8, 9].

Agradecimentos

Os autores agradecem o apoio financeiro das seguintes agências de fomento à pesquisa do Brasil: CNPq, CAPES, FINEP e FAPESP. Ademais, os autores agradecem as suas respectivas instituições.

Referências Bibliográficas

[1] Albrecht, J., Hümmer, W., Lehner, W. and Schlesinger, L. Query Optimization by Using Derivability in a Data Warehouse Environment. In *Proc. 3rd DOLAP Workshop*. (November, 2000), pp. 49-56.

[2] Baralis, E., Paraboschi, S. and Teniente, E. Materialized View Selection in a Multidimensional Database. In Proc. 23rd VLDB Conference. (August, 1997), pp. 25-29.

[3] Beeri, C., Levy, A.Y. and Rousset, M-C. Rewriting Queries Using Views in Description Logics. In *Proc. 16th ACM PODS*. (May, 1997), pp. 99-108.

[4] Chaudhuri, S. and Dayal, U. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Record*. Vol 26, No. 1, (March, 1997), pp. 65-74.

[5] Ciferri, C.D.A. and Souza, F.F. Data Warehousing: Estado da Arte e Direções Futuras. In *Proc. XXVI CLEI Conference*. (September, 2000), 12 pp.

[6] Ciferri, C.D.A., Souza, F.F. Materialized Views in Data Warehousing Environments. In *Proc. XXI SCCC Conference*. (November, 2001), pp. 3-12.

[7] Ciferri, C.D.A., Ciferri, R.R. and Souza, F.F. Uma Arquitetura de Replicação/Fragmentação para Acesso Distribuído a Data Warehouse via Web. In *Proc. XXV CLEI Conference*. (September, 1999), pp.139-150.

[8] Ciferri, C.D.A., Souza, F.F. Distribuição dos Dados em Ambientes de Data Warehousing. In *Proc. XXVII CLEI Conference*. (September, 2001), 12 pp.

[9] Ciferri, C.D.A. and Souza, F.F. Focusing on Data Distribution in the WebD^{2}W System. *Proc. 4th DaWaK Conference*. (September, 2002), pp. 265-274.

[10] Corey, M., Abbey, M., Abramson, I. and Taub, B. Oracle 8i Data Warehouse. Editora Campus, 2001.

[11] Garcia, A., Sampaio. R., Xéxeo, G., Passos, L.C., Reis, F., Lobo, N., Ximenes, Rabelo, L. and Rabelo Júnior, A. FBCDataWare: Um Data Warehouse para Cardiologia, In *Proc. 4th SADIO Symposium*. (September, 2001), 5 pp.

[12] Golfarelli, M.; Rizzi, S. and Cella, I. Beyond Data Warehousing: What´s Next in Business Intelligence? In: *Proc. 7th DOLAP Workshop*. (November, 2004), pp.1-6.

[13] Inmon, W.H. *Building the Data Warehouse*. John Wiley & Sons, Inc. 1996.

[14] Kimball, R. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. USA: John Wiley & Sons, Inc. 1996.

[15] Kimball, R. and Ross, M. The Data Warehouse Toolkit. Editora Campus, 2002.

[16] Kotidis, Y., Roussopoulos and N. DynaMat: A Dynamic View Management System for Data Warehouses. In Proc. ACM Sigmod Conference. (June, 1999), pp. 371-382.

[17] Lai, S.Y. and Nicollet, P. A Clinical Data Warehouse Effort to Support the Measurable Organizational Patient Care and Health Outcomes Improvement Initiative. In *Proc. AMIA Symposium* (November, 2000).

[18] Lima, A. M. P. Implementação de um Data Warehouse para a Área de Saúde. *Trabalho de Graduação*: UEM, Brasil, (2004).

[19] Miranda, R. M. Utilização do Modelo Dimensional para Diagnóstico de Saúde no Município de Belo Horizonte. *Monografia de Especialização*: Prodabel/IRT-PUC/MG, Brasil, (2000).

[20] Mohania, M., Samtani, S., Roddick, J. and Kambayashi, Y. Advances and Research Directions in Data Warehousing Technology. *The Australian Journal of Information Systems*, 1999.

[21] Murphy, S. N., Morgan, M. M., Barnett, G. O. and Chueh, H. C. Optimizing Healthcare Research Data Warehouse Design through Past COSTAR Query Analysis. In *Proc. AMIA Symposium*. (November, 1999), pp. 892-896.

[22] Oracle9i Discoverer Administrator - Administration Guide - Version 9.0.2. 2002, 704 pp.

[23] Oracle9i Discoverer Desktop - User's Guide - Version 9.0.2 for Windows. 2002, 350 pp.

[24] Oracle Oracle9i - Data Warehousing Guide - Release 2 (9.2). 2002, 666 pp.

[25] Oracle9i Warehouse Builder - User's Guide - Release 9.2. 2003, 1020 pp.

[26] Song, I,-Y.; Rowen, W.; Medsker, C.; Ewen, E. An Analysis of Many-to-Many Relationship between Fact and Dimension Tables in Dimensional Modeling. In *Proc. 3rd DMDW Workshop*. (June, 2001), pp. 6.1-6.13.

[27] Theodoratos, D. and Xu, W. Constructing Search Spaces for Materialized View Selection. In: Proc. *7th DOLAP Workshop*. (Novembro, 2004), pp. 112-121.

[28] Vassiliadis, P. Gulliver in the land of data warehousing: practical experiences and observations of a researcher. In: *Proc. 2nd DMDW Workshop*. (Junho, 2000). 16 pp.

[29] Yu, H. and Hripcsak, G. Hereditary disease discovery from a clinical data warehouse. In *Proc. AMIA Symposium* (November, 2000).