MODELAGEM INCREMENTAL NO AMBIENTE DE DATA WAREHOUSE

VÂNIA JESUS DE ARAUJO SOARES

Universidade Federal do Rio De Janeiro – UFRJ
Instituto de Matemática - IM
Núcleo de Computação Eletrônica - NCE

Tese de Mestrado

Grau: Mestrado em Informática

Orientadora: Maria Luiza Machado Campos
Ph.D. em Ciência da Computação

RIO DE JANEIRO – RJ
DEZEMBRO/1998

MODELAGEM INCREMENTAL NO
AMBIENTE DE DATA WAREHOUSE

VÂNIA JESUS DE ARAUJO SOARES

Dissertação (Mestrado) submetida ao corpo docente do Instituto de Matemática e Núcleo de Computação Eletrônica (IM/NCE) – Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários à obtenção do grau de Mestre**.**

Aprovada por:

_____

Prof. ª Maria Luiza Machado Campos - Orientadora
Ph.D. em Ciência da Computação

_____

Prof. Pedro Manoel da Silveira
Ph.D. em Ciência da Computação

_____

Prof. ª Ana Maria de Carvalho Moura
Dr.Ing. em Ciência da Computação

RIO DE JANEIRO – RJ
DEZEMBRO/1998

Soares, Vânia Jesus de Araujo.
Modelagem Incremental no Ambiente de Data Warehouse
/ Vânia Jesus de Araujo Soares. Rio de Janeiro: UFRJ/IM/NCE, 1998.
xiv, 216p.il
Dissertação(Mestrado) – Universidade Federal do Rio de Janeiro, IM/NCE, 1998.
1. Modelagem de ADW. 2 Diretrizes de Modelagem. I. Título. II. Tese (Mestr. – UFRJ/IM/NCE).

AGRADECIMENTOS

Ao meu amado esposo e à minha mãe
pelo apoio e carinho,
e
aos familiares pela compreensão.

RESUMO

SOARES, Vânia Jesus de Araujo. **Modelagem Incremental no Ambiente de Data Warehouse**.
Orientadora: Maria Luiza Machado Campos. Rio de Janeiro:UFRJ/IM/NCE, 1998. Diss**.**

Esta tese apresenta um conjunto de diretrizes para a modelagem incremental em um ambiente de Data Warehouse a partir de Data Marts, empregando os modelos de dados existentes no ambiente operativo. O conjunto de diretrizes permite a criação de um Data Mart através da derivação dos modelos existentes. Os modelos existentes no ambiente operativo de interesse para o Data Mart são integrados e tratados gerando um pré-modelo que representa o modelo base para a derivação de modelos dimensionais. A criação dos modelos dimensionais é realizada através da construção de uma árvore com base no pré-modelo, sobre a qual serão aplicadas técnicas de poda e enxerto definindo o esboço do modelo dimensional. Este esboço é refinado através do emprego de técnicas dimensionais, gerando o modelo dimensional final. Este modelo é então integrado ao Data Warehouse consolidando um novo ciclo da modelagem do ambiente de Data Warehouse.

ABSTRACT

SOARES, Vânia Jesus of Araújo. **Guidelines for Incremental Modeling on Data Warehouse Environments.**
Orientadora: Maria Luiza Machado Campos. Rio de Janeiro: UFRJ/IM/NCE, 1998. Diss.

This thesis presents a group of guidelines for incremental modeling of Data Warehouse Environments from Data Marts, using the conceptual models of operative environment. The existing models in the operative environment related to the interest area of Data Mart are integrated and treated generating a pre-model that represents the base model for dimensional models derivation. The creation of the dimensional models is accomplished through the construction of a tree based in that pre-model, on which pruning and graft techniques will be applied, defining the sketch of the dimensional model. That sketch is refined employing dimensional techniques, generating the final dimensional model. That model is then integrated into the Data Warehouse, completing the modeling of Data Warehouse Environment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ANEXOS

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

As inovações tecnológicas em hardware e software, aliadas às mudanças nas perspectivas de negócios, levaram as corporações a se preocuparem com o controle de suas informações. Os Sistemas de Suporte à Decisão (SSD) que compõem o Ambiente de Apoio a Decisão (AAD), passaram a ser cada vez mais requisitados, tornando-se peças chaves para análises estratégicas. Entretanto, a proliferação de sistemas nas empresas acarretou um aumento do volume de informações processadas e uma maior distribuição das mesmas, causando uma complexidade em seu gerenciamento para o AAD. A complexidade do gerenciamento de informações resultou em necessidade de mudanças que foram atendidas com a tecnologia de Data Warehousing. Esta tecnologia está centrada na figura do Data Warehouse (DW), uma base de dados capaz de manipular um grande volume de informações com bom desempenho. A abordagem de DW permitiu melhorar a gerência, o controle e o acesso aos dados, e tornou este ambiente conhecido como Ambiente de Data Warehouse (ADW) (INMON,1997). A figura 1.1 apresenta a arquitetura usual do ADW.

Apesar da crescente absorção da tecnologia e do considerável número de referências relacionadas ao assunto, a tecnologia de Data Warehousing pode ser considerada muito recente (GUPTA, 1997). Muitos são os casos de insucesso no projeto de um ADW em decorrência da falta de uma abordagem sistêmica no nível de sua complexidade. Atualmente, observa-se o surgimento de metodologias que tentam estabelecer uma melhor orientação ao desenvolvimento deste ambiente. Entretanto, as abordagens continuam superficiais, dedicando pouco atenção à modelagem de dados, uma das questões mais críticas do ADW.

1.1 Justificativa do Estudo

A literatura especializada é pobre quanto ao processo de modelagem de dados do ADW. Normalmente, são apresentadas técnicas e formas de transformações empregando a modelagem "top-down", considerada a modelagem padrão, que

primeiramente desenvolve o modelo do DW para depois modelar os Data Marts (DM). Entretanto, essa modelagem representa um processo custoso e pouco prático, não apresentando, normalmente, bons resultados, em virtude do tempo gasto na sua implementação. As empresas, em sua grande maioria, vêm adotando o desenvolvimento de ADW empregando uma abordagem incremental, também conhecida como "bottom-up". Essa abordagem permite justificar o custo do desenvolvimento deste ambiente pela rápida apresentação de resultados, atraindo a atenção e investimento por parte dos usuários finais.

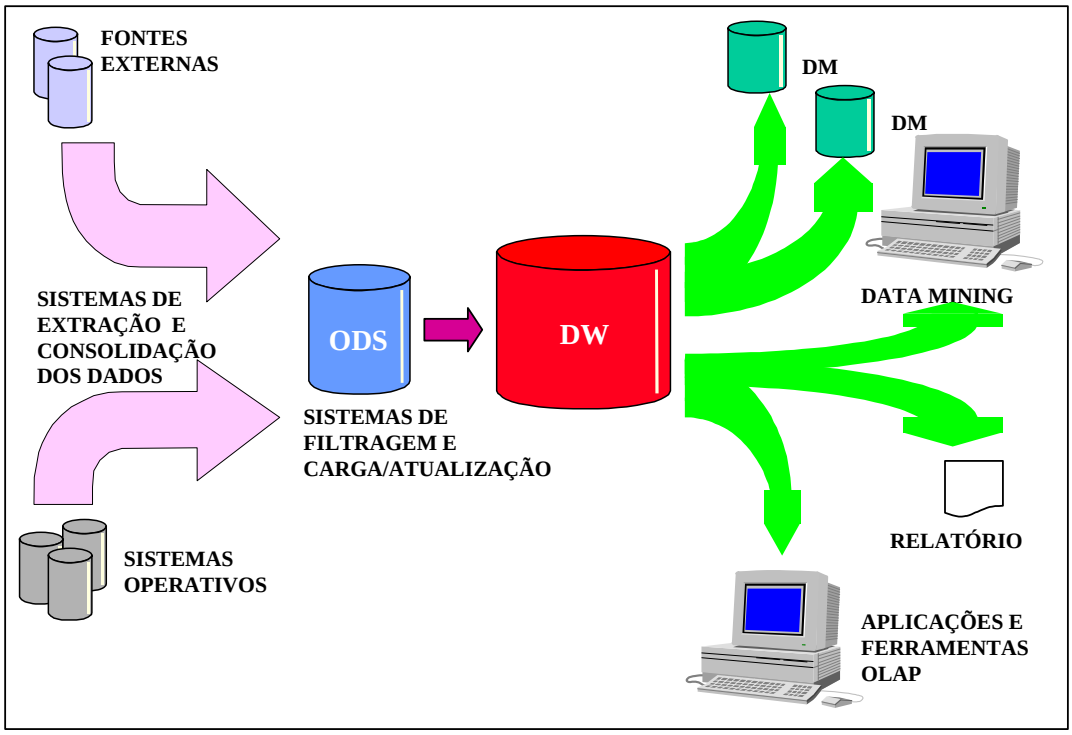

Figura 1.1 – Arquitetura de um ADW

De modo geral, é possível observar uma carência de diretrizes que tratem da modelagem de dados do ADW. Esta modelagem abrange a modelagem de dados do DW e de seus DM. Essa carência leva a uma modelagem de dados empírica sob grande influência das decisões do projetista.

1.2 Objetivo do Estudo

O propósito deste estudo é estabelecer um conjunto de diretrizes que permita realizar a modelagem de dados de um ADW de forma incremental, a partir dos modelos

de dados existentes no ambiente operativo. Segundo estas diretrizes, a modelagem do ADW será realizada a partir da criação de DM, por um processo de derivação dos modelos de dados do ambiente operativo, e sua posterior integração ao DW. Para tornar a seqüência das diretrizes mais clara a modelagem do ambiente foi dividida em quatro fases. A primeira fase é responsável pela análise dos modelos de dados existentes no ambiente operativo. Na segunda fase é realizada uma transformação do modelo de dados, até então orientado a processo, em um modelo de dados mais próximo ao negócio. Na terceira fase é realizada a derivação de um ou mais modelos dimensionais para o DM, a partir do modelo de dados resultante da fase anterior. A quarta fase é a responsável por integrar o modelo de dados do DM ao modelo de dados do DW, finalizando dessa forma a modelagem do ADW.

Apesar da abordagem empregada para o desenvolvimento do ADW ser "bottom-up", o seu processo de alimentação (carga e atualização) funciona segundo os padrões normais do ambiente. Dessa forma, os dados são extraídos das fontes, sistemas operativos e fontes externas, e carregados para o DW, após um processo de limpeza e transformação de dados. A seguir, os dados são disponibilizados para os DM.

1.3 Organização da Tese

Esta dissertação encontra-se organizada em sete capítulos.

O capítulo 2 apresenta as características, os componentes e as arquiteturas do ADW, incluindo as metodologias empregadas para o seu desenvolvimento.

No capítulo 3 são apresentados importantes conceitos, técnicas empregadas e uma série de questões relacionadas à modelagem de dados do ADW.

O capítulo 4 estabelece as diretrizes propostas para o desenvolvimento incremental de um ADW.

O capítulo 5 apresenta a aplicação das diretrizes em um estudo de caso no ambiente de registro acadêmico e resultados do vestibular. Este estudo consiste no desenvolvimento do ADW da universidade, através da modelagem de dois DM: o DM Graduação e o DM Vestibular e a integração dos mesmos ao DW Universidade.

No capítulo 6 as técnicas e os recursos empregados pelo conjunto de diretrizes são apresentados e é realizada uma análise das diretrizes propostas.

Finalmente, o capítulo 7 apresenta uma conclusão geral do trabalho, com sua contribuição e sugestões de trabalhos futuros.

Também fazem parte deste trabalho:

Anexo 1 – Histórico da Evolução do AAD;

Anexo 2 – Peculiaridades do Modelo Físico do ADW;

Anexo 3 – Árvores;

Anexo 4 – Apoio ao Estudo de Caso Universidade; e

Anexo 5 – Dicionário de Dados referente ao Estudo de Caso Universidade.

CAPÍTULO 2

AMBIENTE DE DATA WAREHOUSE

A tecnologia de Data Warehousing é considerada a evolução natural do Ambiente de Apoio à Decisão (AAD) (WU, BUCHMANN, 1997). Sua rápida absorção pelas empresas está relacionada à necessidade do domínio de informações para garantir respostas e ações rápidas, assegurando a competitividade de mercado (GUPTA, 1997). Dentre os fatores que contribuíram para essa absorção, merecem destaque: os avanços tecnológicos, as mudanças organizacionais e estruturais nos negócios, a abertura de mercados e a globalização da economia. Com o advento da tecnologia de Data Warehousing, os AAD passaram a ser denominados Ambientes de Data Warehouse (ADW). Este novo ambiente contém como repositórios principais o Data Warehouse (DW) e os Data Marts (DM).

O propósito deste capítulo é apresentar características, arquiteturas e metodologias de desenvolvimento. Dentre estes aspectos, um maior nível de detalhe será dado às metodologias empregadas para o desenvolvimento do ADW. Por ser considerada uma tecnologia recente, está aberta a propostas que venham a aprimorar o seu emprego.

2.1 Tecnologia de Data Warehousing

A tecnologia de Data Warehousing tem o propósito de capacitar as empresas a acessarem seus dados de forma rápida e fácil, proporcionando um apoio aos tomadores de decisão.

Nesse contexto, o Data Warehousing proporciona ao AAD uma sólida e concisa integração dos dados da empresa, para a realização de análises gerenciais (ADELMAN,1992). Ele se preocupa em integrar e consolidar as informações de fontes heterogêneas e fontes externas, sumarizando, filtrando e limpando estes dados, preparando-os para análise e suporte à decisão.

Esta tecnologia originou o ADW, que possui um conjunto de características, conforme apresentado a seguir, que o distingue de outros ambientes de sistemas

convencionais (POE, KLAUER,BROBST, 1998):

- Extração de dados de fontes heterogêneas (existentes ou externas);
- Transformação e integração dos dados antes de sua carga;
- Normalmente requer máquina e suporte próprio;
- Visualização dos dados em diferentes níveis. Os dados do DW podem ou não ser extraídos para um nível mais específico, os DM, e a partir deste para um banco de dados individual; e
- Utilização de ferramentas voltadas para acesso com diferentes níveis de apresentação.

Um histórico da evolução do ADW, desde o AAD até os dias atuais, encontra-se no anexo 1 .

2.2 Arquitetura do Ambiente

A arquitetura do ADW inclui, além de estrutura de dados, mecanismos de comunicação, processamento e apresentação da informação para o usuário final. A figura 2.1 apresenta a arquitetura padrão deste ambiente. De uma forma geral, as arquiteturas orientadas a este ambiente são constituídas por um conjunto de ferramentas que respondem desde a carga até ao processamento de consultas, assim como por repositórios de dados , como o DW e DM. As ferramentas existentes podem ser divididas em dois grupos: as relacionadas à carga inicial e às atualizações periódicas do DW , que são responsáveis pela extração dos dados de múltiplos sistemas operativos e fontes externas, e pela limpeza, transformação e integração dados; e as relacionadas às consultas, orientadas para o usuário final, que são responsáveis pela elaboração de relatórios, pesquisas informativas, análise e "data mining".

Quanto aos repositórios, o DW funciona como um grande centralizador dos dados, enquanto os DM permitem um visão mais direcionada de um problema, funcionando como repositórios menores, orientados a áreas específicas. A seguir são apresentados os principais componentes e tipos de arquitetura para este ambiente.

2.2.1 Componentes

Com uma visão mais abrangente, é possível analisar os componentes do ADW

com relação aos seguintes aspectos: papéis, processos/ferramentas associadas e dados.

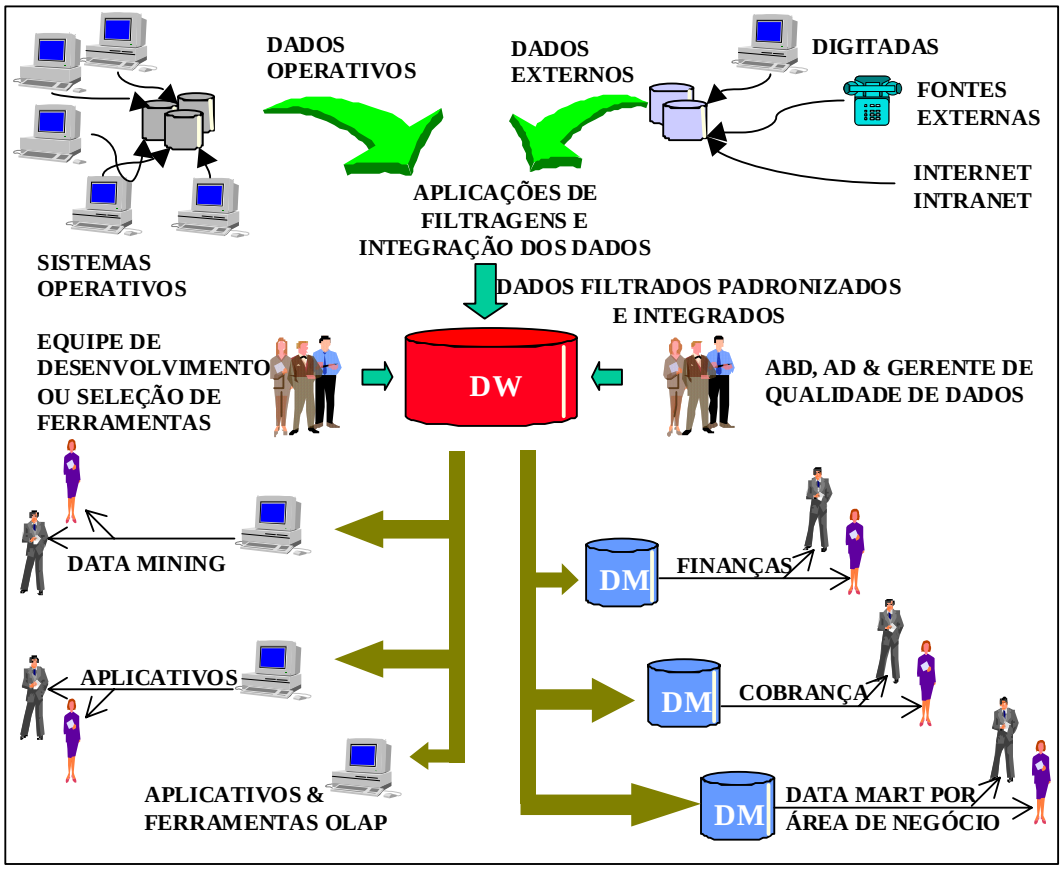

Figura 2.1 – Arquitetura Padrão do ADW

a) Papéis

Por sua abrangência, este ambiente envolve desde profissionais de processamento de dados até analistas de negócios que podem ser considerados como "usuários" do ambiente. Este ambiente inclui: os administradores do projeto; os projetistas do banco; os Administradores de Bancos de Dados (ABD) dos sistemas operativos; os programadores e os analistas dos sistemas de conversão e dos aplicativos que acessarão as informações; e os usuários finais. Esta diversidade de "usuários" implica em uma preocupação maior do que a observada nos sistemas operativos. Nesses, os analistas e projetistas dos sistemas apenas atendiam ao grupo de usuários que efetivamente utilizariam o sistema. No ADW estes "usuários" encontram-se agrupados por papéis da seguinte forma:

- Responsáveis pela carga dos dados: representam os programadores que necessitam conhecer o mapeamento entre o DW e os sistemas operativos, além de todos os requisitos necessários à filtragem e à integração dos dados.
- Usuários finais: são os especialistas, gerentes, executivos e analistas de negócio, que utilizam a informação para apoio à tomada de decisão. Estes usuários apresentam uma grande familiaridade com os termos do negócio e estão sempre em busca da solução de um problema ou de novas oportunidades (DEVLIN,1997). Estes usuários podem ser divididos em dois grupos: os usuários diretos e os usuários indiretos. Os usuários diretos são aqueles que acessam livremente o DW enquanto os indiretos acessam os DM especializados (INMON, HACKATHORN, 1997).
- Responsáveis pelo desenvolvimento e manutenção do DW e dos DM: eqüivalem aos Administradores de Bancos de Dados (ABD) e Administradores de dados (AD) dos Sistemas Gerenciadores de Bancos de Dados (SGBD) dos sistemas operativos. Estabelecem o nível de preocupação com os metadados, com a arquitetura de armazenamento e com a estrutura dos dados, visando, principalmente, melhorar o desempenho das consultas. É comum o estabelecimento de equipes diferentes. Para estes casos, um maior grau de rigidez se faz necessário na elaboração dos DM, evitando inconsistências entre as estruturas dos DM e do DW.

b) Processos e Ferramentas Associadas

Os processos do ADW consistem na extração dos dados das fontes operativas, na organização e integração destes dados de forma consistente para o DW, e no acesso aos dados integrados de modo eficiente e flexível, tanto para consulta direta, como para acesso por DM (GOLFARELLI, MAIO, RIZZI, 1998).

A extração, organização e integração dos dados devem ser realizadas com o propósito de garantir a consistência e integridade das informações. Normalmente, faz-se necessário o desenvolvimento de sistemas ou avaliação de ferramentas para extração de dados e atualização do DW. Estes sistemas/aplicações são responsáveis pela filtragem, limpeza, sumarização e concentração dos dados espalhados pelas fontes externas e nos sistemas operativos. Sua elaboração requer, dos analistas envolvidos, um razoável conhecimento tanto das bases de dados de onde as informações serão extraídas como da base onde serão armazenadas. Muitas propostas vêm sendo desenvolvidas nesta área,

com a finalidade de acelerar o processo de carga e atualização do DW (RADEN, 1996), (MEYER, CANNON, 1998).

No ADW, as ferramentas devem permitir um acesso intuitivo aos dados, possibilitando a análise daqueles mais significativos. O sucesso de um DW pode depender da disponibilidade da ferramenta certa para as necessidades de seus usuários. Para garantir esta flexibilidade, normalmente são empregados (CAMPOS, FILHO, 1997):

- Ferramentas para pesquisa e relatórios: ferramentas simples que oferecem uma interface gráfica para a geração de relatórios e análise de dados históricos. Também permitem, ao usuário, avaliar "o que aconteceu".
- Ferramentas do tipo "On-line Analytical Processing" - OLAP: estas ferramentas permitem ao usuário analisar o porquê dos resultados obtidos. Atualmente existem disponíveis no mercado uma variedade destas ferramentas com diferentes abordagens:
  - ROLAP (OLAP Relacional) : ferramentas OLAP que acessam bancos de dados relacionais;
  - MOLAP (OLAP Multidimensional) : ferramentas OLAP que acessam bancos de dados multidimensionais, através de cubos e hipercubos;
  - HOLAP (OLAP Híbrida) : ferramentas OLAP que permitem acesso tanto aos bancos de dados relacionais como aos multidimensionais; e
  - DOLAP (OLAP Desktop) : ferramentas OLAP voltadas para computadores pessoais. Este tipo de ferramenta vem sendo empregado nos bancos de dados individuais, para análises mais específicas do que as realizadas no DM. Os dados, normalmente, são carregados a partir de DM.
- Sistemas de informações executivas: apresentam uma visualização dos dados mais simplificada. As informações são apresentadas de forma consolidada, não requerendo do usuário experiência e tempo para executar uma análise, como é o caso das ferramentas OLAP.
- "Data Mining": é uma categoria de ferramentas de análise denominada "Open-End". Permitem ao usuário avaliar tendências e padrões entre os dados. Este tipo de ferramenta se utiliza das mais modernas técnicas de computação, como redes neurais, algoritmos genéticos e lógica nebulosa.

c) Dados

Neste ambiente os dados podem ser armazenados em diferentes níveis de agregação, como: dados detalhados, configurando o nível operativo; dados levemente sumarizados; e dados altamente sumarizados (INMON, 1997). Os dados encontram-se em repositórios que constituem uma das maiores preocupações deste ambiente. O emprego ou não de qualquer dos repositórios apresentados a seguir depende, exclusivamente, da arquitetura a ser adotada pela empresa. O ADW pode apresentar os seguintes repositórios de dados:

- "Operational Data Storage" – ODS: representa um armazenamento intermediário dos dados, facilitando a integração dos dados do ambiente operativo antes da sua atualização no DW. Em sua proposta original, o ODS era um repositório temporário, que armazenava apenas as informações correntes, antes de serem carregadas para o DW. Atualmente, alguns autores passaram a denominá-lo Armazenamento Dinâmico de Dados – "Dynamic Data Storage – DDS". Esta nova concepção difere da original quanto à periodicidade de armazenamento. Ao contrário do ODS original, ele não armazena dados apenas para a carga do DW. O DDS não é volátil, e seus dados são armazenados ao longo do tempo. O DDS sofre alterações incrementais e com o decorrer do tempo, pode se tornar o DW.

  O ODS pode servir de base para análises do ambiente operativo, pois sua granularidade é normalmente compatível com os sistemas deste ambiente. Sua função não é sumarizar dados, mas agilizar o processo de consolidação, proporcionando um melhor desempenho na fase da atualização dos dados. O ODS não é um componente indispensável em um ADW, sendo a sua criação uma decisão

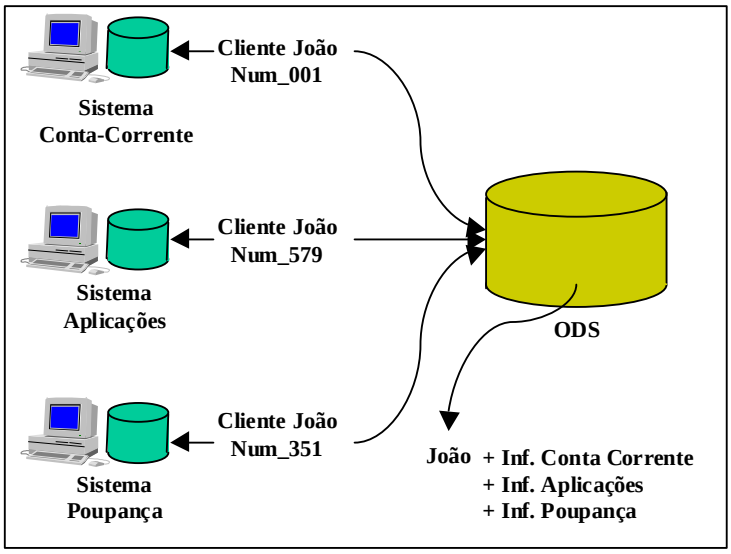

Figura 2.2 – Aplicação de ODS

de projeto. Por combinar a tecnologia de DW com os sistemas operativos tradicionais (OLTP), permite análises e apoio à tomada de decisões que requeira respostas em tempo real. As empresas que optam por sua utilização normalmente empregam informações similares em diversos sistemas operativos. Esta dispersão de informações requer um primeiro tratamento, que consiste na consolidação dos dados, antes de sua integração no DW.

A figura 2.2 apresenta um exemplo de aplicação de ODS em sistemas bancários. No exemplo, observa-se a existência do cliente João em três sistemas diferentes. As informações do cliente João referentes aos três sistemas são integradas no ODS antes de serem transportadas para o DW.

- DATA WAREHOUSE (DW): É a espinha dorsal deste ambiente. Ele representa uma grande base de dados capaz de integrar, de forma concisa e confiável, as informações de interesse para a empresa, que se encontram espalhadas pelos sistemas operativos e em fontes externas, para posterior utilização nos SSD.
- DATA MART (DM): Representa um subconjunto de dados do DW. Permite acesso descentralizado e atualmente serve de fonte para os dados que comporão os bancos de dados individuais. Os dados do DM são direcionados a um departamento ou uma área específica do negócio. O DM, normalmente, é modelado em um esquema estrela, de acordo com as necessidades específicas do usuário final (COREY, 1996). Uma das principais vantagens de seu emprego é a possibilidade de retorno rápido, garantindo um maior envolvimento do usuário final, capaz de avaliar os benefícios extraídos de seu investimento.
- BD INDIVIDUAIS: Estes bancos permitem ao usuário armazenar, em caracter temporário, apenas os dados de seu interesse, reduzindo o escopo da informação e acelerando seu processamento. Normalmente representam um subconjunto do DM. Esta modalidade vem merecendo destaque graças ao desenvolvimento de ferramentas OLAP para desktop (DOLAP).

2.2.2 Tipos de Arquitetura

De uma forma geral, a arquitetura do ADW ainda está em evolução. Esta evolução pode ser considerada como uma resposta à crescente complexidade deste

ambiente e às dificuldades de integração entre todos os componentes. Os desenvolvedores deste ambiente devem se preocupar em como integrar o DW às diversas fontes heterogêneas e externas, aos DM, ODS, aplicações servidoras, "WEB" e "data mining", entre outros tipos de ferramentas disponíveis (FIRESTONE, 1998-b).

As arquiteturas "top-down", "bottom-up" e intermediária são propostas para o desenvolvimento deste ambiente. Variações destas arquiteturas estão sendo avaliadas, sendo que uma arquitetura não inviabiliza a outra. Entretanto, a variedade de opções requer uma análise mais apurada do problema, para se avaliar qual é a arquitetura mais adequada à empresa. A escolha da arquitetura é fator importante na seleção da tecnologia apropriada para o desenvolvimento e a implantação deste ambiente. Atualmente, considera-se que os problemas do ADW estão mais relacionados com a arquitetura do que com a tecnologia disponível (MELLO, 1997).

2.2.2.1 Arquitetura "Top_Down"

A arquitetura "Top-Down" é a arquitetura conhecida como arquitetura padrão (INMON, HACKATHORN, 1997). Nesta arquitetura o processo se inicia com a extração, a transformação e com a integração das informações dos sistemas operativos e dados externos para um ODS. A seguir, os dados e metadados são transferidos para o DW.

Figura 2.3 – Arquitetura Padrão, empregando ODS

Nesta concepção, o DW armazena uma camada de dados atômicos e detalhes históricos. A partir do DW são extraídos os dados e metadados para os DM. Nos DM, as

informações estão em um maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no DW (INMON, 1998). A figura 2.3 apresenta a arquitetura padrão empregando o ODS. A seguir são apresentadas as vantagens e desvantagens desta arquitetura segundo Hackney (HACKNEY, 1998):

VANTAGENS**:**

- Herança de arquitetura - Todos os DM originados a partir de um DW, utilizam a arquitetura e os dados deste DW, permitindo uma fácil manutenção;
- Visão de empreendimento - O DW concentra todos os negócios da empresa, sendo possível a partir dele extrair níveis menores de informações;
- Repositório de metadados centralizado e simples - O DW provê um repositório de metadados central para o sistema. Esta centralização permite manutenções mais simples do que aquelas realizadas em múltiplos repositórios; e
- Controle e centralização de regras - A arquitetura "top-down" garante a existência de um único conjunto de aplicações para extração, limpeza e integração dos dados, além de processos centralizados de manutenção e monitoração.

DESVANTAGENS:

- Implementação muito longa - Os DW são, normalmente, desenvolvidos de modo iterativo, por áreas de assuntos, como por exemplo, vendas, finanças e recursos humanos. Mesmo assim, são necessários, em média, 15 ou mais meses para que a primeira área de assunto entre em produção, dificultando a garantia de apoio político e orçamentário;
- Alta taxa de risco - Não existem garantias para o investimento neste tipo de ambiente;
- Heranças de cruzamentos funcionais. É necessário uma equipe de desenvolvedores e usuários finais, altamente capacitados para avaliar as informações e consultas que garantam à empresa habilidade para sobreviver e prosperar na arena de mudanças de competições políticas, geográficas e organizacionais; e
- Expectativas Relacionadas ao Ambiente - A demora do projeto e a falta de retorno pode induzir expectativas nos usuários.

2.2.2.2 Arquitetura "Bottom-Up"

Em virtude da arquitetura "top-down" ser politicamente difícil de ser definida e muito cara, requerendo um tempo grande para implementação, investimento e sem apresentar retorno rápido, a arquitetura "bottom-up" vem se tornando muito popular. A figura 2.4 apresenta a arquitetura "bottom-up".

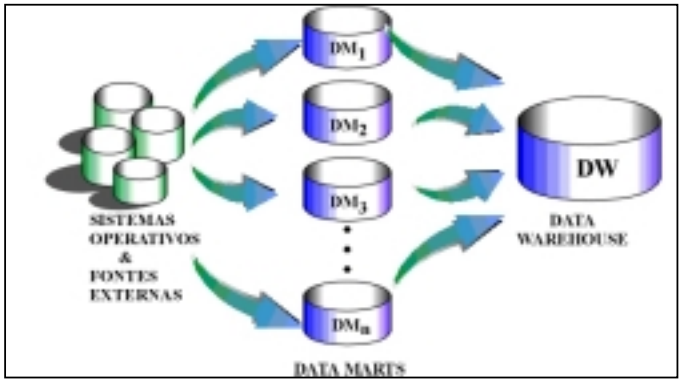

Figura 2.4 – Arquitetura "Bottom-Up"

O propósito desta arquitetura é a construção de um DW incremental a partir do desenvolvimento de DM independentes. O processo começa com a extração, a transformação e a integração dos dados para um ou mais DM, sendo estes DM modelados, normalmente, através de um modelo dimensional. Um dos grandes problemas desta arquitetura é a falta de um gerenciador que garanta padrões únicos de metadados, mesmo com a independência dos DM. A dificuldade em se garantir essa padronização é responsável pela falha na elaboração incremental do DW. A seguir são apresentadas as vantagens e desvantagens desta arquitetura segundo Hackney (HACKNEY, 1998):

VANTAGENS:

- Implementação rápida - A construção dos DM é altamente direcionada, permitindo um rápido desenvolvimento. Normalmente, um DM pode ser colocado em produção em um período de seis a nove meses;
- Retorno Rápido - A arquitetura baseada em DM com incremento demonstra rapidamente seu valor, permitindo uma base para investimentos adicionais, com um nível mais elevado de confiança;
- Manutenção do Enfoque da Equipe - Um dos maiores desafios do desenvolvimento de um ADW é a manutenção do mesmo enfoque por toda a equipe. A elaboração de DM incrementais, permite que os principais negócios sejam enfocados inicialmente, sem que haja gastos no desenvolvimento de áreas que não são essenciais ao problema; e

- Herança Incremental - A estratégia de DM incrementais obriga a entrega de recursos de informação, passo a passo. Isto permite a equipe crescer e aprender, reduzindo os risco. A avaliação de ferramentas, tecnologias, consultores e vendedores só deve ser realizado uma vez, a não ser que existam restrições que impeçam o reaproveitamento.

DESVANTAGENS:

- Perigo de LegaMarts - Um dos maiores perigos no ADW é a criação de DM independentes. O advento de ferramentas de "drag-and-drop" facilitou o desenvolvimento de soluções individuais, de acordo com as necessidades da empresa. Estas soluções podem não considerar a arquitetura de forma global. Desta forma, os DM independentes transformam-se em DM legados, ou LegaMarts. Os LegaMarts dificultam, quando não inviabilizam futuras integrações. Eles são parte do problema e não da solução;
- Desafio de Possuir a Visão de Empreendimento - Durante a construção dos DM incrementais é necessário que se mantenha um rígido controle do negócio como um todo. Este controle requer um maior trabalho ao extrair e combinar as fontes individuais do que utilizar um DW;
- Administrar e Coordenar Múltiplas Equipes e Iniciativas - Normalmente, esse tipo de arquitetura emprega o desenvolvimento de DM em paralelo. Isto pode conduzir a uma rígida administração tentando coordenar os esforços e recursos das múltiplas equipes, especialmente nas áreas de regras e semântica empresariais; e
- A Maldição de Sucesso - A arquitetura com DM incrementais carrega a " maldição de sucesso". Nestes casos, os usuários finais do DM encontram-se felizes querendo mais informação para seus DM. Ao mesmo tempo, outros usuários de outros DM aguardam o incremento de seus DM. Isto conduz a equipe de DM a vencer desafios políticos, de recurso e de administração.

Muitas das novas abordagens propostas, baseiam-se na arquitetura "bottom-up". Elas procuram otimizar o processo de desenvolvimento e garantir a consistência dos metadados e facilidade de integração do ambiente.

A arquitetura de DM para empresa ("Enterprise Data Mart Architecture" - EDMA) e a arquitetura de armazenamento de dados/ DM ("Data Storage/Data Mart" -

DS/DM) são bons exemplos dessas novas abordagens (FIRESTONE, 1998-b). A seguir essas arquiteturas são apresentadas em maiores detalhes.

a) EDMA: A EDMA é uma evolução da arquitetura "bottom-up", que emprega a abordagem incremental para o DW pelo desenvolvimento de DM, utilizando um "framework" compartilhado.

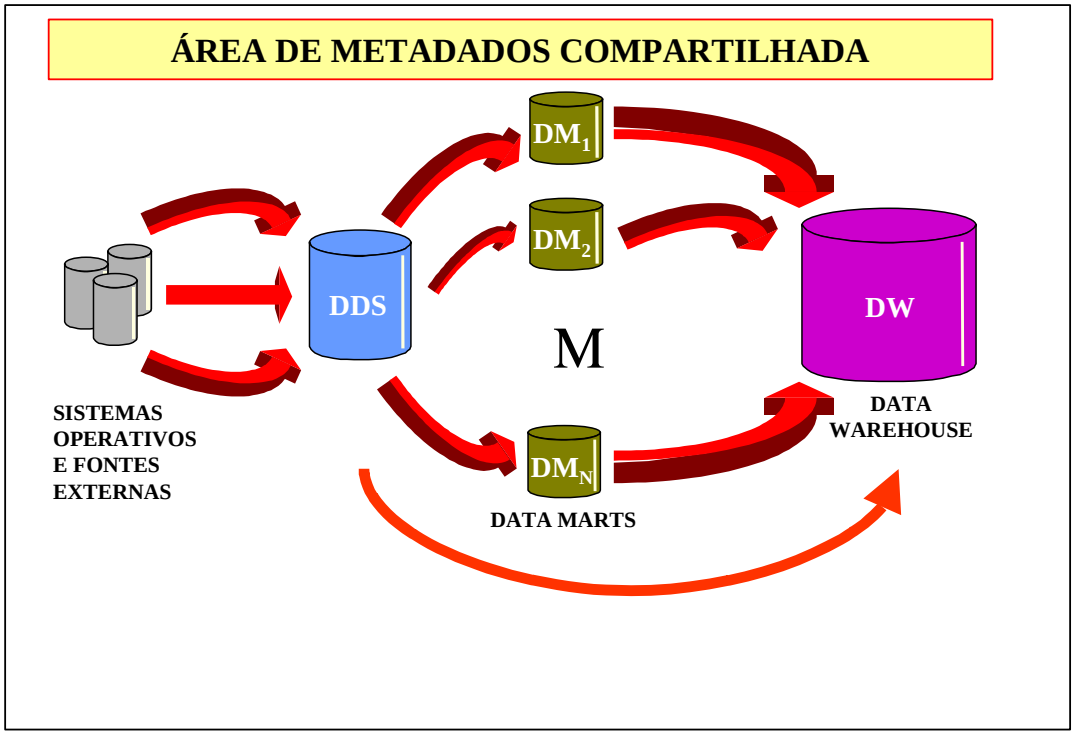

Figura 2.5 - "Enterprise Data Mart Architecture" - EDMA

Este "framework" inclui áreas de assunto da empresa, dimensões comuns, métricas, regras de negócio e fontes de dados. Seu principal objetivo é garantir uma padronização dos metadados utilizados na construção do ambiente, permitindo o desenvolvimento incremental do DW, com margens mínimas de duplicidade e inconsistência de informações. A EDMA é uma arquitetura que introduz o DDS substituindo o conceito do ODS original. A figura 2.5 apresenta esta arquitetura.

b) Arquitetura DS/DM: A arquitetura DS/DM é muito similar à arquitetura EDMA, entretanto ela substitui o DW por uma visão que representa uma conjunção lógica de DM. A arquitetura DS/DM é representada pela figura 2.6.

Para autores como Inmon (INMON, 1998), a idéia de um DW como um conjunto integrado de DM é muito difícil de ser implantada, dadas as características específicas de cada DM.

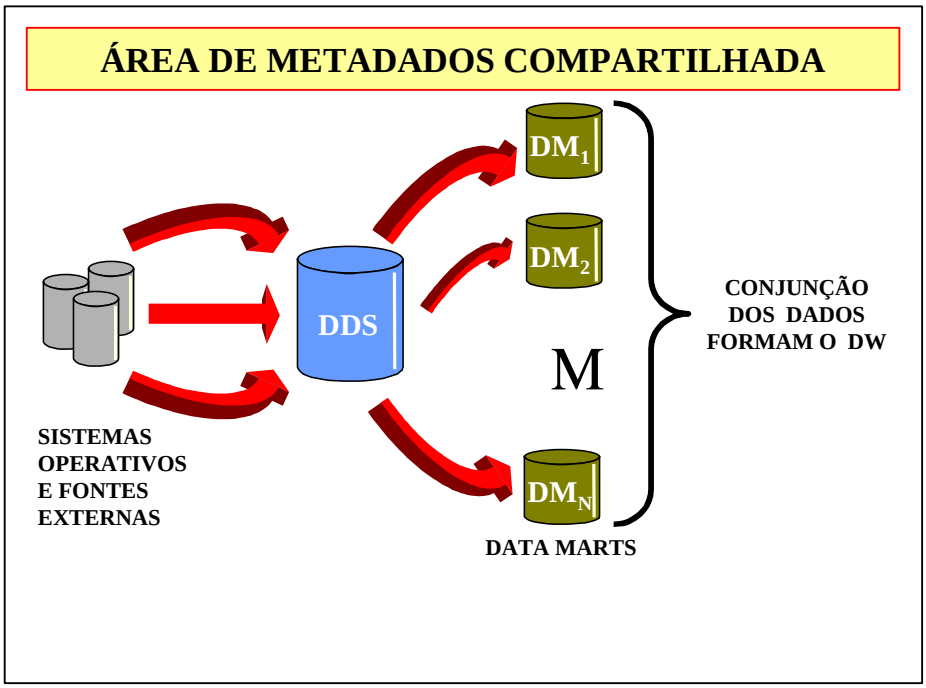

Figura 2.6 -Arquitetura "Data Storage/Data Mart" – DS/DM

2.2.2.3 Arquitetura Intermediária

Esta arquitetura tem o propósito de integrar a arquitetura "top-down" com a "bottom-up". Nesta abordagem efetua-se a modelagem de dados do DW, sendo o passo seguinte a implementação de partes desse modelo. Estas partes são escolhidas por área de interesse e constituem os DM. Cada DM gerado a partir do modelo de dados do DW é integrado no modelo físico do DW. A principal vantagem desta abordagem é a garantia da consistência dos dados. Esta garantia é obtida em virtude do modelo de dados para os DM ser único, possibilitando realizar o mapeamento e o controle dos dados.

2.3 Metodologia de Desenvolvimento

A construção do ADW é um processo longo e complexo, que requer extensa modelagem do negócio da empresa, podendo levar anos até ser bem sucedido

(CHAUDHURI, DAYAL, 1996). Este processo abrange a integração e transformação das fontes de dados; a modelagem e construção do DW; a construção dos DM e o desenvolvimento das aplicações OLAP para atender os usuários finais (MELO, 1997).

As primeiras implementações tentaram empregar as metodologias de desenvolvimento aplicadas aos sistemas tradicionais do ambiente operativo. Entretanto, por ser o ADW muito diferente do ambiente operativo, estas tentativas fracassaram. Cada um dos ambientes apresenta características que afetam a metodologia a ser empregada. A seguir estas diferenças serão detalhadas, seguidas pela apresentação das propostas de metodologia de Meyer & Cannon e da metodologia de Inmon.

2.3.1 Diferenças entre o Ambiente Operativo e o Ambiente de Data Warehouse

O ADW difere do ambiente operativo quanto aos sistemas, aos tipos de usuários, às tecnologias de suporte, ao volume de dados, ao histórico, à utilização da informação dentro do negócio e aos metadados. A tabela 2.1 apresenta as principais diferenças existentes entre eles (WU, BUCHMANN, 1997) (MEYER,CANNON, 1998).

Enquanto o ambiente operativo trabalha com sistemas orientados a processos (On-line Transaction Processing – OLTP), que têm como objetivo automatizar as principais tarefas de uma empresa, o ADW tem seu foco em sistemas capazes de acessarem e fornecerem, de forma rápida e correta, informações que proporcionem apoio à tomada de decisão. Estes sistemas são conhecidos como sistemas ou ferramentas OLAP.

Analisando as diferenças apresentadas na tabela 2.1, é possível avaliar que problemas existiriam se os dados operativos e os informacionais fossem mantidos juntos. Uma grande capacidade e velocidade de processamento seriam necessários para permitir dois tipos de acessos diferentes simultaneamente.

A abrangência e características associadas ao número de componentes envolvidos permitem avaliar que o desenvolvimento deste ambiente não é uma tarefa fácil. Até pouco tempo atrás, pouca literatura se dedicava a oferecer metodologias para o desenvolvimento deste ambiente. As metodologias que vêm sendo propostas apresentam uma preocupação com a modelagem dos dados, com a elaboração e/ou seleção de ferramentas e com a avaliação de hardware e software. Outro ponto importante diz respeito à apresentação dos resultados, tanto por questões relacionadas a

sua origem ou às próprias características da informática: os resultados de um DW devem aparecer rapidamente.

	AMBIENTE OPERATIVO	ADW
Quanto às Aplicações	Orientadas a transações	Orientadas a consultas sobre todo o banco
	Automatizam os processos da empresa, facilitando a gerência/controle dos negócios	Voltadas para o apoio e tomada de decisão
	Automatizam as operações do dia-a-dia como validar e armazenar um grande número de transações individuais.	Utilizadas para comparar e analisar padrões e tendências.
	Trabalham com acesso a registros individuais. Grande volume de transações rápidas.	Trabalham em grupo de informações
	Tempo de resposta em segundos	Tempo de resposta de segundos a minutos
	Acessa informações de forma pré-definida	A busca da informação é feita de acordo com a necessidade do usuário.
	Otimização para desempenho e disponibilidade dos dados	Otimização para interações flexíveis com o usuário final
Usuários	Representado por operadores, treinados para manipular entrada de informação	Analistas, gerentes, executivos e quaisquer usuários que necessitem de informações para tomada de decisão.
	Acessam aplicativos pré-definidos.	São capazes de elaborar consultas *ad-hoc*
Suporte ambiente	Preocupações com consistência	Preocupações com dinamismo
	Acessa informações de forma pré-definida	A busca da informação é feita de acordo com a necessidade do usuário.
Banco de Dados	Orientado para atualização de transações. Enfoque na consistência dos dados.	Orientado para consultas. Enfoque na confiabilidade dos dados.
	Volume de dados razoável. Armazena Dados detalhados	Grande volume de dados. Armazena Dados sumarizados
	Otimização do BD está voltada para atender às transações	Otimização do DW orientada para atender às consultas que, normalmente, trabalham grande quantidade de registros
	Atualização das informações ocorre com frequência.	Atualização dos dados é feita em períodos estipulados.
	Dados correntes, atômicos, isolados, "up-to-date"	Dados históricos, integrados, sumarizados.

Tabela 2.1 – Ambiente Operativo x ADW.

De uma forma geral, para auxiliar o projeto de um ADW, os desenvolvedores e projetistas devem estar atentos com relação aos seguintes itens:

- Evitar projetos muito longos. Quanto antes os resultados forem apresentados, maior será a confiabilidade e investimentos do usuário final. Dessa forma, deve se iniciar por projetos pequenos, porém importantes. *"PENSE GRANDE, comece pequeno"* (RUBINI, 1997). Deve ser dada preferência a projetos com implicações financeiras;
- Estabelecer, o quanto antes, os objetivos a serem alcançados, não permitindo o surgimento de falsas expectativas. A melhor forma de se fazer isso, são reuniões com o(s) usuário(s) final(is), formalizando o que será realizado;
- Avaliar de forma realística o volume de dados. O armazenamento ainda é um dos grandes limitadores deste ambiente, tanto pelo custo do hardware necessário, como pelos problemas de desempenho que poderão existir;
- Manter preocupação com os Metadados. É importante observar que os metadados estarão presente durante todo o processo de desenvolvimento, sendo considerado imprescindível neste ambiente; e
- Manter preocupação com a complexidade da carga dos dados e com o tempo de resposta das consultas. Apesar do tempo de resposta para as consultas ser normalmente maior que o tempo das atualizações do ambiente operativo, ele não deverá inviabilizar ou mesmo prejudicar a análise de apoio a decisão.

Os autores se preocupam em afirmar que, dentre as metodologias oferecidas, as aplicadas no ambiente operativo não servem para o ADW (INMON, HACKATHORN, 1997) (MEYER, CANNON, 1998). De uma forma geral, observa-se que essas metodologias apresentam preocupação com os seguinte itens:

a) analisar os requisitos de negócio: a equipe envolvida no desenvolvimento deste ambiente deve conhecer o negócio da empresa, evitando problemas básicos como, por exemplo, detectar as entidades mais relevantes;
b) delimitar o escopo a ser analisado: pela própria característica de tamanho do DW aliada à necessidade das empresas em apresentarem resultado, normalmente se emprega a técnica de delimitar uma primeira área para a implementação do DW. Através desta área, esta nova tecnologia será melhor apresentada aos usuários que poderão então dismistificar algumas questões e participarem mais ativamente da sua elaboração;

c) modelar os dados: ponto de destaque nas metodologias. O DW é o grande integrador dos dados que serão analisados e trabalhados neste ambiente e os DM representam as bases que serão acessadas pelos SSD;
d) elaborar e/ou selecionar ferramentas para extração e carga: ao contrário dos sistemas operativos, este ambiente é carregado a partir de fontes existentes ou externas. Uma vez definida a base de dados, se faz necessário garantir a integração e a confiabilidade dos dados. Esta garantia é obtida por meio de softwares responsáveis pela extração e carga do DW; e
e) elaborar e selecionar ferramentas para análise: por ser a análise das informações para a tomada de decisão o principal enfoque deste ambiente, a seleção de uma equipe para a elaboração de um aplicativo que tenha capacidade de analisar os dados armazenados e fornecer respostas rápidas e confiáveis é fundamental.

2.3.2 Metodologias de Desenvolvimento de um ADW

A seguir são apresentadas duas metodologias de desenvolvimento de ADW. Analisando estas metodologias é possível observar a importância da modelagem de dados neste ambiente. Uma outra questão importante a ser observada, relaciona-se a abordagem apresentada pelas duas metodologias. Ambas apresentam o desenvolvimento do ADW com uma arquitetura "top-down". Conforme discutido anteriormente, esta abordagem, apesar de academicamente correta, é considerada de difícil implementação para grandes corporações. As questões relacionadas à modelagem serão discutidas, com maiores detalhes, no próximo capítulo.

2.3.2.1 Metodologia de Meyer & Cannon

O objetivo da tecnologia de DW é a criação de uma visão simples e lógica dos dados da empresa acessível aos desenvolvedores e analistas de negócio. Estes dados, normalmente encontram-se armazenados em bases de dados fisicamente separadas. O desenvolvimento deste ambiente requer uma análise dos sistemas de interesse e um planejamento com base nessa análise. Para esta metodologia, a fase de análise e planejamento envolvem os seguintes passos (MEYER, CANNON, 1998):

1. Estabelecer a equipe A: O gerente de projeto deve estabelecer uma equipe inicial que conterá modeladores de dados, analistas de negócios e os usuários chaves do

negócio. Esta equipe irá determinar os requisitos do negócio e será a responsável por criar a definição e o escopo do projeto.

2. Determinar os requisitos iniciais do negócio: O início do projeto é feito mediante o levantamento dos requisitos de alto nível do DW e da determinação das expectativas dos usuários principais, usuários casuais e outros tipos de usuários do sistema.
3. Construir o diagrama de alto nível da área do assunto: Caso não exista é recomendável que se construa um modelo de dados de alto nível (corporativo) que represente a área de assunto (ou entidades). Este modelo de dados geralmente apresenta as área de assunto e seus relacionamentos com outras áreas. Uma destas áreas deve ser escolhida para piloto.
4. Definir o projeto e o escopo: Esta definição envolve o planejamento, identificação da descrição do projeto, objetivos, fatores críticos de sucesso, assunções e questões.
5. Construir o planejamento de Projeto: Esta fase é a responsável em detalhar o plano de desenvolvimento e construção do DW. O planejamento envolve tarefas para a construção do DW, o tempo estimado para cada atividade, avaliação de recursos e custos.
6. Escolher as ferramentas e infraestrutura: Consiste na avaliação das ferramentas e seus fabricantes/distribuidores. Os fabricantes e as ferramentas devem ser avaliados quanto a capacidade de evolução, relação entre atributos, facilidade de uso, desempenho e estabilidade do fabricante.
7. Estabelecer a equipe B: Completados os passos acima, a equipe pode ser reorganizada para trabalhar em paralelo.
8. a) Obter requisitos adicionais: Este passo difere do anterior porque a obtenção dos requisitos neste ponto está relacionada a uma área específica.

   b) Determinar os requisitos dos relatórios e análises para o usuário final.
9. Identificar as fontes do sistema: Depois que o modelo de dados é desenvolvido, os sistemas fontes dos dados são identificados e definidos .
10. Criar o modelo de dados conceitual: Depois de compor o diagrama de alto nível da área de assunto, cria-se o modelo de dados conceitual, refletindo os atributos e entidades para a área selecionada para o DW.
11. Estimar o tamanho do DW: Esta fase consiste do planejamento de armazenamento

em disco e recursos de processamento.

12. Construir o modelo de dados físico: O ABD implementa o modelo físico de dados, convertendo o modelo de dados lógico para o modelo físico através de: remoção dos dados puramente operacionais; adicionando tempo, índices e restrições de integridade referencial; efetuando merging entre tabelas; e adicionando níveis de sumarização, agregação, ou derivação dos dados. Questões de desempenho e requisitos do usuário final serão entradas para este processo.
13. Construir o banco de dados: Esta fase é responsável por criar o banco de dados e as tabelas para armazenar os dados.
14. Extrair, transformar e limpar: Esta fase é responsável por mapear os dados das fontes dos sistemas para as tabelas do DW. Tipicamente esta fase requer a transformação e "merging" de dados das fontes (OLTP) para o DW.
15. Efetuar a Carga dos Dados: Como o próprio nome diz, refere-se à etapa de carregar os dados para o banco de dados criado no item 13.
16. Desenvolver "templates" (modelos) para as ferramentas dos usuários finais: Garantir que as consultas, relatórios, gráficos, consultas *ad hoc* e análise de requisitos elaborados pelo usuário final sejam feitos via pacotes ou aplicações desenvolvidas pela própria empresa.
17. Criar o repositório e a documentação do METADADO do banco de dados: O catálogo contendo a descrição de informações armazenadas no DW e o mapeamento e transformações realizadas nos dados devem ser registradas.
18. Avaliar o "Tuning" da Base de Dados: Consiste em analisar os dados carregados para o DW com o propósito de otimizar o desempenho da carga dos dados e o acesso do usuário final.
19. Garantir a segurança dos dados: O data warehouse é desenvolvido para permitir acesso aos dados. É considerada uma falha se ele não possibilita facilidades e acessos. Entretanto, a segurança dos dados no DW requer que o acesso aos mesmos seja limitado ou controlado.
20. Documentar os processos de operação e procedimentos.
21. Realizar o treinamento: Treinar o usuário final para utilizar as ferramentas de acesso a dados de forma eficiente.
22. Realizar testes e verificar se os resultados atendem às necessidades de reduzir ou

aumentar o escopo da análise: Esta fase é responsável por criar um plano de teste, implementar e modificar de acordo com o analista de negócios (usuário final). Conduzir as avaliações pós projeto para determinar melhorias a serem aplicadas aos próximos.

2.3.2.2 Metodologia Inmon

O ADW é desenvolvido com o objetivo de atender às necessidades da organização quanto ao suporte à decisão. Esta metodologia se divide em 3 fases:

a) METOD1 :

Fase voltada para análise dos sistemas e processamentos operativos. Representa a fase operativa. Encontra-se subdividida em:

M1 - Atividades Iniciais de Projeto: Tem o propósito de obter os requisitos brutos do sistema. Emprega as técnicas de entrevistas, coleta dos dados, sessões de JAD ("Joint Application Design"). Realiza análise do plano estratégico de negócios (se houver) e analisa os sistemas existentes. O resultado desta atividade é uma descrição do impacto e influência dos sistemas existentes sobre os requisitos do sistema que está sendo desenvolvido.

M2 - Uso de Código/Dados Existentes: Tem o propósito de garantir a reutilização de código e de dados, sendo crucial para a integração do ambiente. Ela identifica os códigos/dados existentes que podem ser reutilizados.

M3 - Dimensionamento, Divisão em Fases: Tem o propósito de dimensionar e dividir o desenvolvimento em fases, com base nos requisitos gerais reunidos.

M4 - Formalização dos Requisitos: Este passo apresenta uma definição formal de requisitos, pronta para ser passada para o projeto detalhado.

CA - Análise da Capacidade: Tem o propósito de garantir a disponibilidade dos recursos necessários.

PREQ1 - Definição do Ambiente Técnico: Esta fase é a responsável pelo estabelecimento dos seguintes itens: plataforma(s) de hardware, sistema(s) operacional(is), SGBD(s), software de rede e linguagem(ns) a ser(em) usada(s) no desenvolvimento.

Dl - DER (Diagrama Entidade-Relacionamento): Este passo identifica os principais

assuntos que constituirão o sistema, bem como os relacionamentos entre eles.

D2 - DIS (Data Item Sets - Conjuntos de itens de Dados): Neste passo é realizado o detalhamento das áreas de interesse identificadas em D1.

D3 - Análise de Desempenho: Este passo se encarrega de tratar a questão da desnormalização física de dados, garantindo uma eficiente utilização dos recursos.

D4 - Projeto Físico de Banco de Dados: Este passo é o responsável pela especificação do banco de dados para o SGBD, ou para qualquer software de gerenciamento de dados que seja adotado.

P1 - Decomposição Funcional: Nesta fase é realizada uma decomposição funcional que descreve, de um nível alto até um nível mais baixo, as diferentes atividades a serem executadas.

P2 - Contexto Nível 0: Neste passo é elaborado o contexto de nível zero da decomposição funcional que descreve, no nível mais alto de abstração, as principais atividades a serem desenvolvidas.

P3 - Contexto Nível 1-n: Representa os passos restantes da decomposição funcional. Descreve as atividades mais detalhadas que ocorrem.

P4 - Diagrama de Fluxo de Dados (DFD): Responsável por elaborar o diagrama de fluxo de dados para cada processo.

P5 - Especificação de Algoritmos e Análise de Desempenho: Neste passo é delineado, passo a passo, o processo que efetivamente deve ocorrer.

P6 – Pseudocódigo: Realiza a especificação de código a ser empregada na codificação.

P7 – Codificação: Elabora o código-fonte a partir do pseudocódigo.

P8 - Walkthrough (Revisão em Grupo): Nesta fase o código é discutido publicamente e tenta-se excluir o maior número de erros possível.

P9 – Compilação: Gera o código compilado, pronto para ser testado.

P10 - Teste de Unidade: Realiza teste de código, até considerá-lo pronto para a execução.

P11 – Implementação: Representa a implementação de um sistema de funcionamento satisfatório.

A seqüência dos passos desta fase está representada na figura 2.7.

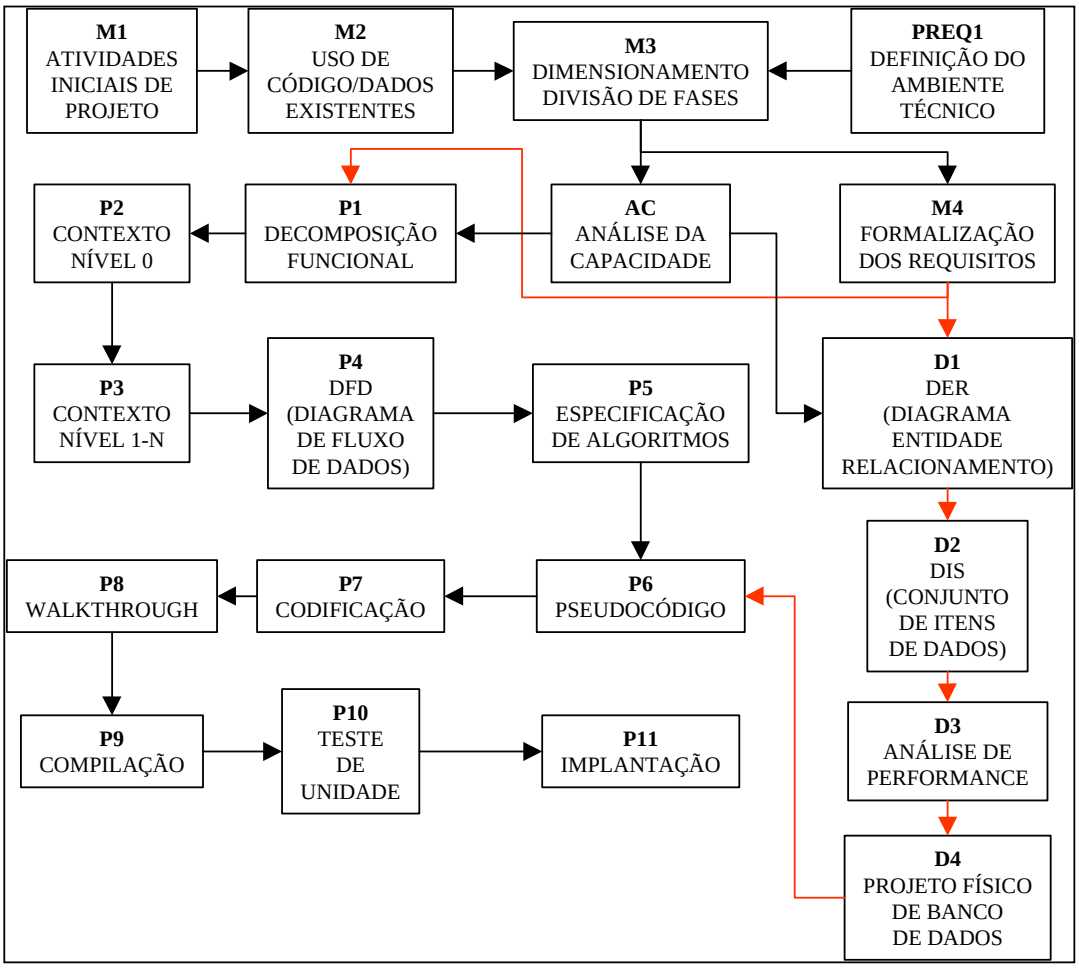

Figura 2.7- Passos da Fase METOD1

b) METOD2

Fase voltada para processamento SAD orientado ao DW. Esta fase se concentra em um modelo de dados que permita organizar a redundância dos dados. É a parte responsável pelo desenvolvimento do DW. Representa a fase de construção.

SAD1 - Análise do Modelo de Dados: Este passo avalia o modelo de dados criado. Se o modelo não atender aos critérios especificados, o andamento deve ser interrompido até que o modelo seja elevado ao padrão aceitável de qualidade.

SAD2 – Dimensionamento: Este passo é o responsável por avaliar o dimensionamento do DW. Se o DW é destinado a conter grandes quantidades de dados, deve se levar em consideração a possibilidade de existência de vários níveis de granularidade. Se o DW

não contiver uma enorme quantidade de dados, não há necessidade de planejar o projeto de vários níveis de granularidade.

SAD3 – Avaliação Técnica: Neste passo os requisitos técnicos para o gerenciamento do DW são avaliados, observando-se os requisitos e considerações técnicas para o gerenciamento de dados e processamento do ambiente operacional.

SAD4 – Preparação do Ambiente Técnico: Identifica tecnicamente como a configuração pode ser acomodada. Trata questões sobre: quantidade de dispositivo de armazenamento de acesso direto ("Direct Access Storage Device" - DASD) necessário, enlace necessário, volume de processamento previsto, dentre outros.

SAD5 – Análise das Áreas de Interesse: Realiza a seleção da área de interesse a ser povoada.

SAD6 – Projeto do Data Warehouse: Realiza o projeto físico de banco de dados para o DW.

SAD7 – Análise do Sistema-fonte: É o passo responsável pelo mapeamento de dados do ambiente operacional para o ADW.

SAD8 – Especificações: Elabora a descrição dos programas que serão usados para efetuar a passagem dos dados do ambiente operacional para o ADW.

SAD9 – Programação: Este passo envolve todas as atividades padrão de programação, desde o desenvolvimento de pseudocódigo até os testes.

SAD10 – Povoamento: Realiza a execução dos programas SAD anteriormente desenvolvidos, gerando um DW povoado e funcional.

A seqüência dos passos desta fase está representada na figura 2.8.

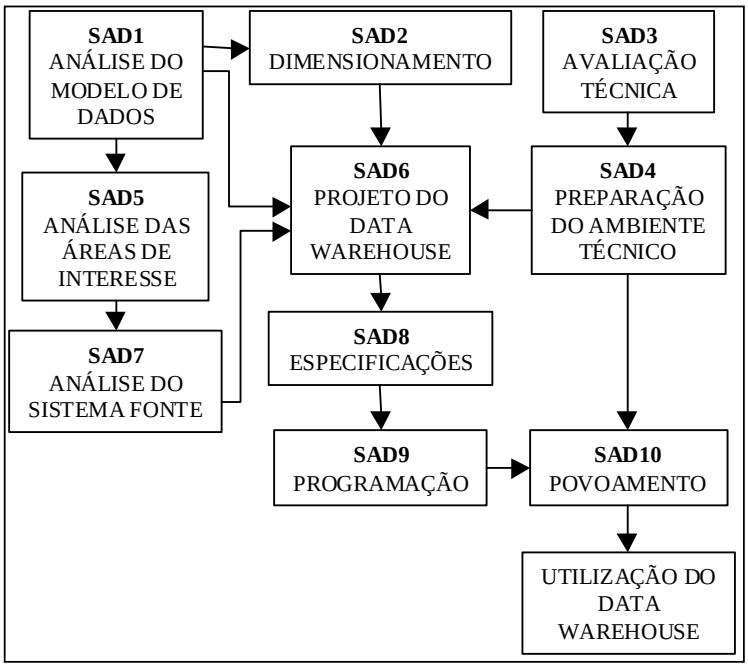

Figura 2.8 - Passos do METOD2

c) METOD3

Esta parte da metodologia descreve a utilização do DW. Representa a Fase de utilização iterativa do DW. Esta fase representa a utilização dos dados do DW para fins de análise. Ela apresenta um desenvolvimento diferente das fases anteriores, pois o processo de desenvolvimento começa pelos dados do DW, os requisitos não são conhecidos no início do processo e o processamento é feito de modo iterativo e heurístico.

IND1 – Determinação dos Dados Necessários: Seleção de dados contidos no DW para possível uso em atenção a requisitos de relatórios.

IND2 – Programação da Extração de dados: Tendo sido escolhidos os dados para o processamento analítico, o próximo passo é escrever um programa para acessar estes dados.

IND3 – Combinações, Intercalações, Análise: Geração dos dados plenamente úteis para análise.

IND4 – Análise de Dados: Analisar se os resultados obtidos atendem às necessidades do analista.

IND5 – Respostas à Questão: Avaliar se o relatório final está atendendo ou quantas

iterações de processamento são empregadas até chegar a um resultado.

IND6 – Institucionalização: Avaliar as necessidades dos relatórios gerados. Caso haja necessidade de executar algum relatório repetitivamente, faz sentido considerar o relatório como um conjunto de requisitos e recriá-lo como uma operação de ocorrência regular.

A seqüência dos passos desta fase está representada na figura 2.9.

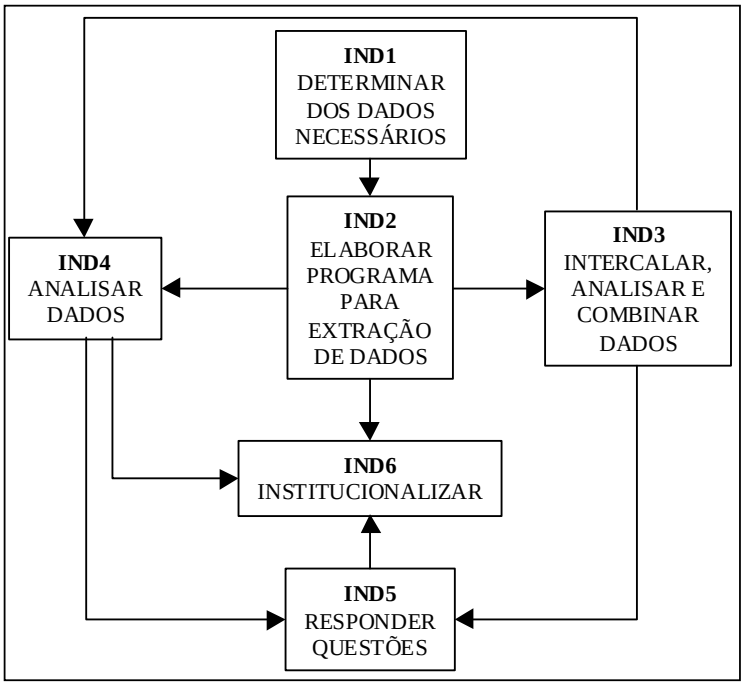

Figura 2.9 - Passos da Fase METOD3

CAPÍTULO 3

O PROBLEMA:
A MODELAGEM DE DADOS NO AMBIENTE DE DATA WAREHOUSE

A informação é a matéria-prima para o ADW, determinando a sua eficácia. O processo de modelagem envolve a integração de dados de diversas fontes e sua transformação em informações consistentes e de qualidade para permitir seu posterior acesso pelo usuário final. Este processo garante o sucesso ou o fracasso do ambiente. Para tornar o processo menos empírico, o emprego de uma modelagem de dados se torna imprescindível.

A modelagem de dados é uma das mais importantes diferenças entre o ambiente operativo e o ADW (KIMBALL, 1996). O ambiente operativo emprega, normalmente, o Diagrama de Entidade e Relacionamento (DER) para garantir o desempenho das transações, eliminado a redundância dos dados. O ADW, em virtude da necessidade de integração dos ambientes e de bons desempenhos nas consultas, apresenta requisitos diferentes quanto à modelagem. O ODS, o DW e os DM, representam repositórios possíveis de serem encontrados neste ambiente, cujas diferentes características repercutem em suas modelagens.

Dentre as possíveis combinações de repositórios existentes para este ambiente, este trabalho enfoca a arquitetura orientada ao DW e DM, por ser uma das mais empregadas pelas empresas. Existem muitas controvérsias sobre que modelos empregar nestes repositórios. Estes modelos variam com as arquiteturas e abordagens de construção e utilização a serem empregadas. Não existem regras definidas que orientem os desenvolvedores na construção de modelos de dados para o ADW. Da mesma forma, perguntas do tipo "Quais as principais preocupações do desenvolvedor?", "Que repositórios utilizar?", "Qual a melhor abordagem?", "Que tipo de modelagem selecionar?", são pouco abordadas na literatura especializada e, normalmente, de modo superficial. Um exemplo clássico é a atualização do DW segundo um desenvolvimento incremental, onde cada novo DM deve ter seus dados integrados àqueles existentes no

DW. "Qual o impacto desta integração sobre os DM existentes?", "Que medidas devem ser tomadas para acelerar possíveis modificações?". Essas e outras questões, são raramente mencionadas.

O propósito deste capítulo é destacar a importância de uma modelagem de dados que envolva todo o ambiente. Esta importância está relacionada às características primordiais do ambiente: integrar, gerenciar e controlar as informações. Além disso, são apresentados os problemas decorrentes da falta de regras formais e estratégias para a modelagem de dados nesse ambiente, ressaltando-se os problemas relacionados à complexidade, às linhas de atuação existentes no mercado e às técnicas empregadas como solução. Ao final, são levantadas as questões e dificuldades mais comuns quanto ao processo da modelagem.

3.1 Modelagem de Dados

A modelagem de dados é um modo eficiente de entender os dados. O seu propósito é prover um registro apurado de alguns aspectos do mundo real para um contexto particular. Através da modelagem, o projetista do banco de dados pode eliminar redundâncias, que representam algumas das fontes de informações inconsistentes e podem levar a sistemas ineficientes.

Um modelo de dados é uma coleção de conceitos que podem ser utilizados para descrever um conjunto de dados e as operações para a sua manipulação (BATINE, CERI, NAVATHE, 1992). A não utilização de um modelo implica em um crescimento desorganizado das aplicações, promovendo altos custos, esforços para a manutenção e dificuldades no crescimento de uma aplicação. O modelo permite ao usuário um melhor entendimento do negócio, com a vantagem de facilitar a visualização das conseqüências de qualquer ação dentro do ambiente e o impacto de qualquer mudança sobre o mesmo (DEVLIN, 1997). Ele representa a definição, caracterização e relacionamento dos dados em um determinado ambiente. Um dos diagramas mais empregados na modelagem de dados é o Diagrama de Entidade e Relacionamento (DER).

Tradicionalmente, a modelagem de dados em três níveis é empregada no ambiente operativo. Esta modelagem se inicia por um diagrama de alto nível relacionado à área do assunto, denominado modelo conceitual, seguido por uma camada intermediária, denominada modelo lógico, até a camada do modelo físico, que

representa a forma como o dados serão armazenados. Atualmente, existe uma tendência das metodologias de desenvolvimento de DW em aplicar este tipo de modelagem. Nestas metodologias, o modelo conceitual do ADW seria representado pelo modelo corporativo e os modelos lógico e físico pelo esquema estrela (modelo dimensional).

A seguir são apresentadas características dos modelos conceitual, lógico e físico empregados no ambiente operativo:

a) Modelo Conceitual

O modelo conceitual é aquele que apresenta os objetos, suas características e relacionamento como uma representação fiel ao ambiente observado. Este modelo não se preocupa com os aspectos relacionados à implementação, como por exemplo, estruturas físicas e formas de acesso de um SGBD específico (COUGO, 1997) (MACHADO, ABREU, 1995). Através deste modelo é possível criar uma descrição da realidade fácil de entender e de interpretar (BATINE, CERI, NAVATHE, 1992).

O modelo conceitual, em particular, tem sido representado através do Diagrama de Entidade/Relacionamento (DER).

b) Modelo Lógico

O modelo lógico é aquele onde os objetos, suas características e relacionamentos apresentam uma representação de acordo com as regras de implementação e limitações impostas por alguma tecnologia (COUGO, 1997). Ele descreve as estruturas que comporão o banco de dados, sem considerar nenhuma característica específica de um SGBD (MACHADO, ABREU, 1996). O modelo lógico é, normalmente representado por uma estrutura relacional, hierárquica ou em rede, sendo a modelagem relacional, a mais empregada atualmente.

A diferença entre o modelo conceitual e o lógico, entretanto, não é tão fácil de ser observada. Com o emprego das ferramentas CASE, o que se observa atualmente é o desenvolvimento de uma modelagem conceitual, muito próxima à modelagem relacional. O que não é incorreto, desde que o modelo conceitual se concentre em representar os conceitos e características de um dado ambiente (COUGO, 1997).

Em um modelo lógico, normalmente se observa informações sobre chaves de acesso, controle de chaves duplicadas, itens de repetição, normalização e integridade

referencial (COUGO, 1997).

c) Modelo Físico

O modelo físico é aquele em que a representação dos objetos é feita sob o foco do nível físico de implementação das ocorrências, ou instâncias, das entidades e seus relacionamentos. O conhecimento do modo físico de implementação das estruturas de dados é ponto básico para o domínio deste modelo (COUGO, 1997).

Este modelo descreve, a partir do modelo lógico, as características físicas associadas ao armazenamento/acesso a dados, como, por exemplo, índices, métodos de acesso e distribuição física.

3.2 Conceitos Importantes sobre Modelagem em ADW

Na modelagem de dados no ADW, termos como visão multidimensional, esquema estrela e modelagem multidimensional são muito empregados. Estes e outros conceitos importantes são abordados nos itens a seguir:

3.2.1 Data Warehouse (DW) X Data Mart (DM)

Nos primórdios do ADW, muitos desenvolvedores de ferramentas de DM pregavam que os conceitos de DW e DM apresentavam o mesmo significado (INMON, 1998). Esta concepção errada fez com que muitas empresas desenvolvessem o ADW partindo dos DM, sem grandes preocupações com o desenvolvimento de um DW e com a gerência dos dados. As observações destas experiências demonstraram que a ausência de um DW implica, dentre outras, em (INMON, 1998):

- redundância de grande volume de dados detalhados e históricos de um DM para outro;
- resultados incompatíveis e irreconciliáveis de um DM para o próximo; e
- uma interface intratável entre os DM e o ambiente operativo.

Atualmente, muitos desenvolvedores de software apresentam o DW como uma coleção de DM integrados (INMON, 1998). Entretanto, a integração de múltiplos DM não parece ser uma tarefa simples. Os DM são desenvolvidos com o propósito de atender a usuários específicos, como, por exemplo, um departamento, sem necessitar da

integração com outros DM. Os conceitos de DW e DM são detalhados a seguir:

a) Data Warehouse (DW)

O DW, por definição, é um conjunto de dados não volátil, organizado por assuntos, integrado, que varia com o passar do tempo, servindo de suporte para o processo de tomada de decisão da empresa (INMON, HACKATHORN, 1997). Os dados do DW devem ser organizados de forma simples, completa e consistente, sendo obtidos a partir de uma variedade de fontes (DEVLIN, 1997).

Por ser um integrador de dados, o DW apresenta características corporativas. A integração é feita mediante a conversão de nomes, a consistência das variáveis, a consistência da codificação das estruturas e dos atributos físicos dos dados e o estabelecimento de valores default, dentre outros. O DW, entretanto, não deve ser confundido com o banco de dados corporativo da empresa. Apesar de consolidar as informações espalhadas pelos sistemas operativos e fontes externas, estas informações apresentam um nível de granularidade diferente e até mesmo formatos diferentes daqueles existentes no ambiente operativo, pois tem o propósito de atender ao nível tático e estratégico dos negócios e não ao nível operativo. Para as empresas que necessitam analisar seus dados operativos é, normalmente, recomendado a utilização de um ODS (INMON, HACKATHORN, 1997) (INMON, 1997).

O principal objetivo do DW é suportar todas as exigências analíticas relacionadas às necessidades de gerenciamento da empresa, que sejam críticas para sua competitividade, ao invés de simplesmente enfocar problemas operacionais. O DW suporta análises de negócios e tomadas de decisão pela integração dos dados de múltiplas fontes, internas e externas à empresa. Ele consolida as informações, eliminando a inconsistência das informações por meio de operações como filtragem e sumarização. Esta transformação dos dados, para uma visão orientada a assuntos, permite realizar análises consistentes, substanciais e acuradas, promovendo não apenas auxílio à tomada de decisão, mas assessorando diversas áreas com relação, por exemplo, a auditorias e análises estratégicas (projeções). Um DW bem projetado contém os dados necessários para solucionar problemas de análise empresarial quanto a questões sobre: "O que?", "Quando?", "Por que?" e "O quê se?", eliminando a possibilidade de um fim prematuro de uma pesquisa, pela falta de determinada informação ou de tempo para o

processamento (TANLER, 1998).

b) Data Mart (DM)

O DM é uma coleção de assuntos de uma área, organizado para apoio à decisão, baseado nas necessidades de um determinado departamento ou setor (INMON, 1998). Por exemplo, uma empresa terá um DM para o departamento de finanças e outro para o departamento de vendas.

Cada DM, normalmente, possui hardware, software, dados e programas específicos. Esta característica de independência torna difícil o controle e coordenação dos dados localizados em DM diferentes. Esse é um dos principais motivos para a elaboração de um DW que funcione como um grande centralizador, ou de uma ferramenta que permita centralizar os dados distribuídos pelos diferentes DM. Segundo Inmon (INMON, 1998) os DM apresentam as seguintes características:

- são especificados para atender a uma área ou conjunto de áreas de interesse;
- empregam normalmente um esquema estrela no projeto de banco de dados. Esta modelagem é elaborada com base nas exigências dos usuários finais;
- contêm uma quantidade razoável de informações históricas, normalmente, menor que o volume histórico do DW;
- apresentam uma granularidade, normalmente, menor que a do DW. Esta granularidade tem o propósito de atender às necessidades do usuário final;
- apresentam, normalmente, o armazenamento em um Sistema Gerenciador de Banco de Dados Multidimensional (SGBDM). Os SGBDM apresentam uma boa flexibilidade de análise, porém não são recomendados para o armazenamento de grandes volumes de dados; e
- apresentam um armazenamento dos dados altamente indexado.

Existem dois tipos de DM, o DM dependente e o DM independente (INMON, 1998). A tabela 3.1 apresenta as diferenças entre os dois tipos de DM.

Um dos maiores problemas dos DM independentes está no fato de suas deficiências se manifestarem com a construção de múltiplos DM independentes, ou seja, só é possível perceber o problema após a implementação de alguns DM.

DM DEPENDENTE	DM INDEPENDENTE
Fonte é o DW	Fonte são os sistemas operativos do ambiente operativo.
Carga centralizada. Todos os DM são atualizados pela mesma fonte.	Carga descentralizada. Cada DM é atualizado de forma separada e exclusiva a partir do ambiente operativo.
Arquitetura de fácil crescimento	Difícil aproveitamento dos dados existentes. A arquitetura dificulta o crescimento.

Tabela 3.1 – Diferenças entre DM dependente e DM independente

A tabela 3.2 apresenta as principais diferenças entre o DW e o DM (INMON, 1998).

DATA WAREHOUSE	DATA MART
Corporativo	Departamental
Granularidade em baixo nível. Dados bem detalhados.	Granularidade em alto nível
Estrutura normalizada (com tratamento).	Emprega o esquema estrela como estrutura de dados
Excelente para processos de exploração.	Excelente para consultas
Grande volume de histórico de dados.	Não armazena grandes volumes de históricos de dados
Emprega tecnologia orientada ao armazenamento de grandes volumes de dados.	Emprega tecnologia multidimensional – Excelente para acesso e análise
Modelagem de dados com o propósito de atender à corporação.	Modelagem de dados com o propósito de atender a um "usuário final".
Levemente indexado.	Altamente indexado

Tabela 3.2 – Diferenças entre o DW e o DM

3.2.2 Metadados

Metadados são dados sobre dados. No ADW apresentam uma importância muito maior que a atribuída no ambiente operativo, representando a figura integradora das bases de dados que constituem o ADW (INMON, HACKATHORN, 1997). Os metadados funcionam como um guia para mapear a origem do dado, o modo como ele é

transformado e o seu conceito dentro do negócio, desde o ambiente operativo até o DW e por, conseguinte, os DM (RUBINI, 1997). Dentre as funções do metadado, merecem destaque:

- Aumentar a produtividade do DW, pois uma vez definido, um atributo poderá ser reutilizado (INMON, HACKATHORN, 1997);
- Permitir ao usuário final e aos desenvolvedores uma "navegação" pelos dados. Possibilita, por exemplo, descobrir a origem de uma informação e as regras empregadas para integrá-la no ADW (MELO, 1997);
- Permitir ao DW e aos DM apresentarem atributos com termos empregados pelos analistas de suporte e apoio à decisão. Empregando por exemplo o termo "Turma" ao invés de "nom_turma" e "Produto" ao invés de "Cod_prod";
- Gerenciar o mapeamento entre o ambiente operativo, o DW e os DM. A gerência de mapeamento de dados engloba as conversões, filtragens, alterações estruturais e qualquer outra informação necessária ao rigoroso acompanhamento das transformações. Serve como um guia para os algoritmos utilizados nos cálculos dos resumos entre os dados atuais e antigos (MELO, 1997); e
- Manter o acompanhamento das alterações estruturais dos dados ao longo dos anos.

Para garantir o cumprimento dessas funções, os metadados devem manter informações sobre (INMON, 1997) (MELO, 1997):

- Estrutura de dados, segundo a visão do programador;
- Estrutura de dados, segundo a visão do usuário final;
- Fonte de dados que alimentam o DW;
- Transformações sofridas pelos dados no momento de sua migração para o DW;
- Transformações sofridas pelos dados no momento de sua migração para o DM;
- Modelo de dados;
- Relacionamento entre o modelo de dados , o DW e os DM; e
- Histórico de extrações.

3.2.3 Visão X Modelagem X SGBD (Multidimensionais)

A diferença entre visão, modelagem e SGBD multidimensionais nem sempre é clara. A seguir são apresentadas as definições destes termos, assumidas neste trabalho:

a) Visão Multidimensional:

Esta visão representa a forma como os executivos, analistas de negócios e especialistas analisam as informações do negócio. Normalmente, estes usuários não trabalham com uma informação específica, mas sim, analisam um cruzamento delas, como por exemplo o volume de vendas de um determinado produto ao longo dos últimos meses por regiões do País. A figura 3.1 apresenta um cubo dimensional, representando uma visão para análise de volume de vendas por Produto X Região X Tempo.

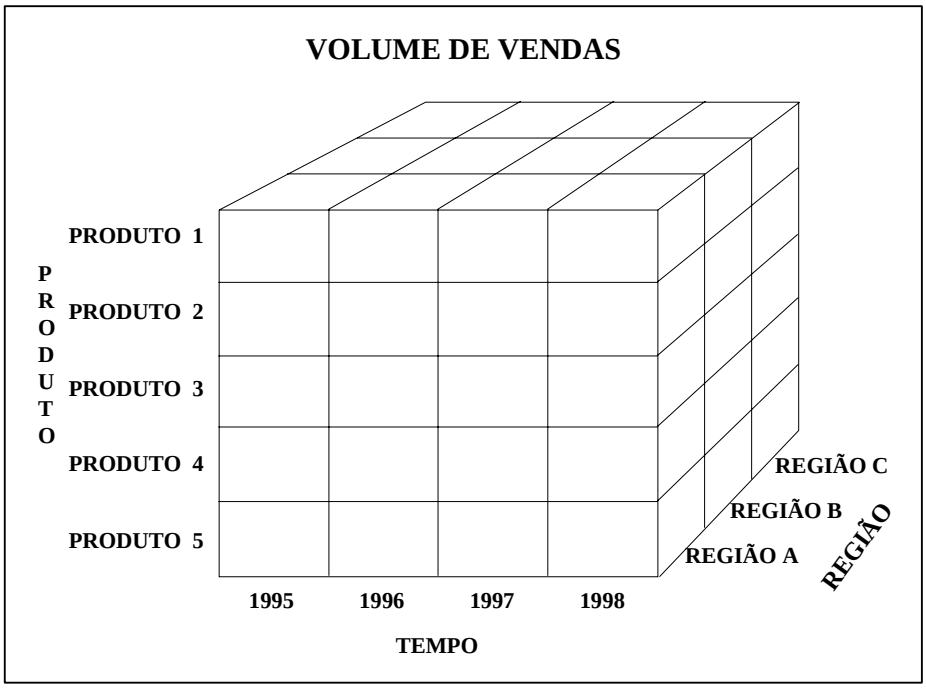

Figura 3.1 - Visão Dimensional para Análise de Volume de Vendas

b) Modelagem Multidimensional:

A modelagem multidimensional representa o modo como as informações existentes na empresa serão modeladas. O objetivo principal desta modelagem é a elaboração de um projeto de banco de dados consistente, que permita atingir, de modo preciso, a visão multidimensional do usuário final. O modelo final deve ser, portanto, facilmente entendido e interpretado pelos desenvolvedores e usuários finais. Este modelo é conhecido por modelo dimensional.

Esta modelagem independe do SGBD onde será implementada. O modelo final pode ser implementado em um SGBD relacional ou em um SGBD multidimensional. O esquema estrela é, normalmente, empregado para representar esta modelagem (TANLER,1997). Este esquema será apresentado em 3.3 – Modelagem Dimensional com Esquema Estrela.

A modelagem multidimensional torna simples as operações realizadas pelas ferramentas do ADW, como por exemplo:

- "Rollup" : permite que o usuário reduza o escopo da análise. Ele sobe o nível de detalhe, ou seja incrementa o nível de agregação.
- "Drill-Down": caminho inverso do "rollup", permitindo que o usuário desça a nível de detalhes ao analisar um problema.

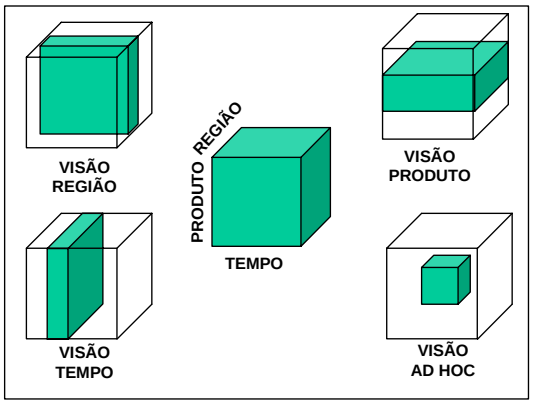

Figura 3.2 - "Slice & Dice"

- "Slice-Dice": seleção e projeção. Os dados no DW devem ser projetados de modo a permitir consultas que combinem e separem os dados, através de qualquer medição possível do negócio. A figura 3.2 apresenta um exemplo de consulta empregando "Slice & Dice". O exemplo apresenta as várias visões para Produto X Região X Tempo.
- "Pivot" (pivoteamento): permite que o usuário reorganize a disposição das dimensões para uma nova visão dos dados.

Operações como o "ranking" (sort), seleções e definição de atributos computados, como média e soma, são também muito aplicadas.

c) SGBD Multidimensional (SGBDM) :

O SGBDM é um gerenciador de bancos de dados que utiliza uma estrutura multidimensional para o armazenamento de suas informações. Normalmente, os SGBDM trabalham com uma estrutura de cubos de informações, que garantem um melhor desempenho na análise e recuperação de informações. Estes cubos são gerados e armazenados previamente. Alguns bancos multidimensionais requerem uma completa recarga do banco quando uma reestruturação ocorre (CAMPOS, FILHO, 1997).

A tabela 3.3 apresenta algumas diferenças entre o emprego de um SGBDM e um SGBDR para tratar com o modelo dimensional (INMON, 1997).

SGBDM – Cubos	SGBDM - Relacional
Não suporta muitos dados. Restrições quanto ao número de dimensões.	Suporta um grande volume de dados.
Tecnologia começa a ser empregada.	Tecnologia comprovada.
Junção dinâmica questionável.	Apresenta junção dinâmica de dados.
Não suporta processamento de atualização de uso geral.	Apresenta bom processamento de atualização.
Desempenho otimizado para processamento de apoio à decisão.	Desempenho não chega a ser excelente.
Estrutura de dados pode ser otimizada para um padrão de acesso conhecido.	Não pode ser otimizada exclusivamente para processamento de acesso.
Não apresenta estrutura flexível para acessar dados por caminho não preparado.	Fácil acesso a dados.

Tabela 3.3 – Diferenças entre SGBDM e SGBDR

3.3 A Modelagem Dimensional com o Esquema Estrela

O esquema estrela é uma estrutura simples, com poucas tabelas e ligações ("joins") bem definidas, que permite (POE, KLAUER, BROBST, 1998):

- Facilidade de leitura e entendimento, não só pelos analistas, como por usuários finais não familiarizados com estruturas de banco de dados ;
- Um projeto de um banco de dados com uma visão mais próxima à do usuário final;
- Criação de um banco de dados que propicie consultas rápidas, realizadas de modo eficiente e intuitivo pelo usuário final;
- Entendimento e "navegação" dos metadados pelos desenvolvedores e usuários finais; e
- Utilização de uma série de ferramentas "front-end", desenvolvidas especialmente para atender a este tipo de modelo.

Este esquema, entretanto, apresenta dificuldades com as dimensões, quando as mesmas são muito grandes ou aparecem em grande número. Os sistemas que apresentam estas características não devem forçar a utilização deste esquema (POE, KLAUER, BROBST, 1998). Além disso, este esquema não apresenta uma forma clara de tratar hierarquias implícitas.

O nome "Estrela" está associado à disposição física do modelo, que consiste de uma tabela central, a tabela de fatos, que se relaciona com "n" tabelas de dimensões. A figura 3.3 apresenta este esquema.

O esquema estrela pode representar tanto o modelo lógico, como o modelo físico do banco de dados (KIMBALL, 1997). A representação mais simples de um modelo dimensional contém um esquema estrela com uma tabela de fatos relacionada com tabelas de dimensões. Na verdade, um modelo dimensional pode ser representado por uma ou mais tabelas de fatos, relacionadas com tabelas de dimensões. Entretanto, a visão de um esquema por vez torna o modelo mais claro. A seguir serão apresentadas as definições e características dos componentes do esquema estrela (RUBINI, 1997) (MELO, 1997) (KIMBALL, 1997) (KIMBALL *et al* , 1998).

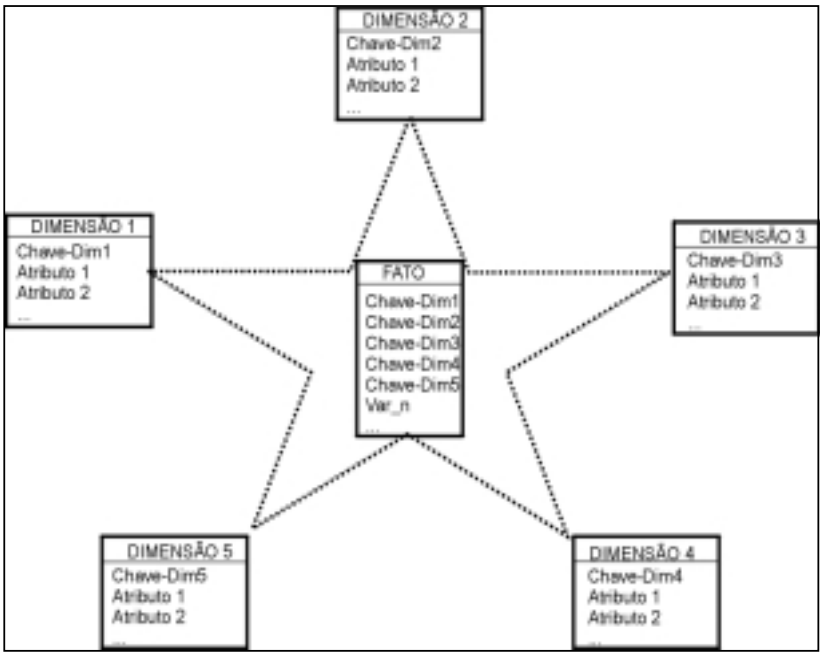

Figura 3.3 – Representação do Esquema Estrela

3.3.1 Tabela de Fatos

A tabela de fatos representa as informações que serão avaliadas, sendo, normalmente, constituída de valores numéricos que representam os objetos da análise, como por exemplo, total de vendas, total de movimentação e média de reservas canceladas. Algumas vezes é possível encontrar tabelas de fatos sem valores numéricos, nesses casos, normalmente, a tabela de fatos é empregada para mapear eventos

(KIMBALL, 1997). A tabela de fatos normalmente é grande, apresentando muitos registros. Esta tabela contêm as informações básicas do nível de transação do negócio, de interesse particular a uma aplicação. Uma característica importante da tabela de fatos é a esparsidade, dessa forma, quando não existe valores para um cruzamento de dimensões, não são armazenados zeros. A tabela de fatos armazena as medições numéricas de interesse para o negócio. Os projetistas devem dar preferência aos atributos que representem valores perfeitamente aditivos (KIMBALL *et al*, 1998).

3.3.1.1 Classificação dos Fatos

Segundo Kimball (KIMBALL *et al*, 1998), os fatos podem ser classificados em transações individuais, "snaphots" e linhas de itens.

As transações individuais, normalmente, apresentam uma estrutura muito simples, com um campo acumulado que contém o valor da transação;

Os fatos "snapshots" representam medidas de atividades extraídas em tempo determinado, como, por exemplo, fim do dia ou fim do mês; e

Os fatos do tipo "linhas de itens" são aqueles que representam exatamente uma linha de item, como, por exemplo, itens de pedido, itens de entrega e itens de apólice de seguro.

3.3.1.2 Fatos Com Produtos Heterogêneos

A área financeira representa um dos melhores exemplos para produtos heterogêneos, porque, normalmente, trabalha com uma variedade de produtos e serviços. Para estes modelos dimensionais, recomenda-se a criação de uma dimensão geral e de dimensões específicas para os produtos/serviços. Da mesma forma, será necessário uma tabela de fatos gerais e tabelas de fatos específicas para cada produto. Estas tabelas de fatos específicas podem apresentar sua chave compondo a chave da tabela de fatos geral. Desse modo, é possível agilizar o processo de consultas.

Os fatos específicos ou "subfatos" representam medições numéricas de uma dimensão específica, podendo ser inseridas na tabela de fatos principal (ou global) (KIMBALL *et al*, 1998).

A figura 3.4 apresenta uma tabela de fatos de análises bancárias, com um subfato

relacionado ao produto específico "depósito".

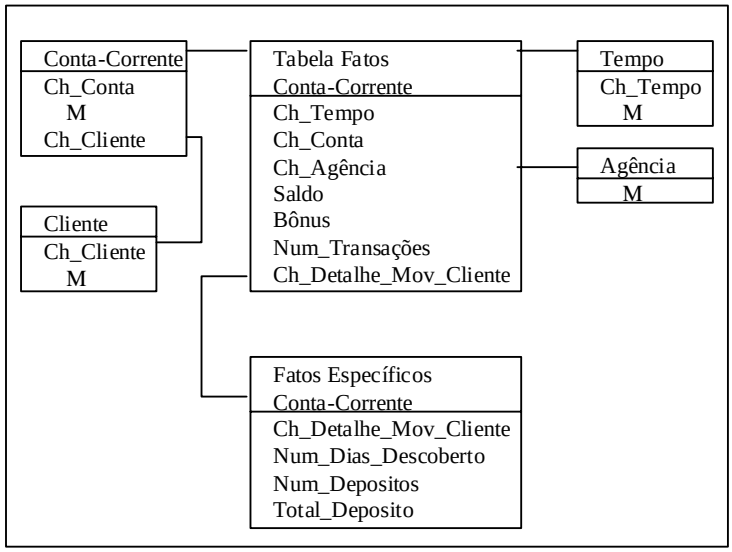

Figura 3.4 -Tabela de fatos específicos

3.3.1.3 Classificação dos Atributos Numéricos em uma Tabela de Fatos

Os atributos mais comuns em uma tabela de fatos são valores numéricos. Estes valores podem ser de três tipos:

a) Valores Aditivos: são valores da tabela de fatos sobre os quais podem ser aplicadas as operações de soma, subtração e média. Os valores, como por exemplo, "total de vendas" e "total de itens vendidos", por Produto X Região X Loja representam valores aditivos.

b) Valores Não Aditivos: são valores da tabela de fatos que não podem ser manipulados livremente, como valores percentuais ou relativos. Para esses tipos de valores, os cálculos devem ser realizados sobre os dados absolutos nos quais se baseiam. Todos os valores que medem um nível de intensidade, são valores estáticos não aditivos. Esses valores são válidos para o momento em que a informação foi obtida, e sua soma através do tempo não tem significado, entretanto, podem ser úteis para futuras manipulações.

Segundo Kimball (KIMBALL, 1997), estes valores, normalmente mostram totais segundo fatores multiplicativos diferentes. Para estes casos, é recomendável

que se mantenham os valores unitários, geradores dos valores não aditivos, na tabela de fatos. As consultas a serem realizadas sobre a tabela de fatos podem realizar os cálculos ou se for o caso, visões podem ser criadas com os valores calculados. Um exemplo de valores não aditivos pode ser representado por informações de totais a preço de custo e a preço de venda, de itens perdidos e danificados, armazenados na tabela de fatos "ESTOCAGEM".

Para não carregar a tabela de fatos, os valores unitários devem ser mantidos, e uma visão pode ser criada para apresentar os valores computados.

c) Valores Semi Aditivos: são valores que envolvem contagem dupla. Portanto, são restritos a uma dimensão. Quando a análise é efetuada sobre a dimensão aditiva, as operações normais podem ser aplicadas sobre o valor.

Segundo Kimball (KIMBALL, 1996), todas as medições que registram um nível estático, como níveis de estoque e saldos de contas financeiras, e medições de intensidade, como de temperatura ambiente, são informações não aditivas ao longo do tempo. Entretanto, podem ser agregadas, de forma útil, ao longo do tempo através do cálculo da média do número de períodos de tempo. Portanto, podem ser considerados valores semi-aditivos. Três soluções são recomendadas para o tratamento deste tipo de valor:

- Qualquer tentativa de processamento sobre valores semi-aditivos deve ser avisada ao usuário;
- Recorrer ao dado básico, isto é, ao dado ainda não agregado, de onde foi extraída a tabela de fatos correspondente; e
- Gerar na tabela de fatos, registros que armazenem o total real, embutindo uma agregação em relação ao atributo semi-aditivo.

Um exemplo de valor semi-aditivo é o valor em estoque por mês de um produto. As manipulações e consultas sobre estes fatos devem restringir ou agregar o período, pois a soma dos totais em estoque em mais de um período, contabilizaria mais de uma vez itens em estoque.

3.3.2 Tabela de Dimensão

A tabela de dimensão armazena as informações necessárias para análises ao longo de dimensões, sendo normalmente, menor que a tabela de fato. Esta tabela apresenta chave simples e seus campos, normalmente descritivos, são empregados como fonte das restrições e linhas de cabeçalhos para relatórios.

A qualidade do banco de dados é proporcional a dos atributos de dimensões, portanto deve ser dedicado tempo e atenção a sua descrição, ao seu preenchimento e a sua garantia de qualidade (KIMBALL, 1996).

3.3.2.1 Dimensões com Itens Heterogêneos

A dimensão que descreve itens heterogêneos, segundo Kimball (KIMBALL, 1996) (KIMBALL *et a*l, 1998), é aquela cujos atributos representam mais de um produto ou serviço. Este tipo de dimensão é comum na área financeira. Para estas dimensões recomenda-se, após a definição do modelo principal, contendo a tabela de fatos central e a tabela dimensional central, a criação de tabelas de fatos e tabelas dimensionais específicas para cada tipo de item (KIMBALL *et a*l, 1998).

A definição da dimensão global, abrangendo os vários tipos de itens, possibilita a elaboração de consultas gerais. Enquanto a definição das dimensões específicas permite consultas específicas com um maior nível de detalhe.

3.3.2.2 Hierarquia de Dimensões

Segundo Thomsen (1997, p.66), "uma hierarquia é um atributo de uma dimensão". As hierarquias são a base para a agregação de dados e para a navegação entre os diferentes níveis de detalhe em um estrutura multidimensional (THOMSEN,1997) (MEYER, CANNON, 1998). As hierarquias descrevem a estrutura organizacional e lógica dos relacionamentos entre os dados (MEYER, CANNON, 1998). A figura 3.5 (I) apresenta os níveis de agregações que podem ser aplicados a dimensão ***Produto***. Muitas dimensões apresentam uma estrutura hierárquica ou multinível. A figura 3.5 (II) apresenta as hierarquias multiníveis para a dimensão ***Tempo***.

Algumas estruturas hierárquicas são facilmente identificadas, como por exemplo, uma estrutura de tempo representada por horas, dias, semanas, meses,

trimestres e anos; e uma estrutura geográfica representada por cidades, municípios, estados, regiões e países (KIMBALL, 1996). Dois tipos de hierarquia podem ser considerados para uma dimensão: explicita e implícita.

- Hierarquias explicitas: segundo Kimball (KIMBALL, 1997), a identificação deste tipo de hierarquia é realizada através de uma análise do DER. As hierarquias são caraterizadas por uma seqüência de entidades interligadas, cujos relacionamentos, entre cada par de entidades na seqüência, sejam N:1. A figura 3.5(I) representa a hierarquia explícita para a dimensão ***Produto***. Essa hierarquia é constituída pelas dimensões ***Tipo*** e ***Categoria***.
- Hierarquias Implícitas: também conhecidas como múltiplas hierarquias, representam as hierarquias embutidas nos atributos das dimensões. Um exemplo para múltiplas dimensões, pode ser observado na figura 3.6. A figura apresenta a classificação de produtos de um supermercado. Essa classificação é feita nas categorias alimentos e material de limpeza. Os alimentos podem ser subcategorizados, quanto à duração, em perecíveis ou não perecíveis, ou, quanto à fórmula, em dietético ou não dietético. Do mesmo modo, os materiais de limpeza podem ser classificados, quanto à sua fórmula, em tóxica ou não tóxica, ou, quanto à sua consistência, em líquida, pastosa ou em pó.

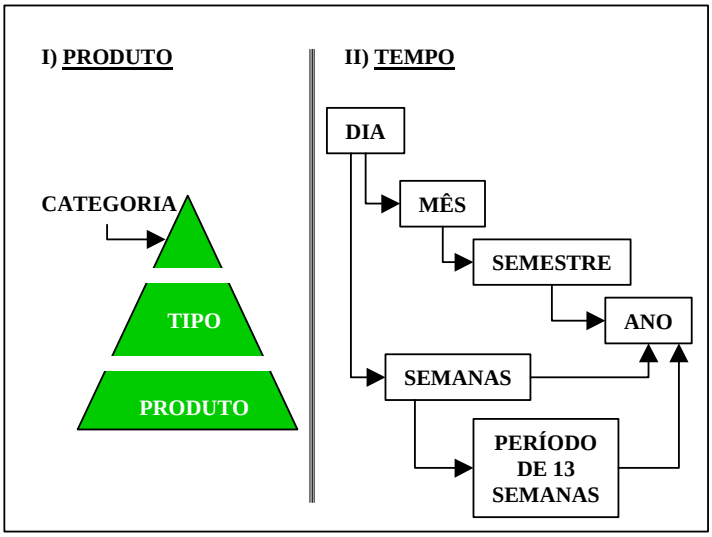

Figura 3.5 - Hierarquias de Produto (I) e de Tempo (II)

Uma questão importante a ser abordada diz respeito à influência da hierarquia das dimensões sobre a tabela de fato. A tabela de fatos deve refletir a menor granularidade das dimensões.

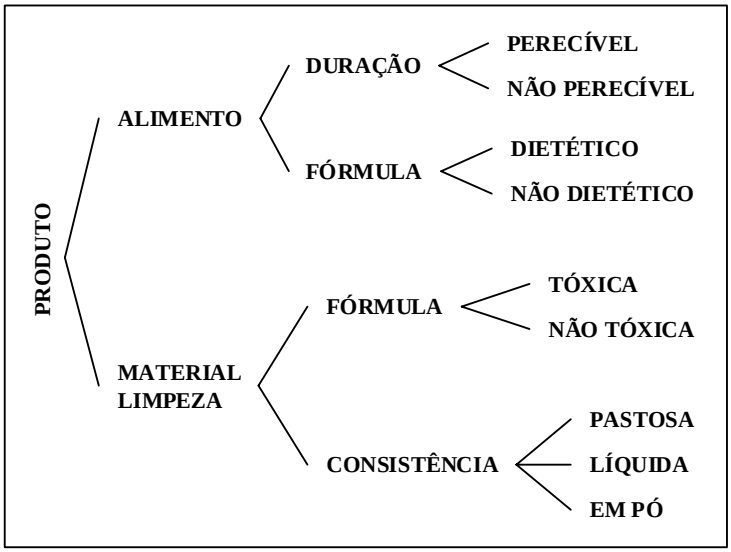

Figura 3.6-Hierarquia implícita para Produto

Ao se estabelecer as hierarquias nas dimensões, a menor granularidade deve ser mantida na tabela de fatos, de modo a garantir que não sejam armazenados registros que representem totais referentes a um nível mais alto na hierarquia de uma dimensão. Dessa forma, se a dimensão ***Produto***, na figura 3.5 (I), apresenta a hierarquia: *CATEGORIA/ TIPO/ PRODUTO*, os registros na tabela de fatos devem indicar totais no nível de produto. Os registros que totalizem por *CATEGORIA* ou *TIPO* não devem ser armazenados.

3.3.2.3 Dimensões Descaracterizadas:

As dimensões descaracterizadas, também conhecidas como dimensões degeneradas, são representadas na tabela de fatos como chaves de dimensão, sem que exista a tabela de dimensão. Um exemplo de dimensão descaraterizada é representada pelo número de controle de documentos, número de pedidos e número de faturas. Este tipo de atributo, normalmente, é empregado em tabelas de fatos onde o grão da tabela representa o documento propriamente dito ou uma linha de item do documento.

3.4 Técnicas de Modelagem Dimensional

Técnica, segundo o dicionário "Aurélio" (FERREIRA, 1986, p.1656), é a

maneira, jeito ou habilidade especial de executar ou fazer algo. Portanto, técnicas de modelagem dimensional, ou multidimensional, representam uma maneira de se desenvolver um modelo dimensional. Estas técnicas têm o propósito de gerar um modelo dimensional que represente a área de interesse e apresente um bom desempenho para as consultas. A seguir são apresentadas algumas das técnicas de modelagem dimensional mais empregadas.

3.4.1 Tratamento de Dimensões e Fatos com Cardinalidade M:N

Segundo Kimball (KIMBALL *et al*, 1998), apesar de, normalmente, a cardinalidade entre as dimensões e a tabela de fatos ser 1:N, podem acontecer casos em que esta cardinalidade seja M:N. Para este tipo de cardinalidade, Kimball (KIMBALL *et al*, 1998) recomenda a adição de uma nova dimensão que representará uma ponte entre a dimensão original e a tabela de fatos. A identificação deste tipo de cardinalidade empregando um desenvolvimento baseado apenas em técnicas dimensionais não é uma tarefa trivial. Neste tipo de desenvolvimento, como será apresentado mais adiante, as dimensões e a própria tabela de fatos são definidas de acordo com a especificação do usuário final. Para identificar a existência da cardinalidade M:N é necessário uma análise mais minuciosa, cruzando o modelo gerado com as regras do negócio.

Quando esta cardinalidade é identificada, deve ser realizada a inserção de uma dimensão ponte. Esta dimensão ponte representa efetivamente a cardinalidade M:N, com a tabela de fatos. Esta dimensão, além de sua chave normal, terá uma chave que permite identificar o conjunto de informações. Esta chave, única para um conjunto, será inserida na tabela de fatos. Dessa forma, será possível realizar as consultas pela dimensão origem, garantindo que não haverão repetições na tabela de fatos.

No capítulo quatro, será apresentada uma variação desta técnica e uma técnica alternativa para lidar com esse problema.

3.4.2 Técnicas de Rastreamento de Alterações

Os modelos de dados do ambiente operativo fazem pouca, ou mesmo, nenhuma distinção entre os dados estáveis e aqueles freqüentemente alterados. Como o ADW é sensível às mudanças, uma organização ótima é aquela que separa os dados de acordo com sua freqüência de atualização. As modificações dinâmicas exigem um grande

controle de sincronismo, o que é difícil, em virtude do grande volume de informações armazenadas neste ambiente.

Os atributos descritivos, normalmente, apresentam informações que evoluem lentamente ao longo do tempo. Por exemplo, o atributo *ESTADO_CIVIL* na dimensão CLIENTE. Ao longo dos anos, as pessoas se casam, enviuvam e se divorciam. O emprego de minidimensões representa uma forma de tratar as alterações em dimensões. Entretanto, outras soluções são apresentadas para esse problema. Para garantir a manutenção do histórico dos dados através do tempo, acompanhando a sua evolução são apresentadas três soluções (KIMBALL, 1996) (McGUFF,1998):

- Não manter o histórico, e simplesmente sobrescrever: a única vantagem desta solução é a facilidade de implementação. A mudança ocorrerá na dimensão responsável pela informação, onde um registro será alterado recebendo um novo valor. Quando as mudanças ocorrem para acertos no cadastro esta solução é válida. Porém, para os demais casos, esta não é a solução ideal, porque não atinge o propósito de manter o acompanhamento histórico dos dados.
- Adicionar um novo registro, com uma nova chave, e a nova descrição: esta solução exige uma chave genérica. Kimball sugere a adoção de um formato de chave que adicione os dígitos de versão ao final. As chaves genéricas devem estar descritas nos metadados e serem tratadas pelas aplicações do usuário final. Esta solução não impõe maiores complexidades às aplicações. Nos aplicativos de navegação pelo DW, as consultas pressupõem uma visão dos dados através do tempo. As consultas são feitas de forma a obter totais por período, particionando naturalmente a tabela de fatos de forma cronológica.
- Criar um campo a mais para o atributo em questão na tabela dimensão, para manter o valor corrente: Esta solução permite visualizar ou restringir a dimensão de acordo com o valor original ou com o valor corrente do atributo que sofreu a alteração. É mais complexa que a solução anterior com relação às aplicações, sendo pouco aplicada na prática. Apesar de nenhum registro novo ser criado, torna-se necessária a manutenção de dois campos para um atributo, um com o valor corrente e o outro com o valor original. É necessária a criação de um campo com a data em que o valor corrente entrou em vigor. Como não existe mudança de chave, a única forma de identificar a mudança é através da referência à data da alteração. Utilizando, como

exemplo, o atributo estado_civil na dimensão ALUNO, a descrição dos registros passaria a ter dois novos campos - est_civil_original, est_civil_data_efetiva - e o campo est_civil passaria a chamar-se est_civil_corrente.

Esta solução é própria para uma fase de adaptação, quando se precisa visualizar os dados com base no valor antigo ou novo do atributo, como se não existissem mudanças. A complexidade de se implementar este tipo de solução reside no tratamento destas considerações, que devem se referir ao campo de data efetiva. Além disso, manter apenas o valor original e o valor corrente de um atributo não atinge, por completo, o objetivo de acompanhar o histórico dos dados, pois os valores intermediários são perdidos.

3.4.3 Criação de Novas Chaves

É comum, a criação de novas chaves identificadoras para as dimensões em um ADW. Esta nova identificação, normalmente, é decorrente das seguintes necessidades (INMON, 1997):

- remapeamento de chave para evitar a dependência da chave original. Essa necessidade ocorre quando existe a possibilidade de alteração de chave e é necessário evitar a sua reutilização. A reutilização de chaves é comum no ambiente operativo devido a sua pequena periodicidade de armazenamento. O ADW, ao contrário armazena os registros por um longo período, exigindo uma nova chave para evitar duplicidades e inconsistências para as consultas;
- remapeamento de chaves, reduzindo chaves longas para obter um melhor desempenho nas consultas;
- estabelecimento de chaves genéricas, permitindo mudanças na descrição dos itens sem provocar alterações nas chaves. A solução normalmente adotada no nível físico, é acrescentar dois ou mais dígitos ao final da chave original. Estes novos dígitos indicam a versão do item. As chaves genéricas permitem o rastreamento de modificações pela generalização da chave primária.

Apesar da modelagem física não pertencer ao escopo deste trabalho é interessante observar que a nível físico, é comum a criação de chaves onde a estrutura representa uma montagem baseada em processos de "hashing", para facilitar o acesso.

3.4.4 Criação de Minidimensões

Segundo Kimball (1996, p.99), "A melhor abordagem para analisar modificações em tabelas dimensionais extremamente grandes é subdividi-las em minidimensões compostas por pequenos conjuntos de atributos estruturados para conter um número limitado de valores."

As minidimensões representam uma das melhores formas de tratar periodicidades de atualização diferentes em dimensões muito grandes (KIMBALL, 1996). Além disso, permitem a manutenção de um bom desempenho porque, normalmente, se relacionam com a tabela de fatos, permitindo consultas diretas. O relacionamento da minidimensão com a dimensão origem permite outros meios de navegação.

A figura 3.7 apresenta um exemplo de minidimensão DEMOGRÁFICA para a dimensão CLIENTE.

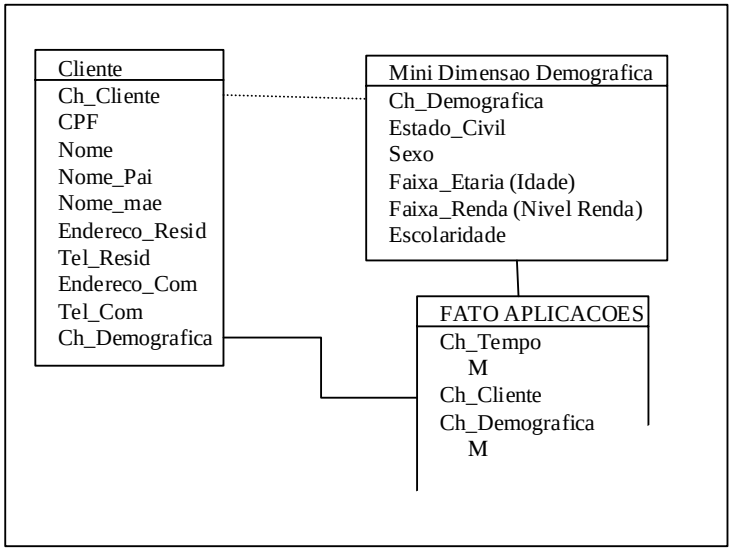

Figura 3.7 - Minidimensão DEMOGRÁFICA

3.5 O Processo de Modelagem do ADW

O projeto de um ADW envolve a seleção de componentes que constituirão o ambiente, a seleção de uma abordagem para a construção dos componentes selecionados e a definição do tipo de modelagem de dados a ser empregada nos repositórios. Uma breve descrição do que trata cada uma destas etapas é descrita a seguir:

- Seleção dos componentes: é realizada através de um estudo das áreas de interesse que irão compor o ADW. De acordo com o número de áreas, o volume de informações e os sistemas operativos existentes é possível realizar uma estimativa de necessidade de repositórios;
- Seleção da abordagem para a construção dos componentes: a abordagem "bottom-up" é a mais recomendada, pois permite balancear a relação custo x benefício, apresentando resultados rápidos e conquistando investimentos a nível pessoal e financeiro; e
- Definição da modelagem a ser empregada: conforme já mencionado, existe um consenso quanto ao emprego da modelagem multidimensional para os DM. Entretanto, para o DW, a modelagem varia de acordo com sua aplicabilidade, podendo ser aplicada a visão Inmoniana ou a visão Kimballiana.

A observação de experiências de desenvolvimento de ADW demonstra que o conhecimento do negócio é alcançado, através da análise dos modelos de dados existentes no ambiente operativo. Esses modelos permitem agilizar o desenvolvimento e garantir a simplicidade para novas interações.

O ADW acessa dados operativos, que são transformados e integrados para gerarem dados informativos e analíticos. A integração entre os modelos do ambiente operativo e o ADW garante que esse processo funcione da melhor forma possível. O processo de modelagem deve, portanto, transformar modelos de dados orientados a processo, os modelos funcionais, para modelos de dados orientados a negócio, os modelos dimensionais.

Através da modelagem de dados realiza-se a transformação de uma visão de processo em uma visão de negócio. Essa transformação está representada na figura 3.8.

Conforme apresentado na figura 3.8, o DW é o centralizador de dados, apresentando diferentes propostas de modelagem: a modelagem dimensional (visão Kimball) e a modelagem não totalmente normalizada, porém sem ser dimensional (visão Inmon-Graziano). A decisão sobre qual melhor modelo aplicar ao DW é uma decisão de projeto, variando caso a caso, de acordo com as necessidades da empresa.

Os dados no ADW representam uma composição de seus repositórios, sendo a fase da modelagem responsável por garantir a integração e consolidação dos mesmos. Essa garantia é obtida através do emprego de metadados que permitem o gerenciamento

e controle dos dados, desde o ambiente operativo até a sua visão pelo usuário final.

O processo de modelagem deste ambiente é realizado com base nos requisitos e

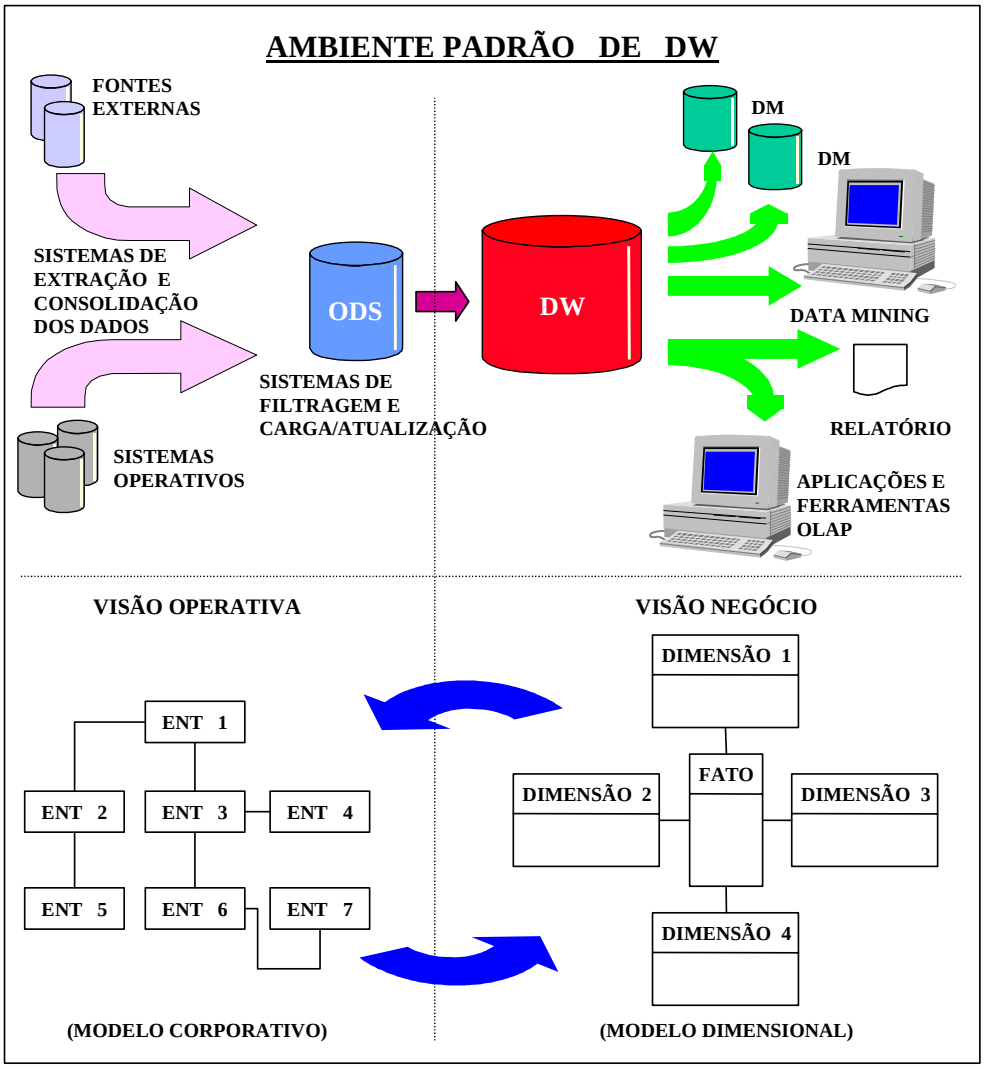

Figura 3.8 -Transformação de Visões

nas necessidades estabelecidos pelos usuários finais. O ponto de partida é a análise dos modelos de dados do ambiente operativo. A partir desta análise são obtidos os dados que serão trabalhados, consolidados e sumarizados. Esses dados comporão o modelo do DW, na abordagem "top-down" ou o modelo do DM, na abordagem "bottom-up".

A abordagem "top-down" cria os DM a partir do modelo de dados do DW. A abordagem "bottom-up", desenvolve os DM independentemente, realizando a integração do seu modelo de dados ao modelo do DW.

Para auxiliar o processo de modelagem, os desenvolvedores devem se manter

atentos com relação aos seguintes itens:

- Estabelecer o escopo do ADW. Este escopo se refere à definição dos componentes que integrarão o ambiente, à seleção da abordagem de desenvolvimento a ser empregada e ao tipo de modelo a ser utilizado em cada um dos repositórios do ambiente.
- Garantir que a modelagem de dados do ADW represente a visão negócio. Deve ser possível extrair do DW ou dos DM, as visões multidimensionais solicitadas pelo usuário final;
- Integrar as fontes de dados dos sistemas operativos e fontes externas. É importante observar que a otimização da integração, implica diretamente em acelerar a disponibilidade dos dados para as consultas dos usuários finais (resultados);
- Garantir a disponibilidade das informações finais para aplicativos, ferramentas OLAP e “Data mining”. Para isso a modelagem dos dados deve se preocupar com questões quanto a facilidades de uso e a transformação dos dados; e
- Garantir que a modelagem do ambiente seja capaz, de rapidamente, ajustar-se a mudanças nos negócios.

A seguir serão apresentadas as abordagens de modelagem de dados para DW e DM existentes na literatura atual. A modelagem do DW é apresentada segundo a abordagem de Kimball, ou seja, empregando um modelo dimensional, e segundo a abordagem de Inmon, nesse caso empregando um modelo relacional tratado. A modelagem empregada no DM, a ser discutida aqui se baseia no modelo dimensional, empregando as mesmas técnicas descritas para o DW dimensional.

3.5.1 A Modelagem do DW

Através da modelagem do DW é possível selecionar as partes que representam o negócio e avaliar se o escopo da informação está completo. Este conhecimento prévio permite uma base sólida para a geração do DW, facilitando futuras adições e permitindo um controle maior sobre as redundâncias.

Construir o DW é o processo de combinar as necessidades de informações de uma comunidade de usuários com os dados que realmente são disponíveis, sempre questionando se as necessidades mais importantes da organização estão sendo atendidas

da forma mais eficiente. Os principais requisitos de um DW são (KIMBALL, 1996) (GOLFARELLI, DECIO, RIZZI, 1998) (MEYER, CANNON, 1998) (INMON, 1998):

- Fornecer acesso a dados corporativos ou organizacionais. O acesso é responsável por possibilitar o desenvolvimento de ferramentas fáceis de serem utilizadas, com um bom desempenho nas consultas. O propósito é garantir a disponibilidade de informações relevantes para o apoio e tomada de decisão;
- Garantir a integração dos dados de fontes heterogêneas;
- Garantir a consistência e qualidade dos dados. Esta garantia pode ser obtida através da introdução de mecanismos que permitam avaliar se as informações disponíveis estão ou não prontas para serem utilizadas. Por exemplo, dispor de um mecanismo que avise que determinada filial deixou de enviar o malote diário;
- Garantir facilidade de manipulação pelo usuário final, que deve ser a mais intuitiva possível;
- Ser compatível com ferramentas para consulta, análise e apresentação de informações; e
- Compartilhar dados.

Um dos aspectos mais importantes da modelagem de um DW está em definir a sua granularidade. Não existem restrições quanto ao armazenamento de dados primitivos no DW, mas esse tipo de armazenamento não é muito comum. A granularidade de um DW está relacionada ao volume de dados a serem mantidos e consultados. Aumentar o nível de granularidade implica em restringir o nível das consultas. Entretanto, com um baixo nível de granularidade, o espaço de armazenamento aumenta e um número maior de questões passam a ser atendidas (KIMBALL, 1996).

Existem duas linhas de atuação para essa modelagem. Estas linhas estão diretamente relacionadas ao seu emprego: a primeira linha, dos seguidores de Kimball, denominada visão Kimballiana, prega a utilização do DW, também, para fornecerem dados para os DM, porém com um maior enfoque para acesso por consultas. A segunda linha, dos seguidores de Inmon, denominada visão Inmoniana, prega a utilização do DW como um grande integrador dos dados a serem utilizados pelos DM para promoverem as consultas necessários aos SSD. Nesta visão, as consultas no DW apresentam um caráter exploratório, empregando, por exemplo, "Data Mining".

3.5.1.1 Visão Kimballiana:

A visão Kimballiana emprega o modelo dimensional, através do esquema estrela, para o desenvolvimento do DW. Para esta concepção, um DW de uma grande empresa deve apresentar em torno de 15 a 20 esquemas, com tabelas de dimensões que possam ser compartilhadas entre os esquemas (KIMBALL, 1996).

O DW é definido pela utilização de estruturas multidimensionais, seja através de um SGBDM ou pela implementação de um esquema estrela em um SGBD relacional. Com esta modelagem, pretende-se atingir os seguintes propósitos :

- Permitir ao usuário final avaliar uma determinada situação sobre diversos aspectos (ângulos), de acordo com a sua necessidade;
- Permitir análises complexas e uma melhor visualização da informação; e
- Disponibilizar a informação em uma estrutura que facilite o trabalho das ferramentas que a manipularão.

Nesta visão, o desenvolvimento do ambiente envolve (KIMBALL, 1996-a) (KIMBALL, 1997):

- Selecionar um processo de negócios para modelar. O processo deve ser uma operação importante na organização e suportado por algum tipo de sistema de onde seja possível coletar dados.
- Selecionar a granularidade do negócio. O grão é o nível atômico que representará esse processo na tabela de fatos. Grãos típicos são transações individuais, instantâneos individuais diários e instantâneos individuais mensais.
- Selecionar as dimensões que serão aplicadas a cada registro da tabela de fato. Para cada dimensão escolhida, deve se descrever todos os atributos de dimensão que preencham cada tabela dimensional.
- Selecionar os atos mensuráveis que irão popular cada registro da tabela de fatos.

Para detalhar o processo, são estabelecidos nove pontos de decisão para o projeto de um DW que consistem em (KIMBALL, 1996-b):

1. Identificar os fatos (tabelas de fatos). Neste ponto são identificados os principais processos da empresa onde os dados serão coletados;
2. Estabelecer a granularidade de cada tabela de fatos. Neste ponto é determinado o

nível de detalhe de interesse para a análise;

3. Descobrir as dimensões de cada tabela de fatos. A partir da granularidade desejada são obtiddas as dimensões primárias. Outras dimensões adicionais poderão surgir ao longo do desenvolvimento. Estas dimensões não são necessárias para a definição da granularidade da tabela de fatos;
4. Obter os fatos, incluindo fatos pré-calculados. Neste ponto serão estabelecidos os fatos mensuráveis;
5. Obter os atributos da dimensão com descrições completas e terminologia apropriada;
6. Rastrear dimensões de modificação lenta. Representa o rastreamento de dimensões que sofram mudanças graduais ao longo dos eixos temporais;
7. Analisar agregados, dimensões heterogêneas, minidimensões, modos de consultas e outras decisões de armazenamento físico;
8. Selecionar a amplitude de tempo do histórico do banco de dados; e
9. Estabelecer os intervalos com que os dados serão extraídos e carregados no DW.

A seguir estão listados alguns problemas desta visão.

- Depende diretamente das necessidades do usuário final, restringindo o DW à visão do usuário. No caso de uma mudança nos negócios, ou redefinição de granularidade, ou alteração de dimensões, o impacto sobre o DW será grande ;
- Esta visão cresceu em ambientes onde os DM eram pouco conhecidos e empregados, sendo a grande maioria das consultas desenvolvidas sobre o próprio DW;
- Apresenta dificuldades na sua implementação. A literatura existente apresenta a busca de informações, dimensões e fatos, como um processo empírico sob a responsabilidade dos projetistas (KIMBALL, 1996);
- Não existe a preocupação em se analisar um modelo de dados seja ele corporativo ou não;

> "Qualquer DER de dados operativo será, de certa forma, útil no estágio da modelagem do DW, mas não influenciará diretamente o processo de identificação das tabelas de fatos ou de sua granularidade. Uma análise ER terá a função de familiarizar a equipe com as complexidades dos dados. Entretanto, se a análise ER dos dados operativos ainda não foi realizada, esse definitivamente não é o momento de interromper o processo para executá-la." (KIMBALL, 1996-b, p.181).

- Os desenvolvedores estão mais interessados nas consultas que este modelo deverá propor em um determinado momento, do que se prepararem para as possíveis consultas que poderão existir. As consultas, atualmente, estão sendo implementadas no nível dos DM; e
- Problemas quanto aos custos com hardware e com o desempenho das consultas. Existem muitas discussões quanto a disponibilidade de hardwares capazes de armazenar e manipular uma grande quantidade de informações. Deve ser levado em consideração que estes hardwares implicam em um vultuoso investimento, sem, a princípio, um retorno garantido. Nesse sentido, avaliar a capacidade de máquina e manter um bom gerenciamento das informações armazenadas, são uma característica importante. A normalização pode, até mesmo, se tornar viável em alguns casos.

3.5.1.2 Visão Inmoniana:

A visão Inmoniana se preocupa em garantir que o usuário final compreenda que o DW não é um banco de dados corporativo. Esta visão prega a utilização do modelo corporativo da empresa como ponto de partida da modelagem que permite (INMON, 1997) (DEVLIN, 1997) (SILVERSTON, INMON, GRAZIANO, 1997) (MEYER,CANNON, 1998):

- Uma visão genérica dos dados do negócio;
- Facilidade de crescimento; e
- Desenvolvimento em paralelo de diferentes necessidades do negócio.

Atualmente, se propõe um investimento inicial para a construção ou atualização dos modelos de dados existentes, antes de se iniciar a modelagem do ADW propriamente dito (MEYER, CANNON, 1998).

A visão Inmoniana trata a modelagem do ADW com 3 modelos distintos:

- Modelo de dados corporativo, representando uma visão de alto nível dos assuntos e áreas de interesse da empresa. Este modelo serve como base para o desenvolvimento dos sistemas operativos e do DW;
- Modelo do DW, representando a modelagem que serve como fonte única do ADW; e
- Modelo de DM, utilizado para manter as informações departamentais, extraindo-as do modelo de dados do DW.

Esta abordagem implementa o DW em um grande banco de dados relacional, que centraliza todos os níveis departamentais da companhia. Os DM são, entretanto, implementados em SGBDM, para atender suas características de análises dimensionais (INMON, HACKATHORN, 1997) (DEVLIN, 1997). Os DM e o DW apresentam um relacionamento interessante e complementar (INMON, 1997). O nível de dados existente no DW proporciona uma fonte de dados mais robusta para os DM. Além disso, permite um nível de análise em menores granularidades para qualquer usuário dos DM.

Ao contrário da visão Kimballiana, esta visão parte do princípio de que as características que tornam o SGBDM excelente para o DM, não são aquelas de vital importância para o DW (INMON, 1998). Do mesmo modo, as características mais importantes do DW, não são as mais importantes para o DM. Esta abordagem apresenta as seguintes vantagens:

- Realiza a separação dos tipos de implementação. Garante o desempenho das consultas nos DM e de atualização no DW. Um exemplo é a análise de vendas que inicialmente foi solicitada como mensal, passar a ser semanal. Se esta for uma decisão a nível de DM, basta buscar os dados no DW com a nova visão;
- Não emprega o esquema estrela para o DW por não considerá-lo genérico o suficiente. Entretanto, o considera ideal para o DM (DEVLIN, 1997); e
- Permite um bom desempenho. O desempenho das consultas no DM, com seus modelos dimensionais, e uma melhor estrutura de armazenamento para o DW, com o modelo relacional.

A abordagem para a transformação do modelo corporativo para o modelo de dados de DW, relacional-transformado é apresentado por Silverston (SILVERSTON, INMON, GRAZIANO, 1997). Esta abordagem emprega as seguintes transformações para a geração do DW:

1. Descobrir os assuntos e áreas de interesse
2. Construir/avaliar os modelos lógicos para cada área identificando:
    - Entidades;
    - Chaves das entidades;
    - Atributos;
    - Sub-tipos; e
    - Relacionamentos entre entidades.

3. Efetuar a transformação:
    - Remover os dados puramente operacionais;
    - Adicionar o elemento tempo à chave da estrutura do DW, caso ela não o possua;
    - Adicionar dados derivados apropriados;
    - Transformar relacionamentos em artefatos de dados;
    - Acomodar os diferentes níveis de granularidade encontrados no DW;
    - Efetuar o merging dos dados de tabelas semelhantes;
    - Criar array de Dados; e
    - Separar atributos de dados de acordo com suas características de estabilidade.

    A seguir são listados alguns problemas da visão Inmoniana.

- Ao se aplicar uma modelagem no DW diferente da existente nos DM, as consultas que necessitam de um nível mais detalhado, são mais trabalhosas de serem produzidas;
- Dificuldade de implementação. Como em várias literaturas, o que fazer é sempre fácil de se encontrar, porém como fazer, continua a ser superficialmente abordado;
- As regras apresentadas anteriormente foram estabelecidas para uma área de interesse, o que provavelmente identifica um DM. Portanto, qual o procedimento ao se desenvolver um novo DM? O que fazer com as informações já existentes no DW? Como integrar os ambientes de forma harmônica?

3.5.2 A Modelagem do Data Mart (DM)

O DM é projetado para atender às necessidades de um setor ou departamento. Em uma empresa são elaborados DM para setores diferentes, com objetivos diferentes (INMON, 1998). Cada DM, ao final de seu desenvolvimento, é destinado a atender a um conjunto de necessidades e aplicações. Pouca atenção é dada às diferenças entre os objetivos de projeto para um determinado DM e os objetivos para corporação. Normalmente, os objetivos serão diferentes, pois o DM é específico, enquanto o DW, responsável pela corporação, deve ser genérico. A modelagem de um DM requer uma boa definição do seu escopo. Essa definição é obtida descobrindo-se, dentre outros, os problemas e as necessidades que o DM irá atender, as formas usuais de utilização da informação pelo usuário final e os benefícios que a empresa espera receber (DYCHÉ,

1998).

Os principais requisitos de um DM são (KIMBALL, 1996) (MEYER, CANNON, 1998) (GOLFARELLI, DECIO, RIZZI, 1998) (INMONl, 1998):

- Fornecer acesso aos dados de uma área de interesse;
- Garantir a consistência dos dados. A melhor forma de se garantir a consistência dos dados é obtendo-os do DW. Entretanto, os DM independentes recebem suas informações diretamente dos sistemas operativos. Para esses casos, a garantia da consistência é um processo mais complicado;
- Garantir facilidade de emprego dos dados pelo usuário final. A manipulação dos dados deve ser a mais intuitiva possível. Os termos devem ser apresentados do modo como são empregados pelo usuário final;
- Permitir as consultas padrões para um modelo multidimensional, como por exemplo, "Slice & Dice", "Drill-down" e pivoteamento; e
- Ser compatível com as ferramentas para consulta, análise e apresentação de informações.

Para o desenvolvimento dos DM existem duas propostas de modelagem: a modelagem a partir do DW e a modelagem a partir dos sistemas operativos. A modelagem de dados do DM derivada a partir do DW garante a consistência e o mapeamento dos dados. A modelagem direta a partir dos sistemas operativos pode, eventualmente, gerar DM independentes. Os DM não serão independentes, nos casos em que exista um controle e gerência dos metadados, durante o processo de criação do DM e da integração com um DW. Os DM independentes não precisam de um DW, sendo desenvolvidos diretamente a partir dos sistemas existentes. Para a modelagem, normalmente é empregado o esquema estrela.

3.6 Questões da Modelagem de Dados

Atualmente, observa-se na literatura o emprego de técnicas de modelagem dimensional para o desenvolvimento de DW ou DM, a partir das necessidades dos usuários finais. Esta abordagem é, normalmente, comparada ao desenvolvimento de sistemas de informação para o ambiente operativo. Porém, como foi apresentado no capítulo 2, muitas são as diferenças entre esses ambientes, principalmente quanto aos

repositórios de dados. Esse tipo de abordagem pode provocar demoras no mapeamento das informações entre o ADW e os sistemas operativos, após a modelagem.

Por outro lado, a abordagem empregando o modelo corporativo como modelo conceitual para o ADW, normalmente é apresentada com a abordagem padrão, modelando primeiramente o DW para a partir dele modelar os DM. Esta abordagem, apesar de conceitualmente correta, não é a recomendada para o desenvolvimento deste ambiente (Capítulo 2). Porém, até a presente data, a literatura especializada não possui uma proposta de modelagem incremental para este ambiente que empregue o modelo de dados existentes no ambiente operativo, servindo de base para a modelagem dos DM.

Analisando os fatos mencionados, é possível observar uma tendência em aplicar na modelagem do ADW as mesmas considerações do ambiente operativo, como, por exemplo, os três níveis de modelagem (modelo conceitual, modelo lógico e modelo físico). Infelizmente, a modelagem inicial deste ambiente surgiu com o esquema estrela, que representa a modelagem lógica e física. Propostas de desenvolvimento de uma modelagem conceitual orientada estão em estudo como, por exemplo, a proposta de Golfarelli (GOLFARELLI, MAIO, RIZZI, 1998), sem no entanto nenhuma garantia de seu emprego ou aceitação.

De uma forma geral, os modelos de dados existentes para os sistemas do ambiente operativo refletem as informações operacionais e as informações do negócio. Portanto, através de transformações aplicadas a esses modelos, é possível remover as características de processo e estabelecer um modelo orientado ao negócio que possa ser aplicado ao ADW. Um dos problemas relacionados ao emprego desses modelos no ADW, segundo Kimball (KIMBALL, 1995) (KIMBALL, 1997), ocorre pela falta de expressividade desses modelos para o usuário final. É muito mais fácil um usuário entender um esquema estrela do que um DER. Entretanto, refinando um DER e delimitando o seu escopo para representar um fato de interesse, seria ele tão incompreensível?

A modelagem do ADW deve ser feita passo a passo, de modo iterativo. A modelagem de áreas de interesse é recomendada com o propósito de reduzir o escopo de informações sobre o negócio a ser analisado.

Dentre as diferentes propostas de modelagem para o ADW, os seguintes itens merecem destaque:

- Falta de profundidade. Normalmente a apresentação da modelagem é feita de forma superficial, a critério do autor;
- Controvérsias com a modelagem dimensional. Um dos principais pontos de controvérsia do emprego da modelagem dimensional, sem uma análise de DER, está no grau de julgamento colocado nas mãos do projetista. Nesta abordagem, cabe ao projetista estabelecer o conjunto mais natural de dimensões para atender ao usuário final (KIMBALL, 1995);
- Discussões sobre o uso do modelo corporativo da empresa. Propostas, como a de Kimball (KIMBALL, 1997) (KIMBALL *et al*, 1998), restringem muito o acesso ao modelo corporativo, não lhe dando devida importância. O apoio total sobre ele, sem nenhuma limpeza de dados, pode ser considerado um outro extremo;
- Soluções pontuais. As propostas se referem à modelagem de DM ou modelagem de DW. Não se observa uma proposta em que se aborde a modelagem do ambiente;
- Divergências de abordagens. Apesar da literatura apresentar a abordagem "bottom-up" como sendo a mais indicada para o desenvolvimento dos repositórios deste ambiente, o que se observa na literatura são técnicas/práticas para a modelagem de DW e posteriormente de DM;
- Atualizações estruturais do DW. A literatura, normalmente, apresenta o processo de integração como trivial (KIMBALL *et al*, 1998). Entretanto, uma avaliação mais criteriosa demonstra que, para se garantir consistência e integridade, o tratamento da integração deve ser feito através de um rigoroso processo. O impacto das alterações sobre DM existentes e sobre os programas de carga e atualização são raramente mencionados;
- Modelagem primária. É comum observar as modelagem tratando um primeiro DM. Porém, como tratar a geração dos próximos? Como lidar com o processo de integração? Estes itens não são mencionados.

CAPÍTULO 4

DIRETRIZES PARA A MODELAGEM INCREMENTAL DE UM AMBIENTE DE DATA WAREHOUSE

O sucesso ou o fracasso da implantação do ADW está associado diretamente a estratégias e arquiteturas da modelagem de dados e ao projeto do DW (MEYER, CANNON, 1998) (KIMBALL, 1996). As metodologias para o desenvolvimento de um ADW, normalmente, apresentam um enfoque superficial com relação a modelagem de dados. Apesar da extensa documentação sobre abordagens e arquiteturas, o que normalmente se observa são abordagens estanques. As técnicas de Kimball (KIMBALL, 1996) para a modelagem dimensional e as transformações de Silverston (SILVERSTON, INMON, GRAZIANO, 1997) para gerar um DW a partir de modelos corporativos representam esse problema.

O propósito deste capítulo é estabelecer diretrizes que permitam uma modelagem incremental do ADW, empregando a abordagem "bottom-up". Essas diretrizes permitem a especificação de DM e do DW, a partir do modelo corporativo. Para a aplicação dessa modelagem, as seguintes considerações são necessárias:

1. Arquitetura: O conjunto de regras está orientado ao desenvolvimento de um ADW de modo incremental;
2. Componentes do ambiente: O ambiente é composto de DW e de DM. É necessário estabelecer a previsão de histórico e a freqüência com que os dados devam ser extraídos e carregados no ADW.
3. Modelos a serem empregados nos repositórios:
   - Dimensional para os DM;
   - Independência de modelo a ser adotado a nível de modelagem do DW. As regras de integração foram desenvolvidas para a visão Kimball e para a visão Inmom.

O conjunto de diretrizes proposto utiliza, em um regime de colaboração, duas abordagens de modelagem existentes na literatura: o emprego de técnicas de modelagem dimensional, com base na especificação do usuário final, para o desenvolvimento de DW e DM; e o desenvolvimento de DW a partir do modelo

corporativo da empresa.

O propósito do conjunto de diretrizes é realizar a modelagem do ambiente de forma incremental (DM→ DW), empregando os modelos de dados existentes no ambiente operativo (o modelo corporativo ou DER isolados). Estes modelos são transformados, permitindo a derivação de um esboço de modelo dimensional. Sobre este esboço são aplicadas técnicas de modelagem dimensional, refinando o modelo. Dessa forma, se estabelece um processo mais sistêmico, reduzindo a dependência do projetista.

Este capítulo apresenta inicialmente uma visão alto nível do processo de modelagem, seguida da definição do pré-modelo – fundamental nessa proposta, e da apresentação do conjunto de diretrizes propostas. Para a apresentação dessas diretrizes, a modelagem do ADW foi dividida em quatro fases. Essas fases encontram-se, em alguns casos, subdivididas em subfases, etapas e subetapas.

4.1 O Processo da Modelagem

O processo de modelagem de dados se inicia com a análise dos modelos de dados existentes no ambiente operativo, sendo ideal a utilização do modelo de dados corporativo da empresa. Algumas empresas, porém, não possuem esses modelos para os sistemas antigos, ou mesmo apresentam o próprio modelo corporativo dividido em modelos menores, em virtude de seu porte. Para estes casos é necessário analisar os DER dos sistemas relacionados à área a ser incorporada ao ADW.

A análise destes modelos permitirá uma melhor base para o conhecimento do negócio e para um ágil desenvolvimento do ADW. Através da análise do(s) DER é elaborado um "pré-modelo de dados" responsável pela integração entre o ambiente operativo (visão processo) e o ADW (visão negócio). O(s) modelo(s) multidimensional(is), que constitui(rão) o DM, é derivado a partir deste "pré-modelo". Com a conclusão do DM, efetua-se a integração do(s) novo(s) modelo(s) ao modelo do DW.

Para melhor tratar a questão de integração entre o ambiente operativo e o ADW, as diretrizes do processo de modelagem foram divididas em quatro (4) fases. As seguintes fases compõem a modelagem:

FASE A - Estudo dos Modelos Existentes : Esta fase é responsável por delimitar o escopo do modelo corporativo ou dos DER relacionados aos sistemas existentes. Este escopo representa a área de interesse para a análise. Para se

iniciar esta fase, as seguintes condições devem ser atendidas:

- área de interesse definida pelo usuário final;
- necessidades de análise. Quando não houver uma definição formal, é necessário o levantamento dos tipos mais comuns de análises solicitadas pelo usuário final; e
- modelo corporativo ou DER de sistemas existentes.

FASE B - Elaboração do pré-modelo: Esta fase é a responsável por analisar e integrar os DER dos sistemas operativos que compõem a área de interesse. Essa integração é responsável pela criação de um pré-modelo. O pré-modelo se caracteriza por ser um DER não normalizado, com características peculiares, conforme será abordado na seção 4.2 (O Pré-Modelo). Para a realização desta fase é necessário que o projetista apresente um conhecimento razoável do negócio, que lhe permita tomar decisões com o auxílio, quando for o caso, de ABD e usuários finais.

FASE C - Modelagem do DM: Esta fase é a responsável por estabelecer um conjunto de visões dimensionais. Após a validação das visões dimensionais, pelos usuários finais, os modelos dimensionais serão derivados, a partir do "pré-modelo", empregando o esquema estrela. O modelo dimensional do DM é representado pela composição dos modelos dimensionais desenvolvidos para atender às visões dimensionais. Para a realização desta fase, o projetista deve avaliar criteriosamente as necessidades levantadas pelo usuário final e possuir o pré-modelo.

FASE D - Integração do DM ao DW: Esta fase corresponde à atualização do DW, de acordo com a construção de um novo DM. Representa o desenvolvimento incremental do DW. Como dois modelos estarão disponíveis, o pré-modelo e o modelo dimensional do DM, a atualização do DW pode ser realizada adotando a abordagem Inmon ou a abordagem Kimball. Esta fase apresenta também considerações sobre as mudanças estruturais usuais durante o desenvolvimento de um ADW.

As fases estabelecidas para a modelagem de dados estão representadas na figura 4.1

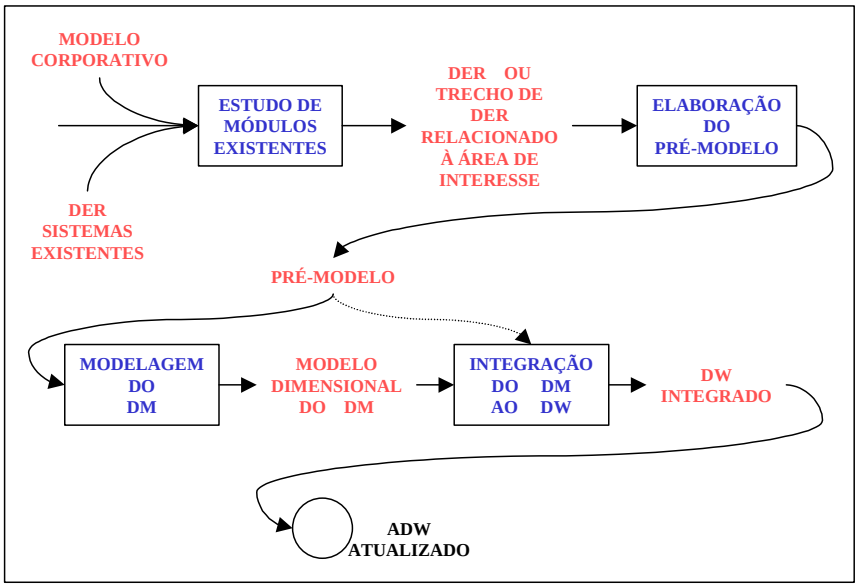

Figura 4.1 - Fases da Modelagem do ADW

4.2 O Pré-Modelo

Segundo Gupta, “As questões relacionadas com a transformação lógica de dados de um sistema operativo para um data warehouse podem requerer considerável análise e esforço.”(GUPTA, 1997, p.13).

O pré-modelo, nessa abordagem, representa a integração dos diversos modelos de dados dos sistemas existentes, representados por um DER, que estejam envolvidos com a área de interesse. O propósito do pré-modelo é reduzir o esforço empregado no processo de transformação dos dados do ambiente operativo para o ADW. Sua construção apresenta, dentre outras, preocupações quanto à:

- Eliminação das informações operativas, entrando nesta questão não apenas os atributos, mas também as entidades empregadas apenas para o processamento normal do sistema;
- Desnormalização de entidades que possam ser tratadas em conjunto e daquelas que apresentem dependência de existência;
- Criação de artefatos, substituindo relacionamentos dos modelos existentes, cuja informação de interesse seja aquela extraída no momento da atualização do DW ("Snapshot"); e

- Criação de atributos derivados que possam ser considerados para o DW.

A necessidade do pré-modelo independe da abordagem a ser adotada para o ADW. Sua finalidade é integrar as diversas informações existentes na empresa em um nível conceitual, antes de desenvolver um modelo dimensional (nível lógico), como normalmente acontece quando se emprega o esquema estrela. O que, a princípio, pode ser considerado um volume maior de trabalho, tem como vantagem o domínio sobre o assunto abordado pelos projetistas. Essa familiaridade com o negócio permite um levantamento de visões multidimensionais até então não consideradas pelo usuário final, por esquecimento ou simplesmente pelo desconhecimento das possibilidades associadas às informações disponíveis para acesso.

4.3 Fases da Modelagem

A seguir será realizado o detalhamento de cada uma das fases da modelagem de dados. Um esquema apresentando o conjunto de diretrizes por fase é mostrado na figura 4.2.

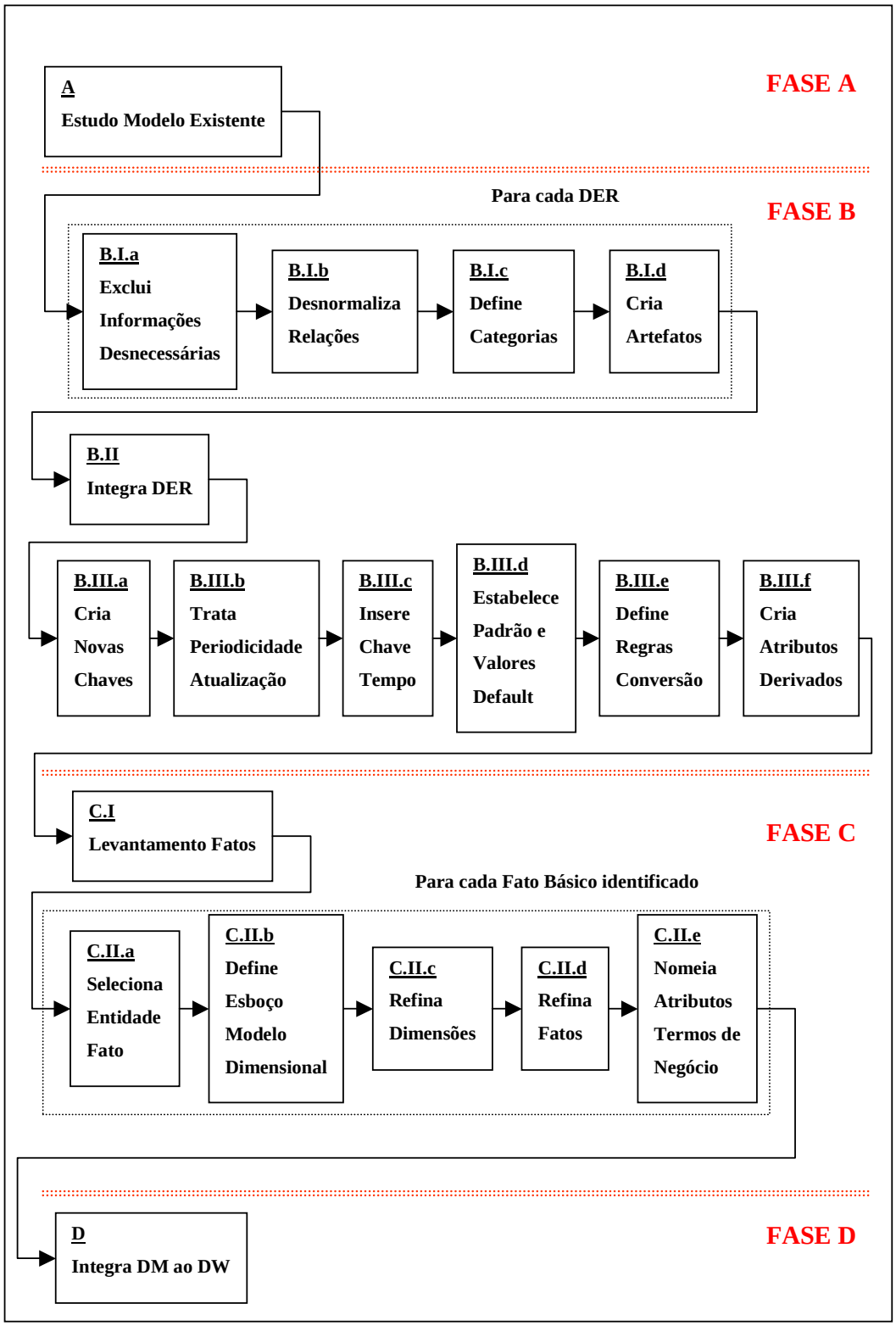

Figura 4.2- Diretrizes para a Modelagem de um ADW

4.3.1 **FASE A - ESTUDAR OS MODELOS EXISTENTES**

Esta fase é responsável por definir o escopo dos modelos de dados a serem analisados. O Estudo dos modelos existentes pode ocorrer de duas formas: estudo do modelo de dados corporativo e estudo dos DER referentes aos sistemas operativos relacionados à área de interesse. Quando for necessário empregar fontes externas e não existir um modelo de dados pronto, este modelo deverá ser criado e passará a ser tratado como um DER dos sistemas existentes.

O estudo do modelo corporativo, consiste em, a partir dele, selecionar as informações de interesse para empresa. Essa seleção é realizada por meio de reuniões entre os projetistas, os usuários finais e os ABD.

Para as empresas que possuam DER dos sistemas existentes ou que apresentem o modelo corporativo desmembrado em modelos menores, a seleção das informações de interesse é realizada em cada modelo.

O resultado desta fase é o escopo do modelo corporativo ou dos de DER dos sistemas operativos, relacionados à área de interesse.

Um exemplo de DER simplificado será apresentado no estudo de caso no capítulo 5.

4.3.2 **FASE B – ELABORAR O PRÉ-MODELO**

Esta fase recebe as entidades/relacionamentos selecionados a partir do DER Corporativo, ou dos DER dos sistemas operativos relacionados à área de interesse, estabelecidos na fase anterior. Nesta fase os DER são analisados, transformados e integrados no pré-modelo. Esta fase encontra-se dividida em três subfases:

- LIMPAR E TRANSFORMAR OS DER: responsável pela análise de cada DER, isoladamente. Esta etapa extrai de cada DER apenas as informações que sejam de interesse para a solução do problema;
- INTEGRAR OS DER: responsável por integrar os DER resultantes da etapa anterior. Só é aplicável nos casos onde existam mais de um DER para a análise;
- REFINAR O DER: responsável por tratar o DER resultante, criando novas chaves e tratando os atributos. O produto final desta subfase é o pré-modelo.

Ao final desta fase, ter-se-á a definição do pré-modelo. Como mencionado, um modelo desnormalizado e transformado, integrando as informações relacionadas com o problema em questão, em uma visão bem mais simples. A análise sobre os modelos do ambiente operativo permite, em um primeiro momento, que o projetista tenha um contato maior com a forma de tratamento da empresa, com suas informações. Conhecendo a forma de tratamento, o projetista pode estabelecer o grau de importância da informação no sistema/negócio.

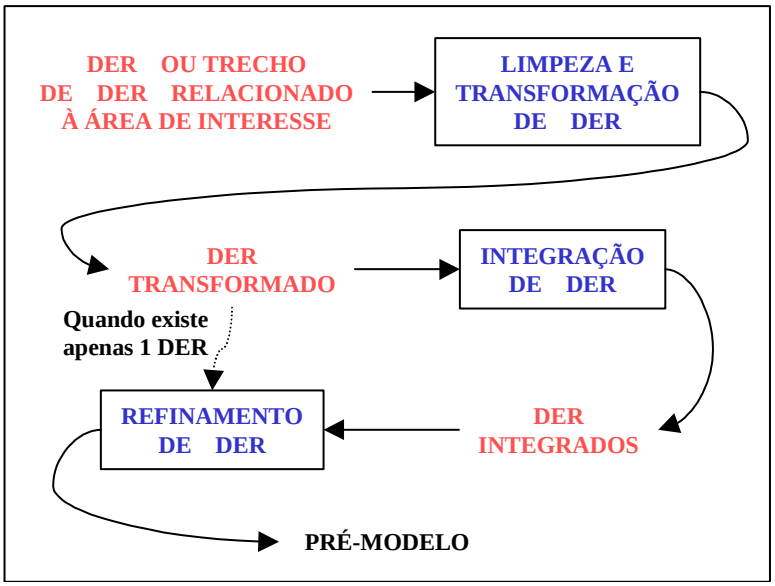

Figura 4.3 - Subfases da Elaboração do Pré-Modelo

Nesta fase a nomenclatura empregada é a mesma do DER, com entidades, atributos e relacionamentos. O pré-modelo, como já mencionado, representa um DER tratado contendo apenas as informações de interesse a análise. A figura 4.3 apresenta as subfases referentes a esta fase.

B.I – Limpar e Transformar o DER

Nesta subfase o DER é reduzido a um novo diagrama que apresenta apenas as entidades relacionadas com o problema em questão. A redução é realizada pela exclusão das informações que não representam interesse para análise, normalmente, as informações operativas. A identificação dessas informações é feita pela análise a nível de entidades e a nível de atributos.

B.I.a) Excluir Informações Desnecessárias:

Esta etapa tem o propósito de remover as informações não relacionadas ao problema, evitando a criação de um pré-modelo sobrecarregado com informações inúteis. A remoção é realizada, primeiramente, retirando-se as entidades que não apresentam informações de interesse ou cujos atributos representem informações

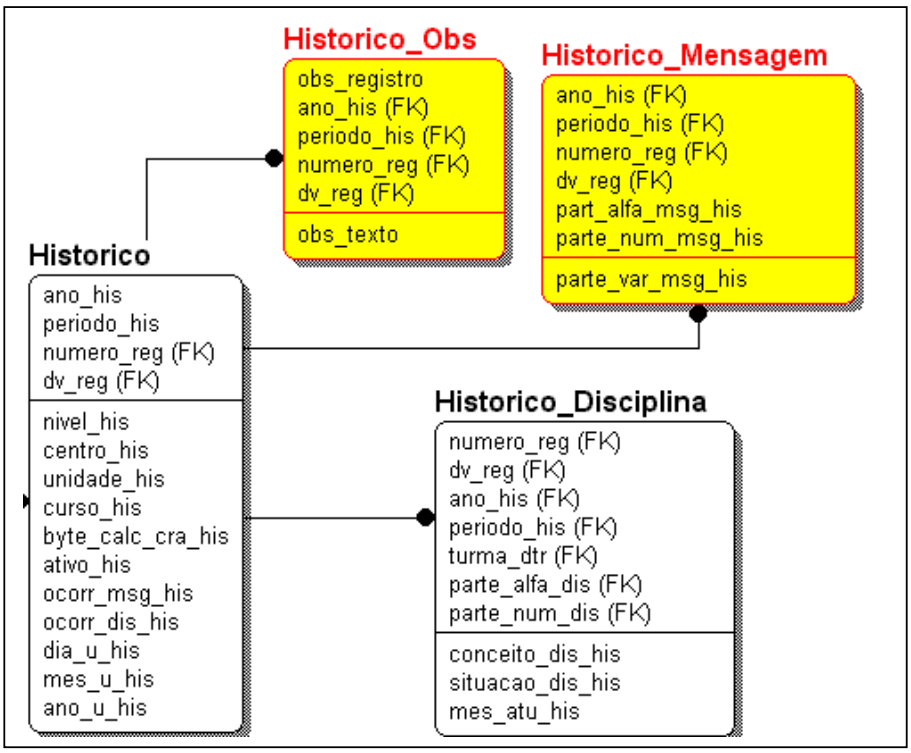

Figura 4.4 - Remoção de entidades operativas

operativas. Para as demais entidades é realizada uma análise atributo a atributo de modo a manter apenas aqueles que representam interesse para a análise.

O processo de exclusão, a nível de entidades, pode adotar os seguintes passos:

- Selecionar as entidades diretamente relacionadas ao problema. Para a universidade, o propósito do DM é avaliar o desempenho dos alunos. Neste caso, uma maior ênfase será atribuída às entidades que trabalham com as informações relacionadas a aluno, disciplina e notas.
- A partir das entidades que tratam a informação de interesse, analisar as que apresentam um relacionamento direto. Este primeiro nível da análise consiste da verificação das entidades diretamente relacionadas. Em um segundo nível, a análise é feita nas entidades relacionadas com aquelas selecionadas no primeiro nível e assim

sucessivamente, até chegar a um nível onde não exista nenhuma informação de interesse. As entidades que não apresentam interesse são removidas do modelo.

As entidades *Historico_Obs e Historico_Mensagem*, representadas na figura 4.4, estão relacionadas à informações adicionais de históricos de alunos. Essas informações não representam interesse para a análise, sendo, portanto, excluídas do modelo.

No nível dos atributos, a remoção é realizada quando os mesmos representam:

- Informações operativas: neste caso, os atributos são utilizados para processamento operativo. As informações operativas estão relacionadas, em sua grande maioria, às transações características dos sistemas operativos. A carga dessas informações no DW pode acarretar problemas de desempenho e análises incorretas, por estarem, constantemente, sendo atualizadas (alta volatilidade). Essa constante atualização dificulta a interpretação dos resultados das consultas, gerando problemas com relação à confiabilidade dos dados armazenados. Desta forma, dados como flags, status e observação devem ser removidos dos modelos existentes; e
- Informações não relacionadas à área de interesse: neste caso os atributos armazenam informações que não são relevantes para a análise. A relevância é estabelecida pelas necessidades do usuário final e deve ser avaliada pelo projetista.

Segundo Silverston (SILVERSTON, INMON, GRAZIANO, 1997), uma abordagem recomendável para verificar se um determinado atributo deve ser selecionado para o ADW, é aplicar a seguinte pergunta:

- "*Qual é a chance deste dado ser utilizado pelo DSS?*".

Uma rígida verificação deve ser realizada para cada atributo a ser excluído, a fim de garantir que sua exclusão não gere informações incompletas e não provoque a perda de informação relevante. A figura 4.5 apresenta a seleção dos atributos operativos da entidade *Histórico.*

A extração dos dados operativos na figura 4.5 não resulta em perda de informações, pelo contrário, garante o tratamento de informações completas, isto é, informações que não estão em transição. É importante, nesse momento, manter um acompanhamento das informações que serão futuramente utilizadas para carga, como: "somente serão carregados para o DW os pedidos com status igual a concluído"

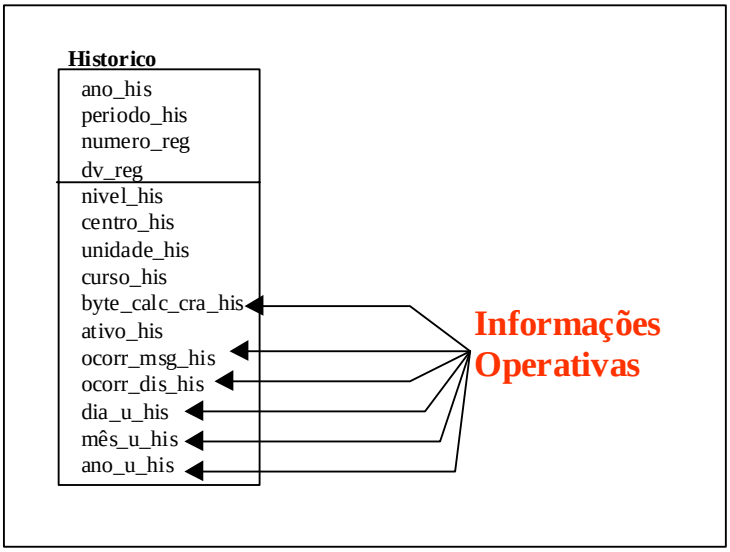

Figura 4.5 – Remoção de Atributos Operativos

B.I.b) Desnormalizar Relações:

Esta etapa tem o propósito de otimizar as consultas e as futuras extrações de informações para os DM, através da desnormalizações de entidades.

O processo de normalização, empregado no ambiente operativo, normalmente desmembra uma entidade em um conjunto de entidades independentes, que se apresentam sem anomalias de atualização (MACHADO, ABREU, 1996). As entidades normalizadas promovem a flexibilidade e não redundância desejadas aos sistemas transacionais do ambiente operativo, mas dificultam e aumentam a complexidade dos sistemas do ADW. No ADW, as tabelas tendem a ser grandes, portanto, qualquer tentativa de reduzir o custo de armazenamento pela normalização, se perde no tempo gasto e na complexidade requerida para a elaboração de consultas. Portanto, o ADW prega a utilização de modelos desnormalizados, facilitando o entendimento pelo usuário final.

A desnormalização reduz a quantidade de junções ("joins") necessárias para a realização de consultas, porque agrega as informações. A desnormalização, entretanto, deve ser aplicada com cuidado, de preferência entre entidades que apresentem relação de dependência de existência como, por exemplo, entre as entidades *Nota_Fiscal* e *Item_NotaFiscal*. A desnormalização entre uma entidade principal e entidades de apoio, também representam um caso típico. Um exemplo é a desnormalização entre a entidade

Cliente e a entidade de apoio *Estado_Civil* e, entre a entidade *Nota_Fiscal* e *Serie_Nota*. Nesses casos, as entidades *Estado_Civil* e *Serie_Nota* representam entidades de apoio. Para os demais casos, é recomendável verificar se as entidades atendem às seguintes condições (SILVERSTON, INMON, GRAZIANO, 1997):

- Compartilham uma chave comum ou parcial;
- Seus atributos (dados) são, freqüentemente, utilizados juntos; e
- Apresentam um padrão de inserção semelhante.

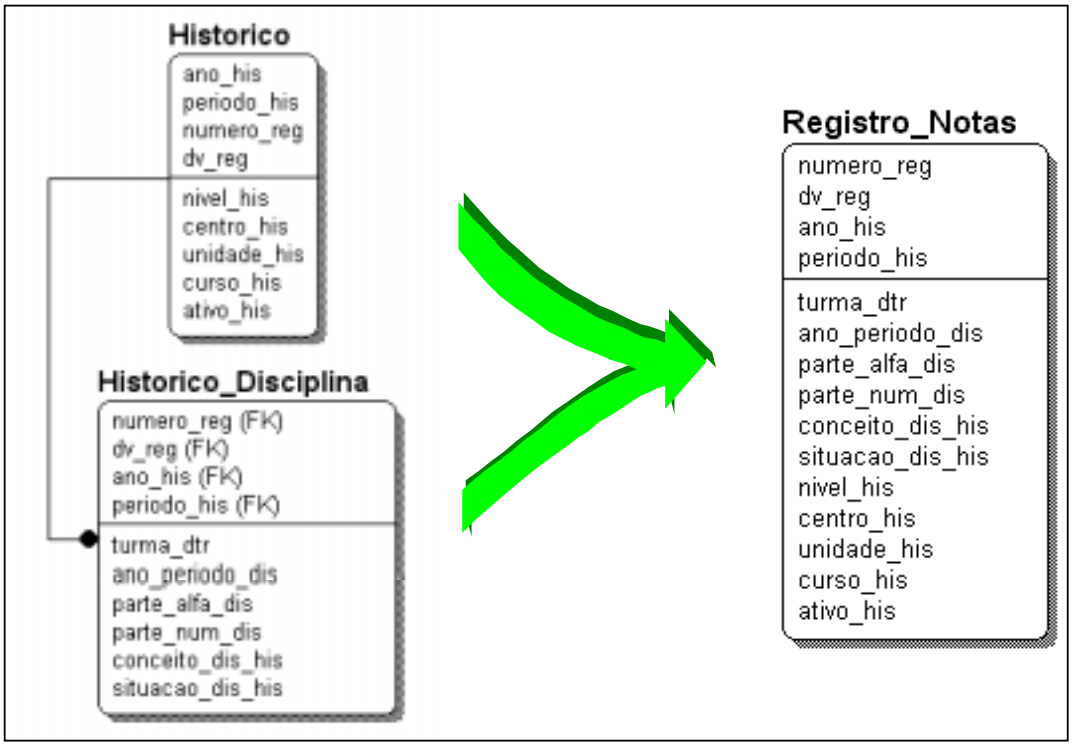

Figura 4.6 – Desnormalização entre entidades

A desnormalização entre uma entidade que represente propriedades estruturais do domínio e que apresente informações quase estáticas com uma entidade freqüentemente atualizada e relacionada aos fatos de interesse não é recomendada. O primeiro grupo de entidades representa fortes candidatas à dimensões, enquanto o segundo grupo representa candidatas a tabelas de fatos.

O exemplo da figura 4.6 representa a desnormalização entre as entidades *Historico_Disciplina* e *Historico* do modelo do SCG.

B.I.c) Definir Categorias:

Esta etapa é responsável por definir categorias de interesse para o ADW. A definição de categorias é um recurso empregado pelo projetista para lidar com informações que podem apresentar uma alta esparsidade para análise, e para criar uma nova hierarquia, de interesse para o DM.

Para tratar as informações que apresentam uma alta esparsidade, o projetista deve identificar os atributos que apresentam valores contínuos, que possam ser substituídos por faixas. Idade e renda representam casos típicos deste tipo de categoria, transformando-se em faixa etária e faixa de renda, respectivamente. Com a transformação em faixas, reduz-se o escopo da análise, favorecendo as consultas dos usuários finais. No modelo da Universidade, por exemplo, o atributo *PONTOS_VEST_REG* será categorizado. A definição dessa categoria é realizada da seguinte forma:

- Substituição do atributo original pelo atributo *FAIXA_PONTOS_VEST*;
- Definição do domínio: "Pontuação inferior a 5000", "Pontuação entre 5000 e 7000", "pontuação entre 7000 e 8000" e "Pontuação acima 8000";
- Definição das regras para seu preenchimento:

  PONTUAÇÃO INFERIOR A 5000 : *PONTOS_VEST_REG* < 5000;

  PONTUAÇÃO ENTRE 5000 e 7000: *PONTOS_VEST_REG* >= 5000 e < 7000;

  PONTUAÇÃO ENTRE 7000 e 8000: *PONTOS_VEST_REG* >= 7000 e < 8000;

  e

  PONTUAÇÃO ACIMA 8000 : *PONTOS_VEST_REG* >= 8000.

O estabelecimento de uma nova hierarquia está associado a criação de um novo atributo. Para este novo atributo, devem ser fixados o domínio e as regras para o seu preenchimento. A necessidade de consultas/relatórios, relacionados a produtos, empregando uma classificação que não seja adotada por nenhum sistema operativo existente é um exemplo deste tipo de categoria. A definição da categoria é realizada na seguinte ordem:

- Criação do atributo *TIPO_ELETRODOMESTICO* na entidade *Produto*;
- Definição do domínio: "Eletrodomésticos Grandes", "Eletrodomésticos Médios" e "Eletrodomésticos Pequenos"; e
- Definição das regras a serem aplicadas na entidade *Produto*, estabelecidas pelo

usuário final:

ELETRODOMÉSTICO GRANDE : Tipo_Prod = 'Eletrodoméstico' e
Peso > 5 quilos;
ELETRODOMÉSTICO MÉDIO : Tipo_Prod = 'Eletrodoméstico' e
Peso > 500 gramas e <= 5 quilos;
ELETRODOMÉSTICO PEQUENO : Tipo_Prod = 'Eletrodoméstico' e
Peso <= 500 gramas.

A regra definida deve ser armazenada para facilitar posteriores alterações, evitando a perda de consistência dos dados históricos já armazenados e, permitindo ao usuário final identificar o momento de alteração de uma regra. Um exemplo de acompanhamento de alteração de regras, seria a criação de um novo domínio "Eletroméstico de Porte". Este novo domínio cria uma nova regra e altera a regra de "Eletrodoméstico grande" da seguinte forma:

ELETRODOMÉSTICO PORTE : Tipo_Prod = 'Eletrodoméstico' e
Peso > 10 quilos;
ELETRODOMÉSTICO GRANDE : Tipo_Prod = 'Eletrodoméstico' e
Peso > 5 quilos e <= 10 quilos;

Os agregados e as informações existentes no ADW antes dessa alteração serão mantidos. As novas cargas e atualizações passarão a empregar a nova regra. Sem um gerenciamento dessas alterações o usuário final pode não entender as mudanças em seus resultados.

B.I.d) Criar Artefatos:

Esta etapa é a responsável por estabelecer as informações do DER que serão representadas como artefatos no ADW. A criação dos artefatos deve ser criteriosa quanto a relevância da informação sendo transformada.

Os artefatos são um recurso do projetista para transformar um relacionamento do ambiente operativo em uma informação de interesse no ADW. Alguns autores (SILVERSTON, INMON, GRAZIANO, 1997) consideram o artefato como uma conseqüência da característica do ADW de armazenar históricos, e não apenas a posição atual, como os bancos de dados operativos. Dessa forma, os artefatos representam a parte de um relacionamento, que seja de interesse no momento da extração dos dados do

ambiente operativo para o ADW – "SNAPSHOT". O artefato pode incluir chaves estrangeiras e outros dados relevantes, tais como atributos de entidades associadas, ou pode incluir somente os dados relevantes, sem incluir as chaves estrangeiras.

Um exemplo deste tratamento é a transformação de *Ajuda_Custo* em um artefato de *Aluno* no modelo UNIVERSIDADE. A entidade *Ajuda_Custo* não apresenta atributos que interessem a análise, porém é importante saber se um aluno recebe ou não ajuda de custo. Dessa forma, a entidade *Ajuda_Custo* transforma-se no artefato *RECEBE_AJUDA*. Esse novo atributo de *Aluno* pode assumir os valores SIM/NÃO. A regra para a definição do valor é : Se existir registro em *Ajuda_Custo* para um registro de *Aluno* então assumir "**SIM"** caso contrário, assumir "**NÃO**".

Outro exemplo de criação de artefato, ocorre entre as entidades *PRODUTO* e *PRODUTOR*, desde que o relacionamento entre as mesmas seja 1:N, isto é, um produto é produzido por apenas um produtor e um produtor pode produzir mais de um produto. Nesse caso, o artefato "PRODUTO_PRODUZIDO_POR" é criado para cada produto, sendo preenchido com o nome do produtor no momento do SNAPSHOT.

As informações transformadas em artefatos, normalmente, apresentam um caráter mais informativo do que analítico. Não existe uma necessidade em se acompanhar a mudança de um artefato ao longo do tempo. Por exemplo, a informação do atributo *RECEBE_AJUDA* será capturada e armazenada, com a situação em vigor no momento do extração das informações do ambiente operativo ("SNAPSHOT"), sem que exista nenhuma preocupação com a informação anterior no ADW.

B.II – Integrar os DER Resultantes

Esta subfase integra os DER tratados, contendo as informações de interesse. O seu resultado final é um modelo intermediário, tratado e orientado para a área de interesse.

A integração é realizada através da criação de um novo diagrama, contendo as entidades e relacionamentos dos DER resultantes da fase anterior. Para os casos onde existam entidades semelhantes, pode ser necessária a aplicação de um "casamento"("merge").

Em algumas referências (SILVERSTON, INMON, GRAZIANO, 1997), o procedimento "casamento" é apresentado para a desnormalização. Entretanto, neste

trabalho ele é empregado com o propósito de combinar duas ou mais entidades semelhantes em uma entidade única para o pré-modelo. Na integração podem ocorrer problemas de conflitos que devem ser tratados. Dentre eles os mais comuns são:

- Atributos de mesmo nome com informações diferentes;
- Atributos de mesmo nome com a mesma informação porém com periodicidade de atualização diferentes;
- Atributos com nomes diferentes, porém com a mesma informação; e
- Atributos com tipos diferentes.

Além dos problemas de conflito, ao realizar o "casamento" de entidades semelhantes, pode ocorrer a necessidade de um novo identificador. Este novo identificador deve ser gerado com base nas regras definidas em B.III.a (Criação de novas chaves).

A integração de entidades semelhantes deve observar os seguintes aspectos:

- As entidades devem apresentar periodicidade de atualização semelhantes. Não é recomendável unir informações que apresentam divergência muito grande quanto à atualização. Essa divergência pode fazer com que um grande volume de dados sejam acessados para realizar uma pequena atualização;
- As entidades devem compartilhar uma chave comum ou parcial;
- As informações das entidades devem ser trabalhadas normalmente juntas; e
- O padrão de tratamento deve ser semelhante. Um exemplo de padrão semelhante é a criação, isto é, sempre que um registro é inserido na entidade 1, será inserido um registro na entidade 2.

Não existe nenhum impedimento, quanto a combinar entidades que apresentem periodicidade de atualização diferentes. Entretanto, os projetistas devem ficar atentos ao estabelecimento de valores "default" para os atributos que não sejam criados no primeiro momento. A definição de valores "default" será tratada, com mais detalhes, na seção B.IV.b (Estabelecer Valores "Default") .

A figura 4.7 apresenta o casamento("merge") entre a entidade *Cliente* do DER de marketing e a entidade *Cliente* do DER de vendas.

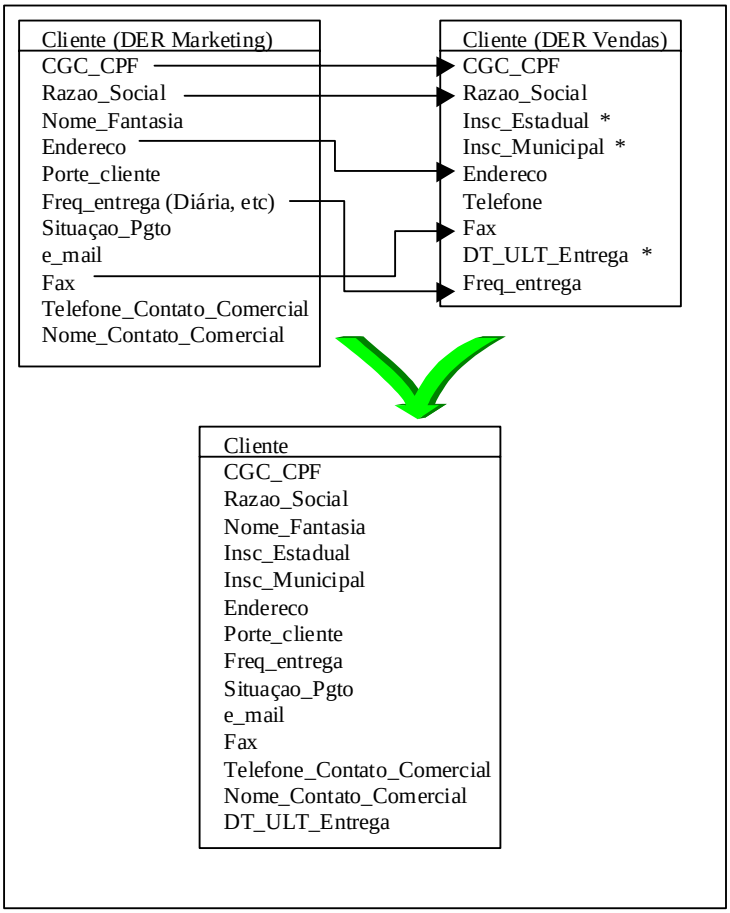

Figura 4.7 - Casamento de entidades Clientes

(*) Atributos de Cliente de Interesse no DER Vendas

B.III – Refinar o DER

O propósito desta subfase é tratar o DER resultante da fase anterior, transformando-o no pré-modelo.

A preocupação das subfases anteriores foi o tratamento a nível conceitual dos dados. Esta subfase recebe um modelo "conceitual" que representa uma composição de modelos conceituais do ambiente operativo, que foram devidamente tratados para representar apenas as informações de interesse. Nesta fase serão abordados aspectos relacionados à modelagem lógica, como a criação de chaves e definição de padrões. A partir desse momento, as informações referentes a modelo lógico, existentes no ambiente operativo para os atributo do DER intermediário são importantes. O emprego de uma ferramenta CASE agiliza esse processo, porque normalmente apresenta a

modelagem completa, ou seja, o modelo conceitual, o modelo lógico e o modelo físico. Dessa forma, enxergar as informações do modelo lógico é uma tarefa simples. Apenas as informações referentes a artefatos e categorias, criadas durante as subfases anteriores, não apresentarão informações no nível lógico. Essas informações merecem uma atenção especial.

B.III.a) Criar Novas Chaves:

O propósito desta etapa é identificar se uma entidade necessita de uma nova chave.

Dessa forma, o projetista deve verificar se as chaves existentes nas entidades requerem um remapeamento. O remapeamento de uma entidade é recomendado quando se observa as seguintes situações:

- Existe a possibilidade de reutilização de chaves no ambiente operativo;
- A entidade apresenta chaves longas, podendo prejudicar o desempenho nas consultas. Este remapeamento é empregado no modelo da Universidade para substituir as chaves de algumas entidades como, por exemplo, da entidade aluno. No ambiente operativo a chave de aluno é representada pelos atributos *NUMERO_REG* e *DV_REG*. Estes atributos serão substituídos pelo atributo *CH_ALUNO*, formado pela combinação dos atributos chaves do ambiente operativo;
- Existe a necessidade de generalização da chave primária, para acompanhar as modificações nos atributos sem realizar alterações nas chaves.

A seção 3.4.3 (Criação de Novas Chaves) apresenta em maiores detalhes as necessidades para a criação de chaves.

Ao ser detectada qualquer uma das situações acima, uma nova chave será criada. Essa nova chave deverá ser registrada, incluindo sua regra de formação.

B.III.b) Tratar Atributos de Acordo com sua Periodicidade de Atualização:

O propósito desta etapa é analisar como os atributos de uma entidade se comportam ao longo do tempo. O tratamento de tempo pode ser considerado um dos mais importantes para o ADW. Portanto, é necessário estabelecer a periodicidade de atualização dos atributos e a importância de seu mapeamento, de modo a garantir que os mesmos sejam devidamente rastreados.

O tratamento da periodicidade de atualização pode ser visto como uma forma de normalização onde, os atributos são separados de acordo com a sua frequência de atualização em atributos que não mudam, atributos que raramente mudam, atributos que mudam algumas vezes e atributos que freqüentemente mudam. Os atributos que nunca mudam, não apresentam nenhum problema para o ADW. Entretanto, os que sofrem alterações ao longo do tempo devem ser analisados quanto a importância dessa alteração para a análise. Deve ser verificado, se essa alteração merece ou não ser rastreada durante a análise.

De acordo com o número de atributos alteráveis, freqüência de atualização e importância de acompanhamento, novas entidades irão surgir. Esse desmembramento evita o acesso a um grande volume de informações para realizar uma pequena atualização. Além disso, ele garante a inserção da chave tempo apenas nas entidades onde ela, realmente, é importante.

Quando, entretanto, a quantidade de atributos a serem rastreados for significativa, mesmo que apresentem periodicidades diferentes, é recomendável adotar a inserção da chave tempo na entidade ou adotar um dos processos de rastreamento.

Um exemplo de desmembramento ocorre na entidade *Aluno* para o atributo *COEF_REND_A_REG*. Este atributo representa o coeficiente de rendimento acumulado do aluno e varia a cada período. Essa é uma informação valiosa para análise de desempenho de alunos. Neste caso, será criada a entidade *Historico_Coef_Rend* que armazena as informações de histórico ao longo dos períodos. Deve ser registrado que o atributo *COEF_REND_A_REG* na entidade *Historico_Coef_Rend* pertence, no modelo operativo à entidade *Aluno*.

B.III.c) Inserir a Chave Tempo:

O propósito desta etapa é avaliar se uma entidade necessita ou não de uma chave tempo. A inserção da chave tempo não é uma tarefa trivial (SILVERSTON, INMON, GRAZIANO, 1997). O controle de tempo está relacionado à necessidade de rastrear as alterações em atributos nas entidades. Para prover esse rastreamento, uma análise da informação sendo armazenada é necessária, permitindo decidir sobre a melhor granularidade de tempo a empregar, como por exemplo, por dia da semana, por semana, por mês, por ano ou por ano fiscal.

A chave tempo é indispensável para as entidades onde as alterações de atributos são importantes e apresentam grande influência sobre o negócio. Nestes casos é importante permitir o rastreamento das mudanças. Existem outras maneiras de acompanhar as alterações, contudo o emprego da chave tempo é a forma mais simples e clara.

A inserção da chave tempo pode ser realizada de duas formas:

- Pontual: quando se adiciona um elemento de tempo à chave das entidades, para que seja possível analisar as modificações ao longo do tempo. Essa informação de tempo representa a data em que a informação foi carregada no banco. Um exemplo desse tipo é a inserção dos atributos *ANO* e *PERIODO* na entidade *Historico_Coef_Rend*, garantindo o armazenamento das informações de coeficiente de rendimento dos alunos ao longo dos períodos;
- Faixas Contínuas: essa técnica permite representar faixas contínuas de tempo, ao invés de pontos ou datas específicas. É realizada pela adição de dois campos do tipo data, um marcando o início e outro o fim de um determinado intervalo de tempo.

B.III.d) Estabelecer Padrão e Valor "Default" de Atributos no ADW:

Esta etapa é a responsável por avaliar os padrões existentes para os atributos, identificando o padrão ideal para o ADW. Além disso, serão estabelecidos valores "default" para atributos, quando houver necessidade.

a) Estabelecer Padrão para Atributos

É comum a existência de um mesmo atributo em DER de sistemas diferentes, apresentando diferentes padrões quanto a tipo e tamanho. Um padrão deve ser estabelecido para o ADW. O padrão escolhido pode não ser encontrado em nenhum dos DER, prevalecendo, neste caso, a decisão de apresentar a informação na forma mais acessível ao usuário final.

A especificação do padrão a ser empregado e da regra de conversão, quando necessária, deve ser registrada. Essa especificação será utilizada pelos desenvolvedores das aplicações de extração e carga dos dados, a ser realizada após a conclusão da modelagem.

O campo DATA-ENTREGA, por exemplo, encontra-se armazenado, em alguns

sistemas operativos, na forma DDMMAAAA com tipo String, porém, no ADW será armazenado como tipo DATE no formato DD/MM/AAAA. A conversão a ser empregada para a transformação do STRING em DATE, considerando um banco de dados ORACLE, será:

TO_DATE(DATA_ENTREGA, 'DDMMAAAA');

b) Estabelecer Valores "Default":

A definição de valores "default" é importante para os casos onde a integração resulta em novos campos, que não são preenchidos em todas as situações e em entidades que apresentam atributos com diferentes periodicidades de atualização. Outro aspecto importante na definição de valores "default" é o estabelecimento da diferença no conceito de valor nulo ("null") no ADW. Essa definição é importante para os SSD que serão elaborados posteriormente. O registro destes valores também é importante, principalmente para os bancos de dados que não permitem a inclusão de valores "default". Neste caso, os desenvolvedores de aplicação terão que se preocupar com o gerenciamento destes valores no momento da carga/atualização.

O atributo *FREQUENTA_ALOJAMENTO* na entidade *Aluno,* por exemplo, pode assumir como "default" o valor "NÃO", sendo alterado apenas quando necessário.

B.III.e) Estabelecer Regras de Conversão para Substituir Códigos e Abreviações

Esta etapa é responsável pela substituição de códigos e de abreviações empregados, nos bancos de dados operativos, por descrições textuais. Essa substituição pode ser realizada como uma adição ou substituição da informação existente.

Para realizar essa substituição, é exigido um controle de qualidade de modo a evitar abreviações e eliminar variações em textos que apresentem valores iguais. O controle permite garantir a qualidade dos dados descritivos. A substituição de código sem um controle rígido pode levar a consequências desastrosas na elaboração de relatórios.

Por exemplo, o atributo "rua" não pode apresentar os seguintes valores: "Nossa Senhora de Copacabana, N.S. Copacabana e N.S. Copa.". Caso o usuário final pretenda analisar um relatório por rua, uma "quebra" indevida é apresentada na sequência. Portanto, a forma diferenciada de escrever pode levar a uma interpretação errada dos

aplicativos de consulta.

Caso não exista a possibilidade de se obter os textos legíveis a partir dos sistemas operativos, é recomendável o registro de uma função de transformação. Estas funções permitem aumentar ou substituir a informação codificada ou abreviada, sempre que a carga for efetuada. Dessa forma, por exemplo, a seguinte função deve ser registrada para a conversão do atributo *NACIONALIDADE_REG* na entidade *Aluno*:

Caso *NACIONALIDADE_REG*:

1 : substituir por "**Brasileiro**";

2 : substituir por "**Naturalizado**";

3 : substituir por "**Estrangeiro**".

B.III.f) Criar Atributos Derivados

Esta etapa se preocupa em avaliar as necessidades de atributos que devem ser calculados e definir o algoritmo a ser executado no momento da sua carga.

No ambiente operativo, os projetistas normalmente não incluem os dados derivados como parte do processo de modelagem de dados, apenas os adicionam no projeto do banco de dados físico, por questões de desempenho e/ou facilidade de acesso (SILVERSTON, INMON, GRAZIANO, 1997). Para o ADW, entretanto, a adição dos dados derivados ao modelo é recomendada, porque reduz o processamento e garante a integridade das informações. Por serem calculados apenas uma vez, não há chances de aplicações distintas empregarem diferentes algoritmos para o cálculo dos dados, comprometendo a credibilidade do DW. Segundo Silverston (SILVERSTON, INMON, GRAZIANO, 1997), é possível criar este tipo de atributo aplicando a seguinte pergunta:

- *"A adição deste dado é importante?"*

A representação de valor total por item na entidade *VENDAS*, é um exemplo de valor derivado, sendo o total_item definido pela seguinte regra:

total_item = ((itens_NF.prc_unit * itens_NF.qtd) – itens_NF.desconto).

4.3.3 **FASE C - ELABORAR O MODELO DIMENSIONAL**

Esta fase recebe, como entrada, o pré-modelo definido na fase anterior. A partir dele, serão derivados um ou mais modelos dimensionais que comporão um DM. Como citado por Kimball e por Firestone (KIMBALL, 1995) (KIMBALL, 1997) (FIRESTONE, 1998-a), um modelo relacional pode ser transformado em um conjunto de modelos dimensionais. Portanto, é importante identificar fatos básicos e visões dimensionais que possam ser extraídas do pré-modelo, identificando a existência ou não de mais de um modelo dimensional.

A elaboração de um modelo dimensional exige que o projetista se preocupe com fatos, atributos, dimensões e hierarquias, que são os elementos básicos desse tipo de modelo. Além disso, detalhes quanto à aditividade dos valores na tabela de fato, à existência ou não de dimensões degeneradas, dentre outros, devem ficar registrados para facilitar o desenvolvimento das aplicações para os usuários finais.

A derivação é realizada, por modelo dimensional a ser gerado, selecionando-se, do pré-modelo, a entidade relacionada ao fato básico e definindo, através do processo de poda e enxerto, suas dimensões. O processo de poda e enxerto empregado se baseia nas técnicas de "Prunning" utilizadas em algoritmos de aprendizado e nas técnicas de "Poda e Enxerto" para a derivação de um modelo conceitual de DW, em desenvolvimento pela Universidade de Bologna (GOLFARELLI, 1998).

Esta fase encontra-se dividida nas seguintes subfases:

- Levantamento dos Fatos Básicos e Visões Dimensionais: Esta subfase tem o propósito de verificar se as necessidades do usuário final são atendidas e estabelecer os fatos básicos e as visões dimensionais a partir do pré-modelo; e
- Derivação dos Modelos Dimensionais: Cada fato básico identificado na subfase anterior, transforma-se em um modelo dimensional. Este fato se refere a uma entidade do pré-modelo, a partir do qual o modelo dimensional será derivado. Ao final desta subfase, o projetista possui um ou mais modelos dimensionais relacionados ao DM em questão; e

Ao final desta fase os modelos dimensionais gerados são integrados estabelecendo um modelo único para DM. O modelo dimensional do DM será empregado na fase seguinte para a integração de DW dimensionais (Visão Kimball).

C.I – Realizar o Levantamento dos Fatos Básicos e das Visões Dimensionais:

O propósito desta subfase é estabelecer os fatos básicos a partir do pré-modelo e realizar um levantamento das possíveis visões dimensionais.

O fato básico, segundo Kimball (KIMBALL *et al*, 1998) é uma descrição, através de uma expressão composta, das informações (fatos) que podem ser extraídas do pré-modelo. Ao se estabelecer os fatos básicos é possível identificar visões dimensionais. No modelo Universidade, por exemplo, o "registro de conceito e situação de aluno por disciplina em ano/período" representa um fato básico. A partir deste fato básico é possível realizar, dentre outras, as seguintes análises: acompanhamento dos alunos por disciplina e verificação dos alunos que apresentam um grande número de trancamentos de disciplina. A seguir são apresentados alguns exemplos de fatos básicos (KIMBALL *et al*, 1998):

- Transações de venda;
- Transações de reivindicações de seguro;
- Total diário de vendas em cada loja;
- Fotografia ("SNAPSHOT") mensal de contas;
- Linhas de item de um pedido; e
- Linhas de item de uma fatura de remessa.

No que se refere às visões dimensionais, nesta subfase, os projetistas normalmente apresentam um melhor conhecimento do problema, o que possibilita a definição de novas visões, além daquelas solicitadas pelo usuário final. As visões dimensionais definidas devem ser apresentadas para a análise e a aprovação do usuário final. A partir desta subfase:

- Não deverão existir requisitos vagos;
- Todas as visões dimensionais de interesse deverão ser extraídas do pré-modelo; e
- usuário final deverá conhecer as limitações para suas análises.

É importante salientar neste ponto, que uma entidade transformada em dimensão, durante a elaboração de um modelo dimensional, será reaproveitada pelos demais modelos, agilizando o processo de desenvolvimento.

C.II – Derivar os Modelos Dimensionais

Esta subfase é responsável por gerar um modelo dimensional para cada fato básico estabelecido na subfase anterior. Ao final desta fase, ter-se-á os modelos dimensionais que comporão o DM em questão.

C.II.a) Selecionar Entidade Chave referente ao Fato Básico (Fatos) :

Esta etapa identifica, no pré-modelo, a entidade que representará a tabela de fatos no modelo dimensional, de acordo com o fato básico. Essa entidade apresenta as informações de interesse para o processo de tomada de decisão. Os projetistas devem dar preferência às entidades que apresentem informações que sejam freqüentemente atualizadas e relacionadas aos fatos de interesse. Devem ser evitadas as entidades que representem propriedades estruturais do domínio e que apresentem informações quase estáticas.

A tabela 4.1 apresenta as entidades candidatas à tabela de fatos de acordo com os fatos básicos apresentados em C.I.

FATO BÁSICO	ENTIDADE
transações de venda	*NOTA FISCAL*
transações de reivindicações de seguro	*PEDIDO ABERTO*
total diário de vendas em cada loja	*NOTA FISCAL*
fotografia ("Snaphot") mensal de contas	*CONTA ou BALANCETE*
linhas de item de um pedido	*PEDIDO*
linhas de item de uma fatura de remessa	*FATURA*

Tabela 4.1 Relação de Fatos Básicos com Entidade Chave.

No modelo Universidade, por exemplo, para representar o fato básico registro de conceito e situação de aluno por disciplina em ano/período a entidade chave é *Registro_Nota.* Essa entidade apresenta uma atualização periódica e está diretamente relacionada ao problema de acompanhar as notas dos alunos por disciplina.

C.II.b) Definir o Esboço do Modelo Dimensional

Esta etapa é responsável por definir o esboço do modelo dimensional. Este esboço é gerado pela construção e tratamento de uma árvore de entidades/ relacionamento seguida da aplicação dos procedimentos de poda e enxerto. A árvore em questão é obtida a partir do pré-modelo, cuja raiz é definida pela entidade selecionada

na etapa anterior, e cujos nós, que compõem as subárvores, são representados pelas demais entidades/relacionamentos.

A operação de poda consiste em eliminar uma subárvore da árvore. Os nós que são retirados não são incluídos no modelo dimensional.

A operação de enxerto é feita quando, apesar de um nó não ser importante para análise, seus descendentes o são, devendo ser preservados. Outro caso de enxerto ocorre quando um nó apresenta uma grande relevância para a análise, devendo ser podado de uma subárvore e enxertado na raiz.

Os passos necessários para a construção e aplicação das operações estão relacionados a seguir.

C.II.b.i)Construir uma Árvore com as Entidades/Relacionamentos:

Esta subetapa é responsável por apresentar a construção de uma árvore de entidades/relacionamentos, sobre a qual será realizado o processo de poda e enxerto. O anexo 3 apresenta a terminologia normalmente empregada no trabalho com estrutura em árvores.

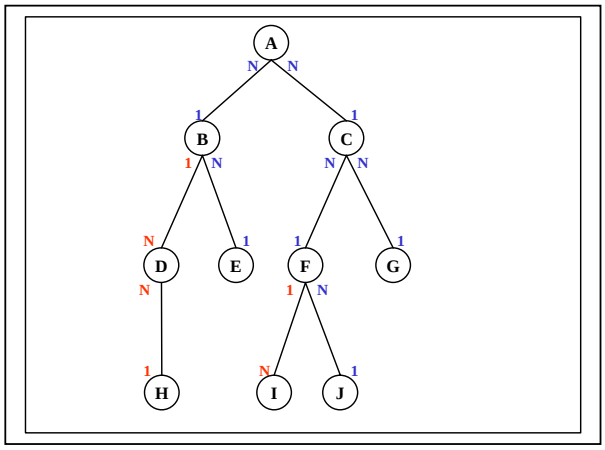

Figura 4.8 - Árvore Gerada do Pré-Modelo com Cardinalidades

A construção da árvore é um processo simples, onde o fato selecionado se torna a raiz da árvore e as entidades relacionadas ao fato compõem as subárvores. Cada uma das entidades passa a representar um dos nós das subárvores. O geração da árvore é realizada da seguinte forma:

Primeiramente é realizada a transformação das entidades/relacionamento relacionadas com a entidade representante do fato básico em uma estrutura de árvore. A raiz da árvore é a entidade selecionada como representante do fato básico. As demais entidades/relacionamentos comporão as subárvores. Nos casos em que ocorrer um ciclo, o nó em questão deve permanecer na subárvore com a qual apresente maior afinidade. A figura 4.8 apresenta um estrutura em árvore resultante da transformação de um conjunto

de entidades/relacionamentos. Conforme se observa, na árvore estão representadas as cardinalidades entre os nós.

Após a criação da árvore é realizado um tratamento com base na cardinalidade entre os nós. A cardinalidade na árvore representa a forma de relacionamento entre os nós. A tabela de fatos representa a raiz dessa árvore. A característica da tabela de fatos é ser a grande centralizadora de informações, portanto, o nó raiz representa uma composição dos demais nós. Dessa forma, é possível caminhar das folhas para a raiz como em um "Drill-up". Navegar por entre os nós de uma subárvore deve permitir aumentar a granularidade até chegar a granularidade máxima permitida (na tabela de fatos). A cardinalidade partindo da raiz para as folhas deve ser N:1, representando a característica de composição.

Dessa forma, a poda de uma subárvore e o enxerto de um nó dessa subárvore na árvore principal, deve ser feito apenas quando, até o nó a ser enxertado, a estrutura apresentava a cardinalidade N:1. Como restrição da operação de enxerto, deve ser garantido que o enxerto de um nó é realizado apenas em nós que pertençam ao caminho entre a raiz (inclusive) e o nó em questão.

Ao gerar a árvore, as seguintes situações podem ocorrer:

- Nó com cardinalidade 1:N: neste caso, toda a subárvore será podada, sem permitir aproveitamento de enxerto de nós. A cardinalidade 1:N quebra a idéia da composição em direção a tabela de fatos. Por exemplo, na árvore da figura 4.8 a subárvore constituída dos nós D e H será removida. O nó H apesar da cardinalidade 1:N com D não poderá ser enxertado na árvore original, pois o nó D quebra a cardinalidade. Ele não representa uma composição para o fato sendo analisado. Na figura 4.9, por exemplo, o nó *Disciplina_Vestibular* não compõe o fato básico representado pela raiz *Registro_Notas*, apresentando uma cardinalidade N:1 com o nó *Aluno.*
- Nó com a cardinalidade N:M: neste caso, uma análise deve ser realizada quanto a importância do nó para a análise. A decisão por manter o nó estabelece uma dimensão com cardinalidade M:N com a tabela de fatos. Essa dimensão será tratada no processo de refinamento de dimensões.

A árvore gerada ao final desta subetapa apresenta apenas as informações de interesse para a análise.

A figura 4.9 (I) representa a árvore original, definida no modelo Universidade, com a raiz na entidade *Registro_Nota.* Analisando a subárvore de aluno, observa-se que a partir dela os nós apresentam a cardinalidade 1:N. Estes nós serão removidos da árvore. A árvore sobre a qual será realizado o processo de poda e enxerto está representada na figura 4.9 (II).

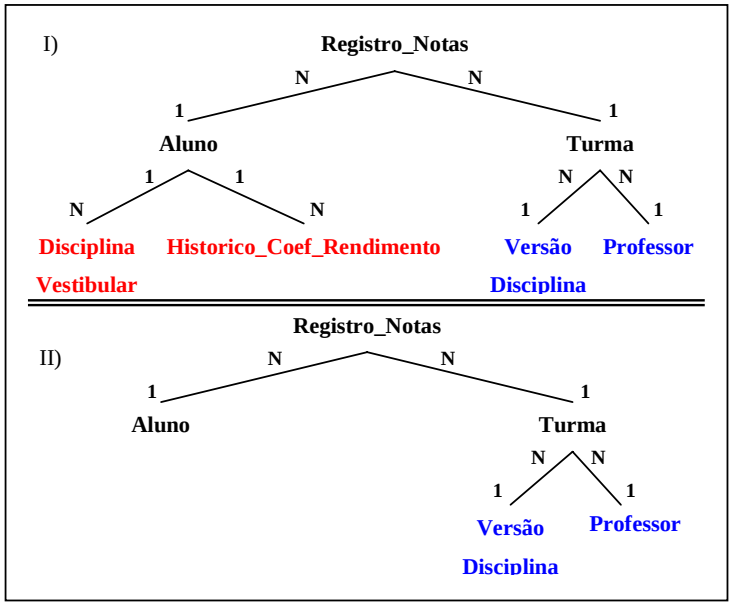

Figura 4.9 - Árvore com raiz em *Registro Notas.*

(I) Original. (II) Tratada.

O procedimento de poda e enxerto deve ser aplicado a cada subárvore a partir da raiz na árvore resultante.

C.II.b.ii)Realizar a Poda e o Enxerto da Árvore:

Esta subetapa é responsável por aplicar o processo de poda e enxerto sobre a árvore gerada na subetapa anterior. A existência, na árvore elaborada, de nós que não são de interesse para o DM é normal. Para excluir estes nós, é aplicada a operação de poda.

A poda também é aplicada quando um nó deve ser elevado de nível. Neste caso ele é podado e enxertado no nível selecionado. As operações de poda e enxerto têm o propósito de eliminar níveis desnecessários de detalhamento.

Funcionamento do Processo

Ao analisar uma subárvore, o projetista deve realizar as seguintes verificações:

- Quando os nós que compõem a subárvore representam informações como características, ou estão fortemente relacionados, são fortes candidatos a se tornarem hierarquias explícitas, conforme será apresentado na seção C.II.c.v (Estabelecer hierarquias). Em uma subárvore de uma entidade *Produto* onde os nós são representados pelas entidades *Tipo* e *Categoria* é possível observar que a subárvore representa características de produto.

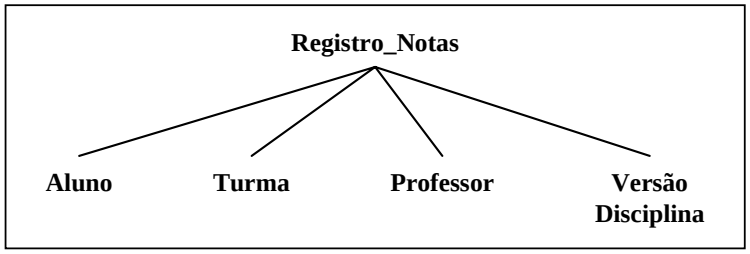

Figura 4.10 - Árvore com raiz em *Registro_Notas* Podada

- Para os casos onde os nós das subárvores representam informações de funcionalidades, de informações geográficas e de informações temporais, o grau de relevância dessas informações, para a análise, deve ser verificado. De acordo com a importância atribuída, o nó pode ser um forte candidato a se transformar em uma dimensão ou ser excluído através de PODA. Na figura 4.9 (II), o nó PROFESSOR na subárvore TURMA, representa uma funcionalidade e a análise por PROFESSOR é de relevância para o DM, portanto, PROFESSOR é um nó que será enxertado na RAIZ.

A árvore representada na figura 4.10 é obtida após o processo de poda e enxerto.

A operação de enxerto não é exclusiva da raiz, podendo ser realizada, também, em nós. Dessa forma uma subárvore extensa pode ser podada e os ramos de interesse enxertados em uma subárvore referente ao caminho da raiz ao nó.

C.II.b.iii) Criar a Dimensão Tempo:

Esta subetapa tem o propósito de estabelecer o nível de granularidade de tempo desejado para análise.

O ADW pressupõe o acompanhamento de dados através do tempo. A dimensão

tempo está relacionada às análises estabelecidas pelos usuários finais sobre a área de interesse. Além disso, essa dimensão, normalmente, emprega atributos para documentar feriados e diferentes marcos de períodos significativos como, por exemplo, dia da semana, semana do ano, estação do ano, flag-feriado, trimestre, quadrimestre e semestre.

Segundo Golfarelli (GOLFARELLI, 1998), os DER, quanto ao tratamento de tempo, podem ser classificados em "snapshots" e temporais.

DER "SNAPSHOT"- descreve o estado atual do domínio da aplicação. As versões antigas dos dados são continuamente substituídas por novas versões.

DER TEMPORAL - descreve a evolução do domínio da aplicação ao longo do tempo. Os dados antigos são explicitamente representados e armazenados.

Quando o modelo de dados é construído a partir de um DER temporal, o tempo é explicitamente representado como um atributo, sendo um forte candidato para definir a dimensão. No modelo CONTROLE_NOTAS, a entidade *Registro_Nota* que representa a tabela de fatos, apresenta granularidade de tempo por ano/período. Essa granularidade é a adotada para o modelo.

Nos modelos construídos a partir de DER SNAPSHOTS, o tempo pode não estar explicitamente representado. Entretanto, uma análise deve ser realizada para avaliar o tratamento de tempo considerado no DER, para a elaboração da dimensão tempo no modelo dimensional.

No exemplo de controle de vendas, o grão é representado pelo próprio pedido. O pedido possui o atributo "DATA-PEDIDO" no formato DD-MM-AAAA, sendo interessante uma análise junto ao usuário final para decidir a granularidade de tempo a ser armazenada e as hierarquias desejadas. Por exemplo, se o usuário final mantém a granularidade DIA, ele pode desejar ver informações agregadas por SEMANA, QUINZENA, MÊS, TRIMESTRE, SEMESTRE e ANO. Além disso, o usuário final deseja realizar análises dos pedidos em épocas de festas, como dia dos pais, dia das mães e Natal, sendo importante criar na dimensão tempo o atributo "PERÍODO", que identificará a data sendo analisada, informando se a mesma pertence ou não e a qual período de festas ela está relacionada.

C.II.c) - Refinar Dimensões:

Esta etapa tem o propósito de analisar as dimensões derivadas a partir do pré-modelo, tratando as grandes dimensões, estabelecendo as hierarquias e verificando a periodicidade de atualização dos atributos em cada dimensão.

A etapa anterior permitiu a elaboração de um esboço do modelo dimensional, com a definição de uma tabela de fatos e suas dimensões. Essas informações, a partir desse momento, passam a ser analisadas em maiores detalhes.

Os atributos das dimensões representam as fontes de todas as restrições interessantes e de todos os cabeçalhos de linha no conjunto de resposta. Kimball (KIMBALL, 1996) afirma que a qualidade do banco de dados é proporcional à dos atributos de dimensão. Portanto, quanto maior for o tempo destinado à descrição dos atributos, ao preenchimento de seus campos e à garantia de qualidade, melhores serão os resultados.

C.II.c.i) Tratar a Cardinalidade M:N no Relacionamento Dimensão X Fato

Esta subetapa tem o propósito de tratar as dimensões que apresentam relacionamentos com cardinalidade M:N com a tabela de fatos.

Normalmente, o relacionamento das tabelas de dimensão com a tabela de fatos apresenta a cardinalidade 1:N, entretanto, em alguns casos, pode acontecer uma cardinalidade M:N.

Um exemplo deste tipo de cardinalidade pode ocorrer, no modelo do DM VESTIBULAR, entre a dimensão OPCAO_CURSO e a tabela de fatos CONTROLE_NOTAS_VESTIBULAR. O propósito do relacionamento é permitir analisar as notas dos vestibulandos por opção de curso. Entretanto, um candidato pode selecionar até três opções, o que estabelece uma cardinalidade M:N entre a dimensão e a tabela de fatos. Para solucionar esse problema, duas soluções são propostas.

A primeira proposta se baseia na sugestão de Kimball para tratar esse problema apresentado na seção 3.4.1 (Tratamento de Dimensões e Fatos com Cardinalidade M:N). Nesta proposta a idéia é um pouco diferente. No exemplo relacionado à tabela de fatos CONTROLE_NOTAS_VESTIBULAR, ao gerar o esboço do modelo dimensional, o projetista conhece a dimensão com a cardinalidade M:N. Quando a dimensão é importante para a análise, deve ser criada uma chave na mesma, que será única na tabela

de fatos. Esta chave deve representar o conjunto de informações. A seguir será inserida uma nova dimensão que emprega a dimensão existente como ponte para acessar a tabela de fatos.

Dessa forma, na dimensão OPCAO_CURSO será inserida a chave *CH_OP_CURSO_VEST* que contém o mesmo valor para todas as opções do candidato. Essa chave substituirá a chave existente para OPCAO_CURSO na tabela de fatos. Com este tratamento, independente do número de opções selecionadas, só haverá uma representação para as mesmas. A seguir será inserida no modelo a dimensão CURSO. Esta dimensão se relaciona com a OPCAO_CURSO, permitindo estabelecer as consultas analisando as notas de disciplinas por curso.

Uma segunda forma de tratamento, válida apenas para os casos onde a cardinalidade é conhecida, é a criação de artefatos. No exemplo mencionado, as três opções de curso seriam transformadas em artefatos para o vestibulando, durante a elaboração do pré-modelo. Na fase da elaboração do modelo dimensional, os atributos referentes aos artefatos definiriam uma minidimensão da dimensão vestibulando, permitindo consultas pela descrição dos cursos.

C.II.c.ii) Analisar Dimensões com Itens Heterogêneos

Esta subetapa é responsável por analisar a existência de itens heterogêneos no modelo gerado. Duas situações podem ocorrer após a derivação: existir uma dimensão com itens heterogêneos ou existir uma dimensão com item específico.

Nos casos onde se observa a existência de uma dimensão com itens heterogêneos, uma análise deve ser feita para verificar a necessidade de criação de dimensões específicas. Essa análise consiste em verificar se os itens específicos estão sendo tratados em modelos de outros fatos básicos.

Para os casos onde existam dimensões específicas tratando de um item relacionado a um conjunto de itens heterogêneos, deve ser realizada uma análise para confirmar a necessidade da criação de uma dimensão global. Essa dimensão global permite atender a visões dimensionais mais genérica.

C.II.c.iii) Elaborar Minidimensões:

Nesta subetapa as dimensões serão analisadas para verificar a necessidade de criação de minidimensões.

Segundo Kimball (KIMBALL, 1996), as dimensões não devem ser desmembradas, mesmo que sejam extensas, pois seu desmembramento gera um desempenho limitado. Entretanto, rastrear modificações em dimensões extremamente grandes é um processo difícil. Com o propósito de resolver esse problema, Kimball (KIMBALL, 1996) recomenda o emprego de minidimensões.

As minidimensões também são recomendadas para melhorar o desempenho das consultas e facilitar a visualização de uma dimensão muito grande.

Neste contexto, as minidimensões devem ser compostas por pequenos conjuntos de atributos relacionados. Estes atributos devem conter um número limitado de valores, sendo recomendável o mínimo de 5 atributos (KIMBALL, 1996). A seleção dos atributos que comporão a minidimensão é realizada com base nos atributos mais utilizados e que permitam a seleção de um conjunto razoável de informações.

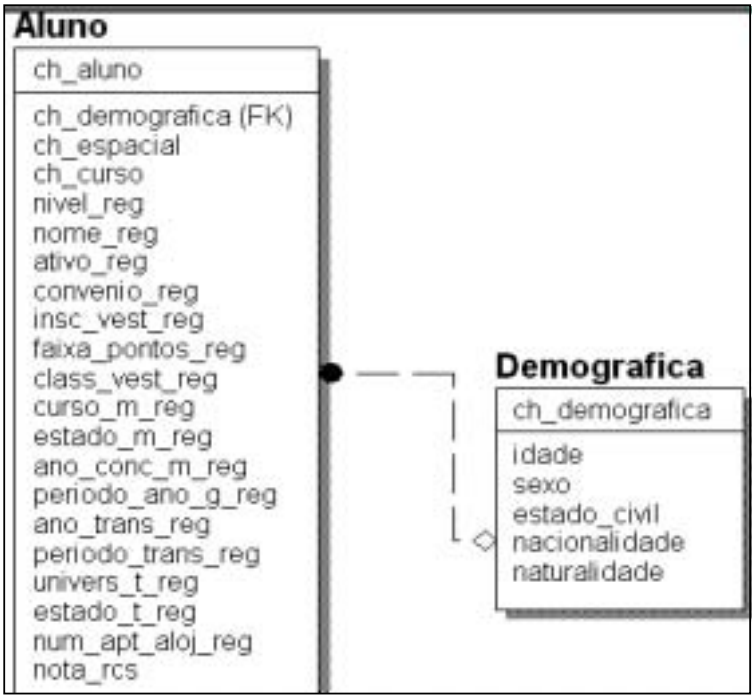

Figura 4.11 - Minidimensão DEMOGRÁFICA

O emprego de minidimensões é realizado na dimensão ALUNO, apresentada na figura 4.11. Neste exemplo, os atributos idade, sexo, estado_civil, nacionalidade e naturalidade, pertencentes a esta dimensão, passam a constituir a minidimensão DEMOGRÁFICA.

Outros meios de rastreamento existem para a análise de modificações lentas ao

longo das dimensões. Esse meios são apresentados em C.II.c.vi (Analisar a periodicidade de atualização e a necessidade de rastreamento de modificações lentas).

C.II.c.iv) Remover Atributos não Relacionados ao Modelo:

Esta subetapa avalia os atributos das dimensões, verificando a sua importância para o modelo. O resultado da avaliação pode gerar a exclusão de um atributo.

Cada DM, normalmente, é relacionado a um único departamento. Nestes casos é interessante que sejam excluídos da dimensão os atributos que classificam ou separam os dados por departamento.

Deve ficar registrado, entretanto, que estes atributos servirão de filtro para os programas de extração de informações para o DM. Portanto, a extração de atributos no DM deve ser cuidadosamente analisada, porque um atributo que aparentemente não apresenta interesse direto por parte do DM, pode ser fundamental para permitir ao usuário realizar "drill-down" até os dados do DW.

A remoção dos atributos que não representam interesse pode gerar uma dimensão "oca". Uma dimensão "oca" é aquela que apresenta a chave da dimensão e um ou poucos atributos, que representam as informações de interesse para análise. Neste caso, o projetista pode decidir por manter a dimensão ou criar uma dimensão descaracterizada, inserindo apenas os atributos de interesse na tabela de fatos.

No modelo Universidade, para a tabela de fatos CONTROLE_NOTAS a única informação da dimensão TURMA que apresenta algum interesse para a análise é o atributo *TURNO*. No modelo em questão, ele foi inserido na tabela de fatos como uma dimensão descaracterizada.

C.II.c.v) Estabelecer Hierarquias:

Esta subetapa se preocupa em estabelecer as hierarquias para cada dimensão do modelo.

Normalmente, as hierarquias explícitas são facilmente identificadas, como por exemplo, as hierarquias relacionadas com tempo e estrutura geográfica. Entretanto, as hierarquias implícitas requerem uma avaliação mais minuciosa e criteriosa. Para estabelecer essas hierarquias é necessária uma análise atributo a atributo, verificando se o mesmo permite uma hierarquia implícita.

Não é necessário criar dimensões separadas para comportar hierarquias diferentes e/ou comportar os atributos que tomem parte da hierarquia embutida em uma dimensão. Porém, é importante que os atributos empregados como hierarquia, assim como seu encadeamento, sejam registrados. Este registro será futuramente utilizado pelos projetistas de SSD e pelos usuários das ferramentas OLAP.

O registro das hierarquias facilita também o trabalho de criação de tabelas sumarizadas e dos níveis de visualização permitidos aos usuários, a ser estabelecido no nível físico.

Uma verificação, ao final da definição das hierarquias, deve ser realizada com o propósito de validar se a granularidade da tabela de fatos está realmente refletindo a menor granularidade das dimensões. Isto evita o armazenamento na tabela de fatos de registros que representem diferentes granularidades.

C.II.c.vi) Tratar a Necessidade de Rastreamento de Atributos de Dimensão

Esta subetapa se encarrega de ajustar a organização dos dados, de acordo com sua estabilidade/volatilidade, buscando uma estrutura confortável para o ADW.

O projetista deve analisar a volatilidade, verificando se uma informação representa apenas um indicador de transição ou uma informação concreta. A existência de dimensões dinâmicas e dimensões que apresentam estados que sofrem mudanças constantes, pode levar a uma reavaliação, a fim de verificar se a dimensão deve ou não compor o DM. É importante ressaltar a dificuldade em gerenciar dados dinâmicos.

O ajuste da organização dos dados é feito com base nos atributos que apresentam volatilidade identificada pelo projetista, com um destaque para os atributos descritivos. De acordo com a importância do atributo, os seguintes tratamentos podem ser oferecidos:

- Não manter o histórico, e simplesmente sobrescrever;
- Criar minidimensões;
- Adicionar um novo registro, com uma nova chave, e a nova descrição; e
- Criar um campo a mais para o atributo em questão na tabela dimensão, para manter o valor corrente.

Estes tratamentos encontram-se em detalhe na seção 3.4.2 (Técnicas de Rastreamento de Alterações).

C.II.d) – Refinar a Tabela de Fatos

Esta etapa é responsável por realizar uma análise dos atributos da tabela de fatos. Operações como exclusão de atributos que não sejam de interesse, adição de novos valores derivados e inclusão de subfatos podem ser realizadas. Essas operações são aplicadas à tabela de fatos, com o propósito de adequá-la ao modelo dimensional.

A tabela de fatos apresenta atributos de uma entidade selecionada no pré-modelo, portanto, alguns atributos originais podem não ser necessários e outros podem ser incluídos.

C.II.d.i) Tratar Atributos de Acordo com o Tipo de Fato

Esta subetapa é responsável por classificar o tipo do fato, analisando seus atributos e tratando-os quando for o caso.

Quando os fatos são do tipo transação ou do tipo linhas de itens, normalmente, a entidade selecionada apresenta a granularidade desejada, sendo necessário apenas excluir os atributos que não interessam à análise.

Quando os fatos são do tipo "snapshot", um trabalho maior é necessário. Neste caso, pode ocorrer da granularidade apresentada na tabela de fato, não ser a desejada pelo usuário. É preciso transformar a entidade "fato", para que ela assuma a granularidade devida. Esta transformação é realizada pela exclusão das informações transacionais e das identificações pontuais. Por exemplo, em "mensal de vendas de itens por loja", a entidade a ser transformada em tabela de fatos é representada por V*ENDAS*, combinação da entidade *NOTA_FISCAL* com a entidade *ITEM_NOTA_FISCAL*. A entidade *VENDAS* apresenta a informação de Nº de nota fiscal e data. O número da nota fiscal é uma identificação pontual que deve ser removida. A informação de data que se encontra no formato dia, mês e ano, deve ser transformada para mês e ano, e os valores agregados para comportarem o valor mês. Essa transformação deve ser registrada para a carga/atualização das informações do DM.

C.II.d.ii) Tratar Tabela de Fatos com Produtos Heterogêneos

Esta subetapa é responsável por identificar as tabelas de fatos que apresentam produtos heterogêneos. Uma tabela de fatos com produtos heterogêneos, normalmente, estabelece a criação de tabelas de fatos específicas para cada produto. Essa criação só é

necessária, quando na fase das visões dimensionais ela não tenha sido identificada.

Existindo ou sendo criada, a chave da tabela de fatos específica ou sub-fatos deve ser introduzida como chave na tabela de fatos geral, permitindo consultas específicas para os produtos.

A elaboração da tabela de fatos específica é realizada através da separação das informações específicas de dimensões, que se encontram na tabela de fatos principal, em tabelas de fatos menores (especializadas). Um exemplo deste tipo de fato foi apresentado em 3.3.1.2 (Fatos com Produtos Heterogêneos).

Durante a separação, novos atributos derivados podem ser criados para melhor atender às necessidades do modelo dimensional. Por exemplo, nos casos de aplicações bancárias como na figura 3.4, ao se extrair as informações associadas ao produto "depósito" transformando-o em subfato, pode ser inserida na tabela de fatos principal a informação de "total geral de depósitos", permitindo uma visão geral sobre esse serviço/produto.

C.II.d.iii) Classificar os Atributos

Esta subetapa tem o propósito de analisar cada atributo da tabela de fatos, classificando-os. Para os casos em que, na subetapa anterior, foram elaborados subfatos, esses também deverão ter seus atributos classificados.

Os atributos de uma tabela de fatos, normalmente, representam chaves de dimensões, dimensões descaracterizadas, chaves de subfatos ou atributos numéricos. A classificação dos atributos facilita o trabalho dos desenvolvedores de aplicativos e dos usuários finais, além de auxiliar a modelagem física quanto à melhor disposição para os atributos numéricos. As tabelas de fatos relacionadas a eventos são uma exceção no referente a atributos, porque normalmente, não os apresenta. A análise deste tipo de fato e a classificação dos demais atributos de acordo com seus tipo serão abordados a seguir.

Uma tabela de fatos pode não apresentar nenhum atributo de medição. Este tipo de tabela, também denominada tabela de fatos sem fatos, é normalmente empregada para modelar eventos. As tabelas de fatos de eventos apresentam como atributos apenas as chaves que representam as dimensões do evento. A única informação armazenada é a ocorrência do fato.

A seguir são apresentadas algumas formas de diferenciar os tipos de atributos

em tabelas de fatos não relacionados a eventos.

a) Chaves de Dimensões:

As chaves de dimensões representam os identificadores das dimensões e minidimensões relacionadas à tabela de fato.

b) Dimensões Descaracterizadas:

As dimensões descaracterizadas como apresentado em 3.3.2.3 representam chaves de dimensões sem dimensões. No momento da classificação dos atributos de uma tabela de fatos, pode ocorrer de surgirem atributos que, apesar de não representarem a chave de um dimensão descaracterizada, contêm informações de uma dimensão. Estes atributos permanecem na tabela de fatos em virtude de sua origem ter sido uma entidade do pré-modelo. Normalmente, quando este fato ocorre, duas opções podem ser consideradas:

- remover atributos: a remoção é permitida quando as informações não representam nenhum interesse para a análise e sua exclusão não provoca problemas, como por exemplo, contagem dupla;
- criar nova dimensão: quando os atributos apresentam interesse para a análise, eles formarão uma nova dimensão. A chave desta dimensão substituirá os atributos na tabela de fatos.

c) Chaves de Subfatos:

As chaves de subfatos representam os identificadores de tabelas de fatos específicas (subfatos). Essas chaves são empregadas nos modelos dimensionais que apresentam fatos relacionados a produtos/serviços heterogêneos.

d) Atributos Numéricos:

Os atributos numéricos, como o somatório do número de instâncias nos fatos, ou expressões envolvendo atributos numéricos relacionados às dimensões, devem ser classificados, de acordo com seus tipo, em aditivos, não aditivos e semi-aditivos.

Os tipos identificados para os atributos devem ser registrados para posterior emprego pelos desenvolvedores de SSD e usuários de ferramentas OLAP. Esse registro permite o desenvolvimento de aplicações que informam ao usuário final a possibilidade

ou não de realizar determinada consulta. A seguir, são apresentados procedimentos que facilitam a identificação dos tipos de valores.

VALORES ADITIVOS:

Identificar os valores que não apresentam dependência com relação a nenhuma dimensão. O emprego da informação com uma dimensão ou com um conjunto de dimensões não resulta em valores incorretos. Os valores "total alunos trancados" e "total alunos transferidos" na tabela de fatos CONTROLE_DISCIPLINA representam valores aditivos.

VALORES NÃO ADITIVOS:

Identificar os valores que apresentam dependência total das dimensões. O valor só é verdadeiro quando analisado por todas as dimensões estabelecidas, não sendo possível aplicar operadores a partir de uma única dimensão ou de um subconjunto delas. O percentual de alunos reprovados por disciplina ao longo dos últimos anos é um exemplo de valor não aditivo.

Quando um valor não aditivo é identificado, recomenda-se o armazenamento dos valores origem. A identificação dos valores não aditivos é importante, porque permite uma economia de espaço na tabela de fatos, quando transportados para o modelo físico.

VALORES SEMI-ADITIVOS:

Identificar os atributos que apresentam dependência parcial com relação a uma dimensão.

Para este tipo de valor é importante registrar qual a dimensão aditiva nas quais as operações podem ser aplicadas. É importante também analisar as informações a nível estático, de modo a verificar se realmente são não-aditivas ou se podem ser agregadas ao longo do tempo. Neste caso elas se enquadram como valores semi-aditivos.

Na figura 4.12 encontra-se a classificação dos atributos da tabela de fatos **VENDAS**.

Durante a identificação do tipo do atributo, é importante realizar o registro da expressão que descreve como ele é calculado a partir dos atributos do DER. Por exemplo:

quantidade vendida = SUM (Item_pedido.qtd);

*total = SUM (Item_pedido.qtd * Item_pedido.Prc_Unit).*

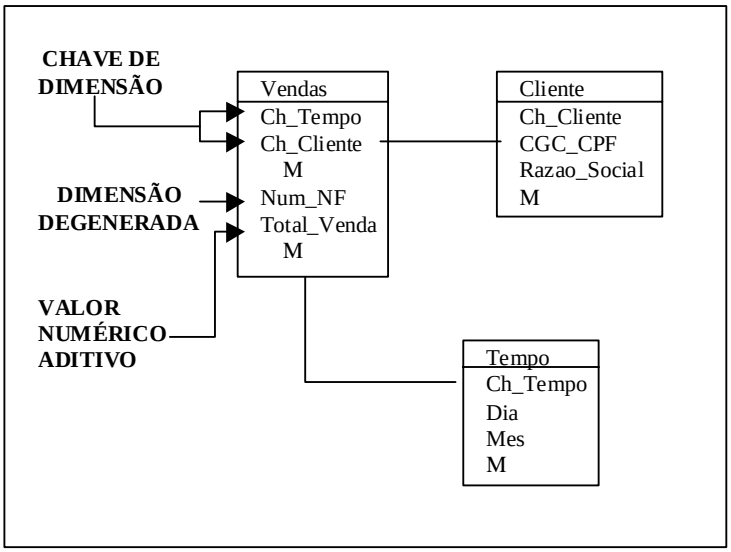

Figura 4.12- Classificação dos Atributos de Tabela de Fatos

Para facilitar a migração de aplicativos entre ferramentas OLAP e o desenvolvimento de novos SSD, é interessante manter também, um controle sobre limites dos valores, definindo sinalizadores. Um exemplo de sinalizador é apresentado a seguir:

Identificar os produtos com TOTAL_VENDAS < $5.000,00.

C.II.e - Nomear Atributos do Modelo com Termos de Negócios

Esta etapa é responsável por transformar os termos operativos utilizados nos atributos dos DER operativos em termos de negócio, facilitando o acesso do usuário. Uma nomenclatura mais próxima do usuário final é ideal para a elaboração de consultas e para a utilização de ferramentas OLAP.

A nova nomenclatura deve ser registrada. Este registro facilita o desenvolvimento de novos DM. Entretanto, deve ficar claro que um novo DM é livre para nomear seus atributos de acordo com seus usuários finais. Denominações

diferentes para um mesmo atributo são comuns em grandes empresas. Portanto, a flexibilidade de nomeação deve ser garantida e devidamente registrada. O registro poderá, por exemplo, permitir que uma modificação no futuro seja realizada rapidamente, localizando-se os DM que necessitarão de mudanças.

4.3.4 **FASE D – INTEGRAR O DM AO DW**

Esta fase é responsável pela integração do modelo do DM ao modelo de dados do DW. O procedimento a ser executado para a integração deve ser capaz de atualizar o modelo de dados do DW com o modelo do DM, mantendo sua consistência e integridade na passagem de um estado inicial para o estado final.

Esta fase está dividida em duas subfases:

- Integração dos dados: Responsável por realizar a integração propriamente dita do DM ao DW; e
- Tratamento de Evoluções no DW: Responsável por analisar o impacto das modificações e das mudanças no DW.

D.I) Integrar Dados:

Em virtude da abordagem dimensional(visão Kimball) e da abordagem relacional tratada (visão Inmon) serem as mais divulgadas e aplicadas na criação de DW, esta subfase apresenta uma versão de integração para cada uma.

D.I.a) Integrar o DM ao DW sendo o DW Dimensional - VISÃO KIMBALL

Esta etapa é responsável por apresentar o processo de integração a um DW dimensional.

É recomendável que o DW apresente o menor nível de granularidade possível, permitindo um projeto mais robusto (KIMBALL *et al*, 1998). Quanto menor a granularidade, maior a possibilidade de atendimento a novas consultas solicitadas pelos usuários finais. Além disso, facilita a introdução de novos elementos de dados (KIMBALL *et al*, 1998).

Segundo Kimball, a rápida integração de tabelas de dimensões e tabelas de fatos a um modelo existente representa uma das vantagens da modelagem dimensional. A seguir serão apresentados os passos necessários à integração. Estes passos demonstram

a necessidade de algumas análises criteriosas, que buscam a rápida integração proposta por Kimball (KIMBALL *et al*, 1998).

O processo de integração em um DW dimensional é realizado de modo incremental, a partir do dimensional do DM gerado na fase anterior. A integração deve começar pelas tabelas de dimensões e, a seguir, pela tabela de fatos. Os procedimentos de integração são apresentados a seguir.

D.I.a.i) Integrar Dimensões

Esta subetapa é responsável por realizar o processo de integração de cada dimensão do modelo dimensional do DM ao modelo do DW.

Duas situações podem ocorrer durante esta integração de dimensões: a tabela de dimensões pode existir ou não. Com o propósito de facilitar o entendimento, a tabela de dimensões do modelo dimensional a ser integrada será denominada DIMENSÃO MD e a do modelo de dados do DW será denominada DIMENSÃO DW.

A verificação da existência ou não da DIMENSÃO MD, no modelo de dados do DW, deve ser realizada com atenção, porque as dimensões podem apresentar nomes diferentes. A existência de dimensões semelhantes, com nomes diferentes, entre o DW e o DM é normal no ADW. Esta variação é aceitável porque cada DM, normalmente, atende a um departamento específico (INMON, 1997) (INMON, 1998), portanto, pode empregar termos próprios para sua área. Se o projetista não realizar uma verificação criteriosa, pode ocorrer uma duplicidade de informações. O problema da duplicidade pode se agravar, quando as dimensões apresentam periodicidades diferentes de atualização, gerando, por exemplo, a inconsistência dos dados e, consequentemente, a perda da confiabilidade do DW.

A inserção de uma nova dimensão ou a alteração das já existentes devem ser registradas. Este registro é de vital importância para os responsáveis pela carga/atualização do DW e dos DM. Uma ferramenta CASE ou um gerenciador de metadados facilitariam a localização de programas que devem ser alterados, além de permitir uma visualização do impacto decorrente das mudanças realizadas.

a) Criar Nova Dimensão

Quando uma DIMENSÃO MD não existe no modelo de dados do DW, ela deve

ser criada com todos os atributos que foram definidos no modelo dimensional. A criação de uma tabela de dimensões no modelo do DW representa o caso mais simples de integração. Entretanto, é necessário verificar se seus atributos não se encontram espalhados em outras dimensões ou minidimensões. Essa verificação deve ser realizada com base na origem do atributo. Os atributos que não existem podem ser criados na dimensão. Para os que existem é necessário verificar se as periodicidades atendem ao propósito ou se geram inconsistências de informações. Nos casos de conflito, a decisão final deve ser tomada através de reuniões entre os interessados pela informação.

A criação da dimensão e seus atributos deve ser registrada, permitindo o mapeamento: ambiente operativo ◄► DW◄► DM.

b) Atualizar uma Dimensão Existente

Para os casos onde exista uma DIMENSÃO DW equivalente à DIMENSÃO MD, os atributos das duas dimensões deverão ser verificados para garantir não apenas a existência, mas a semântica e estrutura. É importante, no momento da verificação de um atributo, certificar que o mesmo não esteja em minidimensões relacionadas a dimensão principal.

O procedimento de verificação é aplicado a cada atributo da DIMENSÃO MD, para a certificação de sua existência ou não na DIMENSÃO DW. Da mesma forma que no levantamento de dimensões, os atributos podem apresentar nomes diferentes, porém com o mesmo significado ou nomes idênticos e significados diferentes. Nos casos onde o atributo não exista, sua inserção deve ser realizada automaticamente.

Quando o atributo existe na DIMENSÃO DW, ele pode apresentar ou não as mesmas características da DIMENSÃO MD. Essas diferenças podem acontecer com relação ao formato, ao valor "default", à freqüência de atualização, às regras, ao domínio, e à origem, sendo que:

- Formato: os atributos devem apresentar tipo e tamanho compatíveis. Por exemplo, se a DIMENSÃO DW apresenta o atributo data com o formato "DD/MM/AAAA", ele poderá ser transformado para um formato, na DIMENSÃO MD, "MM/AAAA";
- Valor "default": os valores estabelecidos para "default", nas dimensões, devem ser compatíveis. No caso de data, se o "default" na DIMENSÃO DW é '01/01/1990' na DIMENSÃO MD deverá ser '01/1990', garantindo a compatibilidade;

- Freqüência de atualização e regras: a frequência de atualização e regras dos atributos devem ser compatíveis;
- Domínio: O domínio do modelo dimensional dever ser menor ou igual ao domínio do DW; e
- Origem: a origem dos dados deve ser a mesma.

Os casos que apresentam discrepância deverão ser discutidos entre os interessados: projetistas, ABD e usuários finais. Nessas discussões deverão ser analisadas as necessidades do usuário final e o impacto das alterações sobre o ambiente. A alteração de um atributo no DW será abordado mais adiante, na seção D.II (Mudanças estruturais no DW).

Após as inserções na tabela de dimensão, deve ser realizada uma análise da dimensão final, com o propósito de avaliar possíveis necessidades de desmembramento por:

- Freqüência de atualização: para os casos onde existam um determinado conjunto de atributos com frequência de atualização diferente;
- Criação de minidimensões: a inclusão de novos atributos pode gerar uma dimensão muito extensa ou pode, juntamente com outros atributos existentes, constituir um grupo representativo para a criação de uma minidimensão; e
- Rastreamento de mudanças: os novos atributos podem exigir um rastreamento até então desnecessário na dimensão. As técnicas de rastreamento podem ser adotadas conforme a abordagem em 3.4.2 (Técnicas de rastreamento).

O esquema da figura 4.13 apresenta a integração entre uma dimensão do DM e uma dimensão do DW. No esquema está representado, também, a definição de uma minidimensão após a atualização da dimensão DW.

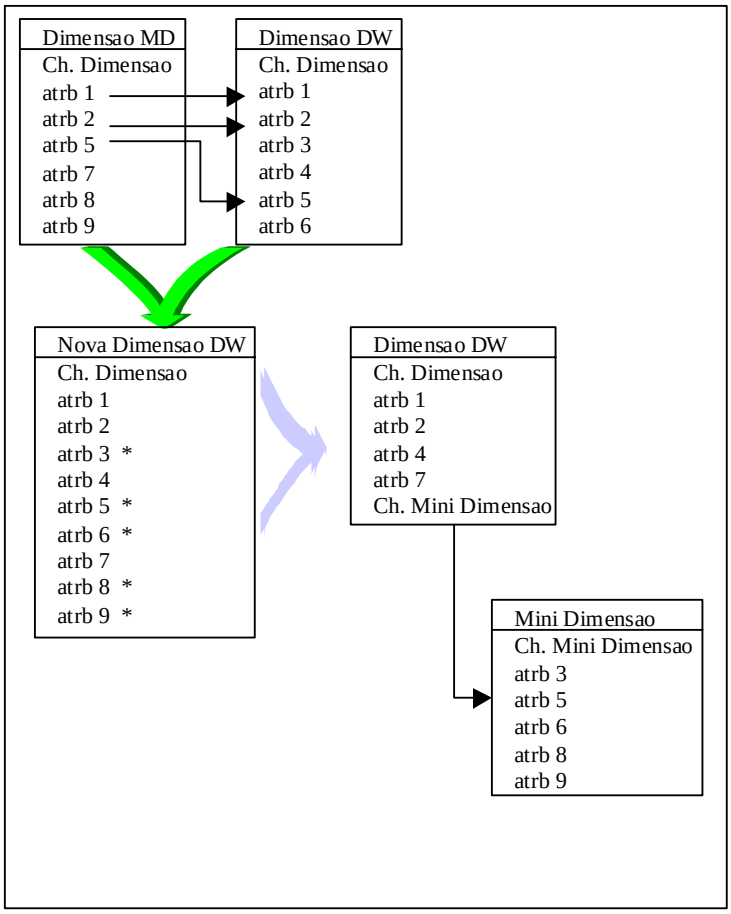

Figura 4.13 - Integração de Dimensões DM x DW

D.I.a.ii) Integrar Fatos

Após a integração das dimensões, inicia-se a integração da tabela de fatos. Para auxiliar o entendimento desse procedimento, a tabela de fatos do modelo dimensional será denominada FATO MD e a tabela de fatos do DW será denominada FATO DW. O procedimento para a integração do FATO MD ao FATO DW não é trivial. A localização de um FATO MD no modelo DW não é simples. Uma avaliação deve ser realizada para evitar a criação deliberada de tabelas de fatos sem necessidade, provocando um aumento no volume de dados e um novo processo para a carga e atualização. Ao se realizar um levantamento da tabela de fatos no modelo de dados do DW com relação ao FATO MD, quatro situações podem ocorrer:

- Nenhum FATO DW se aproxima do FATO MD;
- Existe FATO DW com a mesma granularidade;
- Existe FATO DW com granularidade superior; e
- Existe FATO DW com granularidade inferior.

Cada uma das situações requer um tratamento específico para a integração. Estes tratamentos são apresentados a seguir.

a) Nenhum FATO DW se aproxima do FATO MD

Quando não existe nenhuma tabela de fatos no modelo de dados do DW que se aproxime do FATO MD, ou quando se constata, após uma análise, que os fatos existentes não atendem as necessidades do FATO MD, a tabela de fatos deverá ser criada. Após a sua criação é necessário estabelecer os relacionamentos com as dimensões.

b) Existe FATO DW com a mesma granularidade do FATO MD

A existência de uma tabela de fatos no DW com a mesma granularidade que o FATO MD, representa o caso mais simples da integração.

Verificada a existência da tabela de fatos, é realizada a inserção de atributos do FATO MD que não existam no FATO DW. O tratamento de atributos é descrito em maiores detalhes após a apresentação das situações observadas.

c) Existe FATO DW com granularidade superior a do FATO MD

Nos casos onde se verificar que o FATO DW apresenta uma granularidade superior à do FATO MD, normalmente, é possível que o FATO MD possa ser obtido a partir do FATO DW.

Todos os atributos do FATO MD devem ser extraídos a partir do FATO DW. Se existirem atributos no FATO MD que não possam ser extraídos do FATO DW, esses atributos deverão ser incluídos. A inclusão de atributos que representem valores numéricos deve considerar a possibilidade de desmembramento. Este desmembramento ocorre em virtude das diferenças entre as granularidades, com o propósito de compatibilizar as informações. Por exemplo, a inserção de médias mensais de descontos ao ser inserida na tabela de fatos que contenha movimentações diárias será

desmembrada em valor de desconto por item. A regra para gerar a média, entretanto, deverá ser registrada no DW para uso dos usuários finais, aplicações e ferramentas OLAP.

O registro de regras no DW permite uma uniformização das regras empregadas no ADW, evitando que cada DM estabeleça uma lei de formação.

Pode acontecer, entretanto, de um atributo não ser obtido pelas vias normais. Nesses casos a solução será a criação de uma nova tabela de fatos (agregado).

d) Existe FATO DW com granularidade inferior a do FATO MD

Nos casos onde exista um FATO DW com uma granularidade menor que o FATO DM, duas alternativas são possíveis:

- substituir o FATO DW pelo FATO MD: se o FATO DW pode ser gerado a partir do FATO MD (vide c), este fato poderá substituir o FATO DW no modelo de dados do DW. Essa substituição deverá ser realizada com base em uma avaliação do impacto sobre as consultas e DM envolvidos, para que a sua validade seja certificada; e
- manter o FATO DW como um agregado e inserir o FATO DM: o registro da duplicidade de informações que possam vir a existir entre as tabelas de fatos deve ser realizado.

TRATAMENTO DE ATRIBUTOS

O tratamento aplicado aos atributos de tabelas de fatos é similar, em alguns pontos, ao tratamento de atributos de dimensões. Cada atributo de FATO DM requer uma verificação quanto à sua existência no FATO DW. Quando o atributo não existir, ele será automaticamente criado. Caso o atributo exista é necessário uma análise criteriosa quanto a:

- Formato: verificar se tipo e formato são os mesmos;
- Origem: a origem do atributo deve ser única;
- Tipo: a classificação, no caso de valores numéricos, quanto a aditivo, semi-aditivo e não-aditivo deve ser a mesma;
- Expressão: para os casos onde os valores são numéricos e apresentam uma expressão para cálculo, ela deve ser única.

Quando ocorrer divergência entre os itens mencionados, é necessária uma

verificação através de reuniões entre os projetistas, os ABD e os usuários finais. A existência de discrepância pode estabelecer a criação de um novo atributo.

D.I.b) Integrar o DM ao DW sendo o DW Relacional - VISÃO INMON

Esta etapa é responsável pela integração do pré-modelo ao modelo de dados de um DW relacional, conforme proposto pela visão Inmon.

A integração é realizada para cada entidade existente no pré-modelo verificando a sua existência ou não no modelo do DW. O levantamento das entidades no DW não é uma tarefa trivial. Como na abordagem dimensional, podem existir entidades com nomes diferentes porém apresentando as mesmas informações ou acessando a mesma entidade no ambiente operativo. Da mesma forma, podem existir entidades com mesmo nome porém mapeando atributos de entidades do ambiente operativo completamente diferentes. É necessária uma verificação com relação à origem dos atributos existentes nas entidades para se certificar de sua igualdade.

Se uma entidade não existe, ela deverá ser criada com seus atributos e relacionamentos. Para os casos em que a entidade exista, uma verificação a nível de atributos é realizada.

Para cada atributo da entidade do pré-modelo, é feito um batimento com o atributo da entidade do DW. Os mesmos cuidados referentes aos atributos de dimensões devem ser aplicados aqui. O levantamento de atributos deve ser realizado para garantir que o atributo não exista em outra entidade. Quando o atributo não existe, ele é automaticamente criado. Quando o atributo existe, uma avaliação criteriosa deve ser realizada quanto aos quesitos: formato, valor "default", freqüência de atualização, regras, domínio e origem. Esses quesitos devem ser iguais.

Qualquer discrepância nestes quesitos deve ser discutida entre projetistas, ABD e usuários finais. Neste aspecto, a discrepância pode indicar um novo atributo. Possíveis alterações nos atributos já existentes deverão ser questionadas quanto às necessidades do usuário final e ao impacto sobre o ambiente. A alteração de um atributo no DW será abordada na próxima seção.

D.II – Mudanças Estruturais no DW

Esta subfase é responsável por apresentar um tratamento para as mudanças estruturais de um DW.

De acordo com Kimball (KIMBALL, 1996), as mudanças estruturais estão associadas ao processo de desenvolvimento incremental do DW. Qualquer mudança estrutural em um DW em uso requer uma apurada análise de impacto.

As mudanças estruturais são representadas pelas correções necessárias por erros de modelagem, por novas solicitações e por alterações com o propósito de melhorar o desempenho. Estas mudanças são aplicadas às dimensões (visão Kimball) e às entidades (visão Inmon) e envolvem a inclusão/exclusão de atributos e alterações de atributos quanto ao: formato, regras, origem e domínio. Estas mudanças e a diferença entre nomenclaturas dos atributos do modelo dimensional com o modelo do DW, devem ser registradas através de metadados, permitindo um mapeamento entre os dados do ambiente operativo e aqueles do DW e os do DM.

Nos casos onde seja empregado um DW dimensional, a inclusão e exclusão de atributos apresentam um outro tipo de mudança estrutural, a criação e eliminação de minidimensões. Quando novos atributos são inseridos em uma dimensão, conforme já abordado, podem gerar, na dimensão original, a necessidade de uma minidimensão. Essa minidimensão pode ser gerada com o propósito de:

- Tornar a dimensão mais clara;
- Auxiliar o processo de consultas; e
- Permitir o rastreamento de modificações lentas.

No processo de exclusão, de forma similar, uma minidimensão pode perder a razão de existir, com os atributos restantes, retornando à dimensão original. Este tipo de mudança, pode chegar ao nível de tabelas de fatos, desde que a chave da minidimensão represente uma das chaves dessa tabela.

a) Inclusão de Atributos

Para realizar a inclusão de um atributo, o projetista deve ficar atento com os possíveis conflitos quanto, por exemplo, a:

- Nomes diferentes e mesmo significado;
- Nomes iguais e significados diferentes;

- Nomes e significados iguais, porém estruturas diferentes;
- Nomes, significados e estruturas iguais, porém periodicidades diferentes.

Dois atributos são considerados idênticos, quando possuem semântica e estrutura iguais. Nos casos em que ocorram discrepância quanto a estrutura e freqüência de atualização, o projetista deve considerar a existência de outros DM que trabalhem com a existente. Alguns ajustes podem ser necessários ao modelo dimensional, para que ele apresente a mesma freqüência do DW. Uma outra solução é assumir a menor freqüência, desde que o impacto dessas atualizações não provoque perdas no desempenho. A melhor forma de tratar essa situação é através de reuniões entre projetistas, ABD e usuários finais.

A inclusão de um novo atributo deve considerar, também, o problema com o passado armazenado no DW e nos DM. Um novo atributo, criado por uma mudança de regra de negócio, pode passar a ser visualizado do momento da sua criação em diante. Os agregados e valores derivados existentes podem não permitir alterações. Essas considerações devem ser apresentadas aos usuários finais, para evitar a perda da confiabilidade nas informações do DW.

b) Exclusão de Atributos

A exclusão de um atributo, normalmente, ocorre quando ele não é mais útil ao ambiente. Para garantir que não exista mais interesse por parte dos DM e das ferramentas e aplicações de consultas, é necessário realizar uma análise sobre o atributo. Essa análise, além de garantir que o atributo pode ser excluído, permite um levantamento do impacto que será provocado com a sua exclusão. Quanto ao seu emprego direto ou indireto de agregados e regras.

c) Alteração da Estrutura do Atributo

Em virtude de um erro de modelagem ou alterações de negócio, um atributo pode sofrer alterações em sua estrutura. Essas alterações representam mudanças, dentre outras, quanto a: formato, regras, origem, domínio e tipo. Da mesma forma, como no processo de exclusão, uma análise se faz necessária para verificar o impacto dessas alterações sobre o ambiente.

4.4 Considerações Gerais

Para cada modelo dimensional de um DM, as entidades que foram tratadas são reaproveitadas no processo da elaboração do esboço do modelo. É a partir do refinamento deste esboço que uma dimensão pode, por exemplo, se transformar em uma tabela de fatos.

O conjunto de subetapas proposto tem o propósito de garantir que o modelo dimensional gerado atenda às características multidimensionais esperadas. As subetapas relacionadas ao refinamento das dimensões e refinamento dos fatos representam tratamentos a serem aplicados para atingir o propósito da etapa. A ordem como são executadas fica a critério do projetista. Esse fato é decorrência da modelagem neste ambiente ser muito mais um processo empírico do que científico.

CAPÍTULO 5

ESTUDO DE CASO
MODELO UNIVERSIDADE

A Universidade Federal do Rio de Janeiro deseja desenvolver um ADW, a partir das informações existentes em seus sistemas, para analisar de modo geral, os cursos mais procurados e o perfil dos alunos da universidade ao longo dos anos.

O desenvolvimento desse ambiente será realizado empregando as diretrizes propostas apresentadas no capítulo 4. Para a aplicação das diretrizes é necessário determinar os DM que serão construídos, bem como a prioridade de desenvolvimento.

Por decisão da Universidade, dois departamentos serão beneficiados: o Departamento de Graduação e o Departamento de Vestibular, cada qual possuindo um sistema operativo que atende às suas necessidades.

O primeiro DM atenderá ao setor da graduação. O propósito básico deste DM é permitir:

- Avaliações sobre o desempenho dos alunos da graduação ao longo dos anos; e
- Análise das disciplinas, verificando, por exemplo, aquelas que apresentam maiores índices de trancamento e cancelamento.

O segundo DM atenderá ao setor de vestibular. O seu propósito é fornecer uma avaliação dos candidatos ao vestibular, permitindo, por exemplo:

- Análises dos cursos mais procurados,
- Perfil dos candidatos por curso e
- Acompanhamento das médias do vestibular ao longo dos anos.

O estudo de caso emprega o conjunto de diretrizes apresentado no capítulo 4 para realizar a modelagem. O emprego das diretrizes segue o modelo apresentado na figura 4.2.

A notação IDEF1X é empregada para a confecção dos diagramas. Ao longo do estudo de caso serão apresentadas aquelas diretrizes que foram utilizadas no caso em questão. No anexo 5 encontra-se o dicionário de dados dos DER empregados neste estudo.

5.1 DM Graduação

FASE A – ESTUDAR OS MODELOS EXISTENTES

O modelo a ser utilizado nesse DM é o DER referente ao Sistema de Controle de Graduação (SCG). A versão simplificada desse DER é apresentada na figura 5.1. Nessa figura, as entidades e relacionamentos que compõem o escopo da análise estão delimitados. Em azul estão as principais entidades para análise e em vermelho as entidades e relacionamentos considerados de interesse. As entidades e relacionamentos selecionados compõem o diagrama que serve de entrada para a próxima fase.

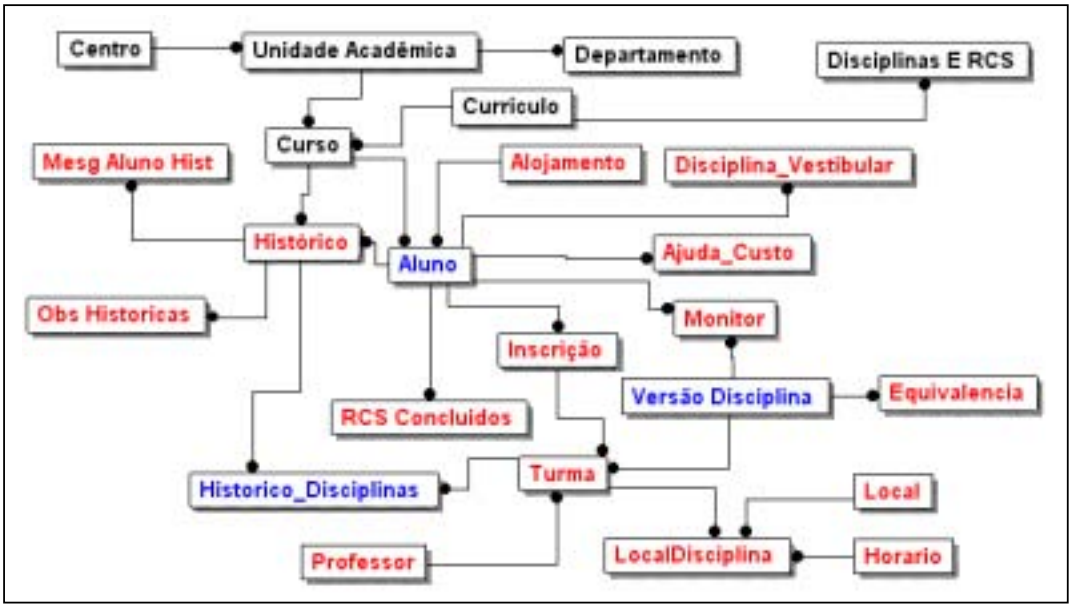

Figura 5.1- DER do Sistema de Controle de Graduação (SCG)

FASE B – ELABORAR O PRÉ-MODELO

Entrada: DER resultante da fase anterior.

Saída : Pré-modelo.

O DER do SCG contendo as entidades/relacionamentos selecionados está representado na figura 5.2.

Figura 5.2 - DER resultante da FASE A

B.I) Limpar e Transformar Modelos:

Entrada: DER resultante da fase anterior.

Saída: DER limpo e transformado.

B.I.a) Excluir Informações Desnecessárias:

Algumas entidades no modelo mantém informações operativas ou que não interessam a análise. Essas entidades são removidas com o propósito de reduzir o escopo para a análise de atributos. As entidades/relacionamentos a serem excluídas do modelo são apresentadas a seguir:

- *Equivalencia*: contém informações operativa de disciplinas;
- *Historico_Obs* e *Historico_Mensagem*: contém informações operativa de histórico de alunos;
- *Local_Disciplina, Horario* e *Sala:* contém informações operativa de turma; e
- *Inscricao_Periodo* : Esta entidade será excluída por conter informações operativas referentes ao período escolar. Entretanto, na extração de informações para o DW, ela será empregada para carregar no ADW as disciplinas que foram trancadas ou canceladas pelos alunos. Estas disciplinas não são armazenadas no histórico pelo SCG.

Para as entidades que permanecem no modelo é necessário uma análise de seus atributos. A relação dos atributos excluídos, por entidade, por representarem informações desnecessárias a análise encontram-se no Anexo 4, item A4-1- Relação de Atributos excluídos no DER do SCG.

B.I.b) Desnormalizar Relações:

No DER é possível observar duas possibilidades de desnormalização: *Historico* com *Historico_Disciplina* e *Turma* com *Versao_Disciplina.*

Analisando a validade da desnormalização para *Historico* e *Historico_Disciplina* é possível observar que compartilham a chave primária, apresentam os dados, normalmente juntas e possuem um padrão de inserção semelhante. Portanto, podem ser desnormalizadas. A entidade gerada pela desnormalização é nomeada como *Registro_Notas.* Esta desnormalização é representada no capítulo 4 pela figura 4.9.

Analisando a validade da desnormalização para *Turma* com *Versao_Disciplina* é possível observar que a entidade *Turma* é uma representação da entidade *Versao_Disciplina* para um período e que essas entidades não apresentam uma padrão de inserção semelhante. Portanto, a desnormalização não é realizada.

B.I.c) Definir Categorias:

A entidade *Aluno* no DER de entrada apresenta as seguintes informações a serem categorizadas: Data de nascimento, pontuação do vestibular e data da graduação. Através da definição de categorias a esparsidade que poderia ser provocada no momento das consultas é controlada.

Os atributos *IDADE, FAIXA_PONTOS_VEST e PERIODO_ANO_GRAD* serão criados na entidade aluno com as informações categorizadas, substituindo os atributos *DIA_NASC, MES_NASC, ANO_NASC, PONTOS_VEST* e *DIA_GRAD, MÊS_GRAD* e *ANO_GRAD*. As regras para estabelecer as categorias estão definidas no anexo 4, item A4-2- Regras de categorias apresentadas no DER do SCG.

B.I.d) Criar Artefatos:

No DER de entrada foram selecionadas as entidades *Alojamento, Ajuda_Custo, Monitor* e *RCS_Concluido* para se tornarem artefatos na entidade *Aluno*. As informações das entidades não representam interesse para a análise. Entretanto, saber se um aluno utiliza alojamento, recebe ajuda de custo ou é monitor é de interesse. Assim como a nota do RCS para os alunos que já o concluíram. Essas informações representam interesse no momento em que são extraídas do sistema operativo. Portanto, podem ser criadas como artefatos. Os artefatos *FREQUENTA_ALOMENTO, RECEBE_AJUDA, MONITOR* e *NOTA_RCS* serão criados na entidade *Aluno.* As regras para estabelecer os artefatos estão definidas no anexo 4, item A4-3- Regras de definição de artefatos apresentados no DER do SCG.

A figura 5.3 apresenta o diagrama resultante da subfase de limpeza e transformação. Este diagrama não necessita de integração, servindo de entrada para a subfase de refinamento de DER

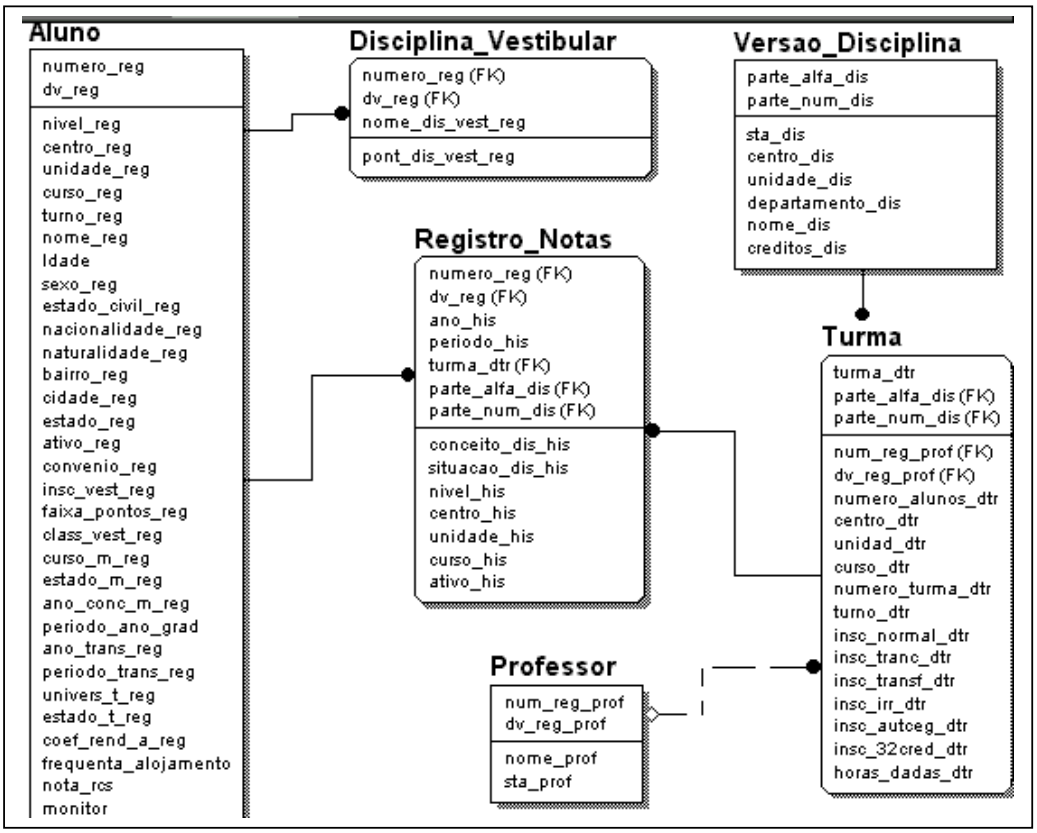

Figura 5.3 - DER resultante da SubFase de Limpeza e Transformação

B.III) Refinar o DER

Entrada: DER limpo das informações desnecessárias e tratado através da definição de categorias e artefatos.

Saída: Pré-Modelo.

B.III.a) Criar Novas Chaves:

As chaves das entidades do DER do SCG não são reutilizadas. Entretanto, as chaves das entidades básicas podem ser simplificadas para facilitar o acesso. Dessa forma, são criados os atributos *CH_ALUNO*, *CH_PROFESSOR* e *CH_DISCIPLINA*, como chaves para as entidades *Aluno*, *Professor* e *Versao_Disciplina*, respectivamente. As regras para a definição das novas chaves estão no anexo 4, item A4-4- Regras de definição de novas chaves de entidades no DER do SCG.

B.III.b) Analisar a Periodicidade de Atualização das Informações das Entidades.

De acordo com uma análise sobre a periodicidade de atualização dos atributos das entidades do DER é possível concluir que o atributo *COEF_REND_A_REG* pertencente a entidade *Aluno* requer mapeamento. Este atributo é periodicamente alterado e sua alteração exige um mapeamento. Para as demais entidades, apesar de possuírem informações que sofram alterações, não requerem mapeamento. As tabelas com as análise dos atributos que sofrem alteração estão no anexo 4. item A4-5- Análise da periodicidade de atualização das entidades no DER do SGC.

No DER será criada a entidade *Historico_Coef_Rendimento.* Esta nova entidade se relaciona com a entidade *Aluno.* O atributo referente a coeficiente de rendimento é excluído da entidade *Aluno.* Deve ser registrado que esse atributo pertence à nova entidade.

B.III.c) Inserir a Chave Tempo:

As informações de coeficiente de rendimento variam a cada período, portanto é necessário inserir na entidade *Historico_Coef_Rendimento*, as informações de ano e período.

B.III.d e B.III.e) Estabelecer Padrões, Valores "Default" e Regras de Conversão para Substituir Código e Abreviaturas dos Atributos do Modelo.

Neste momento as entidades são analisadas para que os atributos de interesse assumam um padrão e um valor "default" no ADW. Além disso é necessário estabelecer as regras de conversão necessárias.

Nesse estudo de caso as entidades *Aluno, Versao_Disciplina, Turma e Professor* necessitam de tratamento para alguns de seus atributos. A relação destes atributos com seus respectivos tratamentos estão no anexo 4, item A4-6- Definição de padrão, valores "default" e regras de conversão para os atributos no DER do SCG.

A figura 5.4 apresenta o pré-modelo gerado na fase B.

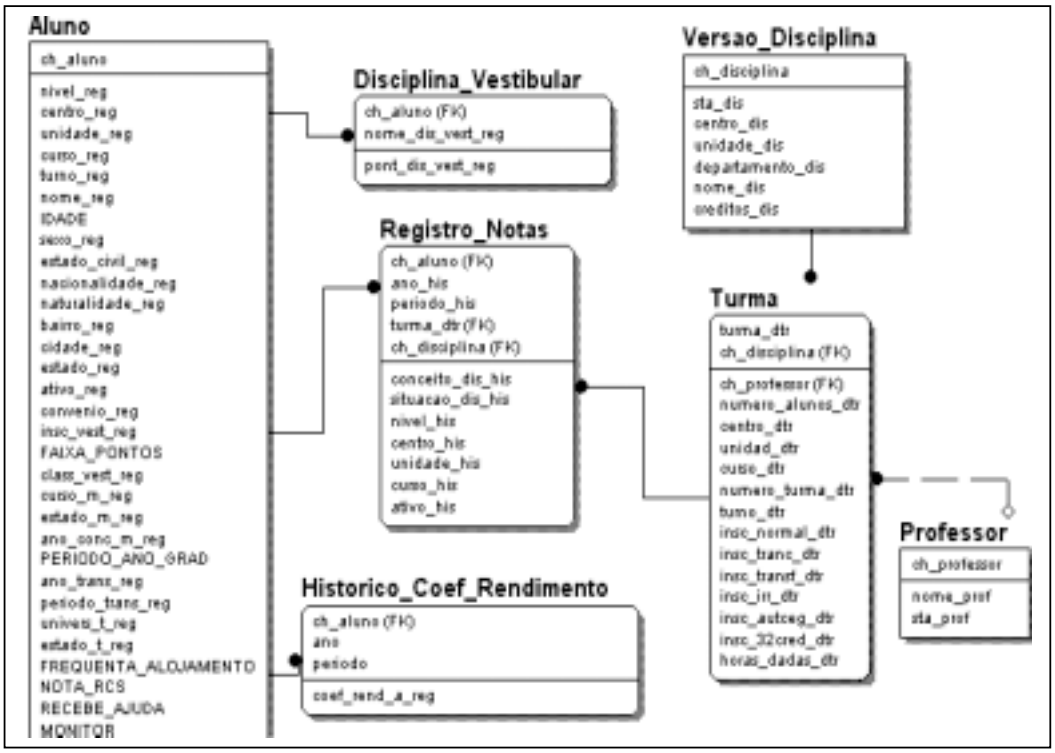

Figura 5.4 – Pré_modelo resultante da Fase B

FASE C – ELABORAR O MODELO DIMENSIONAL

C.I) – Realizar o Levantamento de Fatos Básicos e Visões Dimensionais:

Entrada: Pré-Modelo.

Saída: Pré-Modelo e lista de fatos básicos.

A partir do pré-modelo gerado na FASE B observa-se os seguintes fatos básicos:

- Registro de coeficiente de rendimento do aluno por ano/período;
- Registro de pontos nas disciplinas do vestibular por aluno;
- Registro de conceito e situação de aluno por disciplina em ano/período; e
- Registro de aluno inscritos, trancados, transferidos e com inscrição irregular em disciplinas por ano/período.

Esses fatos básicos permitem estabelecer as seguintes visões:

a) Para atender à avaliação de desempenho:

- Acompanhamento dos alunos através do coeficiente de rendimento;
- Acompanhamento com base nas notas das disciplinas; e
- Acompanhamento de cancelamento e trancamento de disciplinas.

b) Para atender à análise de disciplinas:

- Avaliação das disciplinas com maior número de trancamento/cancelamento; e
- Avaliação do desempenho das disciplinas/turmas com relação à média dos alunos.

c) Para analisar o desempenho no vestibular:

- Avaliação das notas no vestibular.

C.II) Derivar os Modelos Dimensionais

Entrada: Pré-modelo e lista de fatos básicos resultantes da subfase anterior.

Saída: Modelos Dimensionais do DM.

Esta subfase será realizada para cada fato básico identificado na subfase anterior.

FATO BÁSICO: registro de coeficiente de rendimento do aluno por ano/período

C.II.a) Selecionar Entidade Chave Relacionada ao Fato Básico

A entidade selecionada para servir como fato básico foi *Historico_Coef_Rendimento.* A tabela de fatos a ser criada será denominada CONTROLE_COEFICIENTE_RENDIMENTO.

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.5(I) foi gerada ao se transformar a entidade *Historico_Coef_Rendimento* em raiz. Aplicando o processo de limpeza com base na cardinalidade, as subárvores referentes aos nós *Disciplina_Vestibular* e *Registro_Notas* são removidas, gerando a árvore representada na figura 5.5(II).

C.II.b.iii) – Criar a Dimensão Tempo

A entidade *Historico_Coef_Rendimento* foi criada com a periodicidade ano/período para ser compatível com o SCG.

A nível de análise, o usuário final deseja manter a mesma periodicidade, portanto será criada a dimensão TEMPO com os atributos: *ANO* e *PERÍODO*.

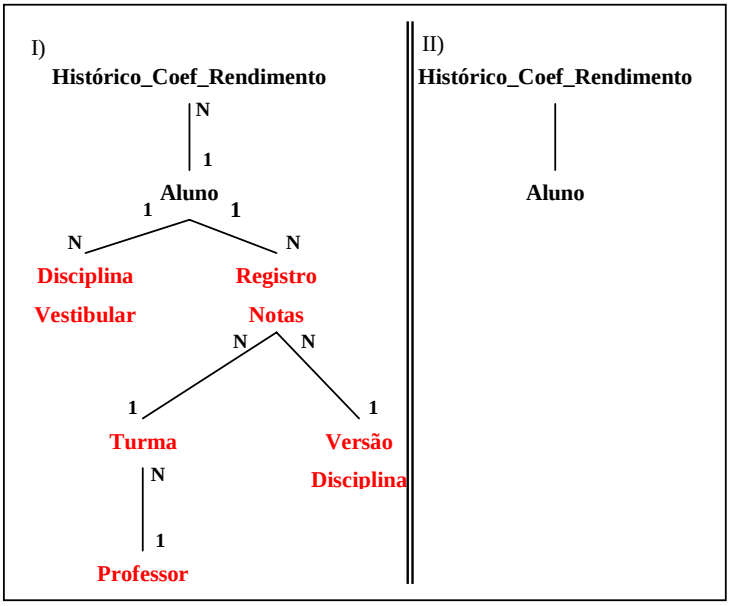

Figura 5.5 - Árvore de *Historico Coef Rendimento.* (I) original. (II) Transformada

C.II.c) – Refinar Dimensões:

O esboço do modelo dimensional gerado apresenta as dimensões **ALUNO** e **TEMPO** e a tabela de fatos **CONTROLE_COEF_RENDIMENTO.** Este modelo passa a ser refinado.

C.II.c.iii) - Elaborar Minidimensões:

A dimensão **ALUNO** permite a criação de minidimensões constituídas por grupos de informações relacionadas. A criação de minidimensões, nesse caso, garante um melhor desempenho nas consultas e uma melhor visualização da dimensão. São criadas três minidimensões. A primeira com atributos referentes a identificação de curso selecionado pelo aluno, a segunda com informações demográficas e a terceira com informações de localidade. Estas minidimensões serão denominadas respectivamente de **CURSO, DEMOGRÁFICA E ESPACIAL**. A relação das minidimensões geradas e seus respectivos atributos estão no anexo 4, item A4-7- Definição de minidimensões para a dimensão Aluno no DER do SCG

C.II.c.iv) Estabelecer Hierarquias:

As dimensões **CURSO**, **ESPACIAL** e **TEMPO** apresentam as hierarquias representadas na figura 5.6 (I, II e III), respectivamente.

As hierarquias estabelecidas para as dimensões da figura 5.6, podem ser definidas como hierarquias explícitas da dimensão **ALUNO**.

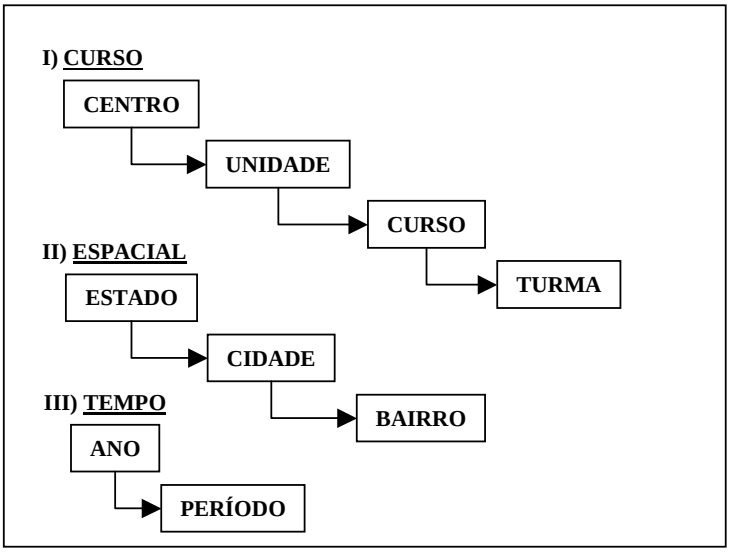

Figura 5.6 - Hierarquias das Minidimensões: CURSO (I), ESPACIAL (II) e da Dimensão TEMPO (III)

C.II.d.i) Tratar Atributos de Acordo com Tipo de Fato

O fato CONTROLE_COEF_RENDIMENTO representa um fato do tipo linha de item, apresentando a granularidade desejada (Ano/Período). Não é necessário nenhum tratamento de atributo.

C.II.d.iii) Classificar os Atributos da Tabela de Fato

Chaves de dimensões: CH_ALUNO,
CH_TEMPO,
CH_CURSO,
CH_DEMOGRAFICA, e
CH_ESPACIAL.

Valor não aditivo: COEF_REND_A_REG

O modelo dimensional para o fato CONTROLE_COEF_RENDIMENTO está representado na figura 5.7.

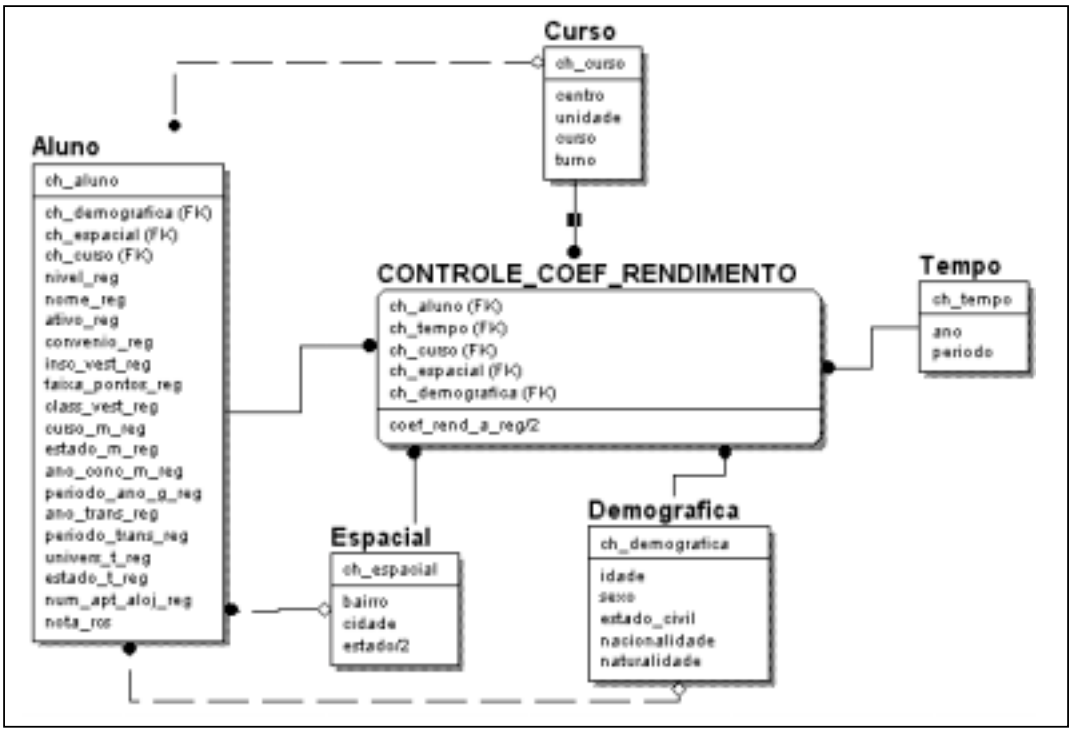

Figura 5.7 – Modelo dimensional para CONTROLE_COEF_RENDIMENTO

FATO BÁSICO: registro de pontos nas disciplinas do vestibular por aluno

C.II.a) Selecionar Entidades Chaves Relacionadas aos Fatos :

A entidade selecionada para servir como fato básico foi *Disciplina_Vestibular*. A tabela de fatos a ser criada será denominada CONTROLE_VESTIBULAR.

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.8(I) foi gerada ao transformar a entidade *Disciplina_Vestibular* em raiz. Aplicando o processo de limpeza com base na cardinalidade, as subárvores referentes aos nós *Historico_Coef_Rend* e *Registro_Notas* são removidas, gerando a árvore representada na figura 5.8(II).

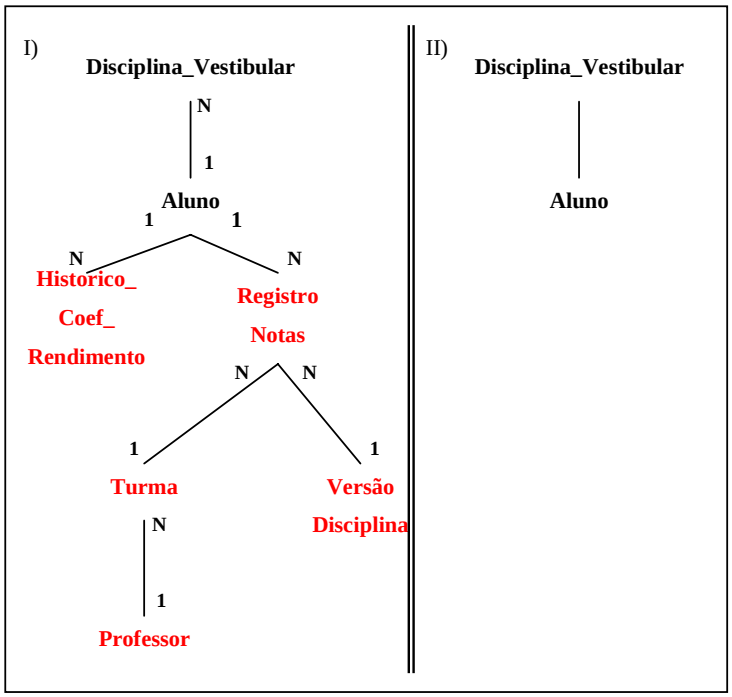

Figura 5.8 - Árvore de *Disciplina_Vestibular*.

(I) original. (II) Transformada

C.II.b.iii) Criar a Dimensão Tempo

As informações de vestibular são armazenadas no primeiro período de cada ano. Entretanto, pode ocorrer da universidade abrir vestibular no meio do ano. Portanto é interessante manter a chave tempo com período/ano.

C.II.c.iii) Elaborar Minidimensões:

Como a entidade *Aluno* foi tratada para a primeira visão dimensional, não precisa ser tratada novamente. Neste caso, a entidade *Aluno* é automaticamente desmembrada em suas minidimensões.

C.II.c.iv) Estabelecer Hierarquias:

As hierarquias são as mesmas válidas no primeiro modelo.

C.II.d.i) Analisar os Atributos com Relação ao Tipo de Fato

O fato CONTROLE_VESTIBULAR representa um fato do tipo linha de item, apresentando a granularidade desejada (Ano/Período). Não é necessário nenhum tratamento de atributo.

C.II.d.iii) Classificar os Atributos

Chaves de Dimensão: *CH_ALUNO, CH_CURSO, CH_DEMOGRAFICA, CH_TEMPO*
Valor Aditivo: *PONT_DIS_VEST_REG*

O atributo *NOME_DIS_VEST_REG* representa o nome da disciplina no vestibular. Esse atributo poderia ser caracterizado como uma dimensão descaracterizada, porém pelo fato de ser descritivo é necessário avaliar se existe a possibilidade de transformá-lo em uma dimensão. No modelo em questão será criada a dimensão **DISCIPLINA_VESTIBULAR** cuja chave será *CH_DISCIPLINA_VESTIBULAR*. Esta dimensão conterá o atributo descritivo *NOME_DISCIPLINA_VESTIBULAR* originado do atributo *NOME_DIST_VEST_REG*.

O atributo descritivo da tabela de fatos será substituído por *CH_DISCIPLINA_VESTIBULAR*. Essa nova dimensão apenas será atualizada quando uma nova disciplina for criada.

O modelo dimensional gerado para o fato CONTROLE_VESTIBULAR está representado na figura 5.9.

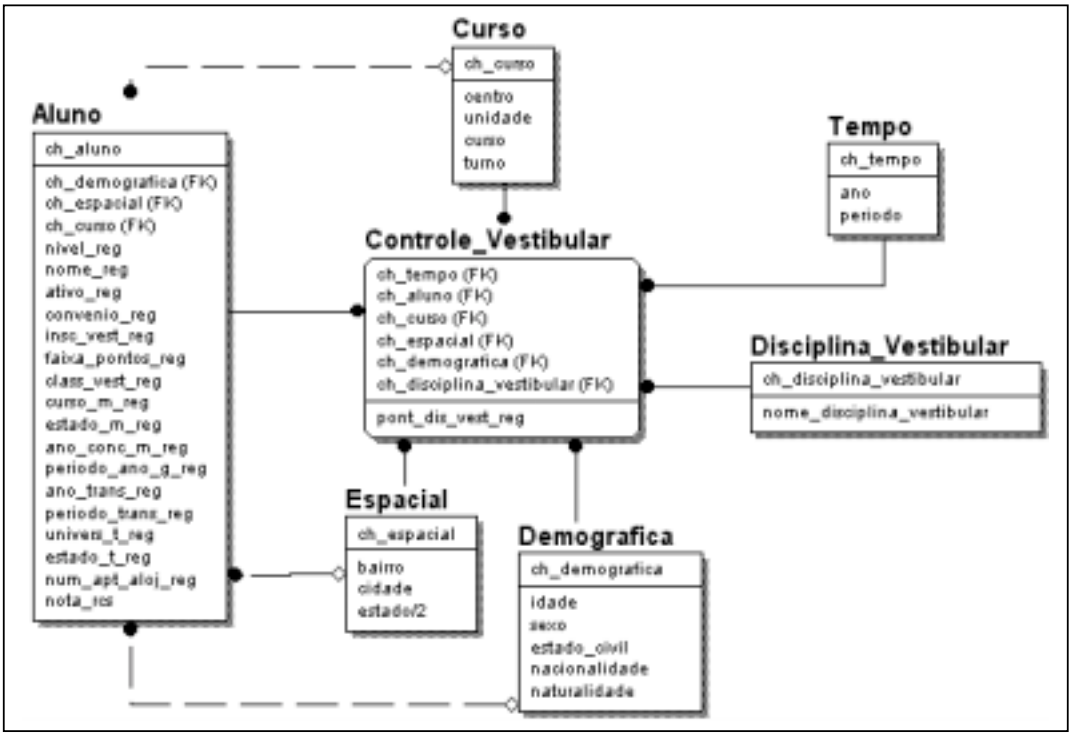

Figura 5.9 – Modelo Dimensional para CONTROLE_VESTIBULAR

FATO BÁSICO: registro de conceito e situação de aluno por disciplina em ano/período

C.II.a) Selecionar Entidades Chaves Relacionadas aos Fatos :

A entidade selecionada para servir como fato básico foi *Registro_Notas.* A tabela de fatos a ser criada será denominada CONTROLE_NOTAS

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.10(I) foi gerada ao se transformar a entidade *Registro_Notas* em raiz. Aplicando o processo de limpeza com base na cardinalidade, as subárvores referentes aos nós *Disciplina_Vestibular* e *Historico_Coef_Rend* são removidas.

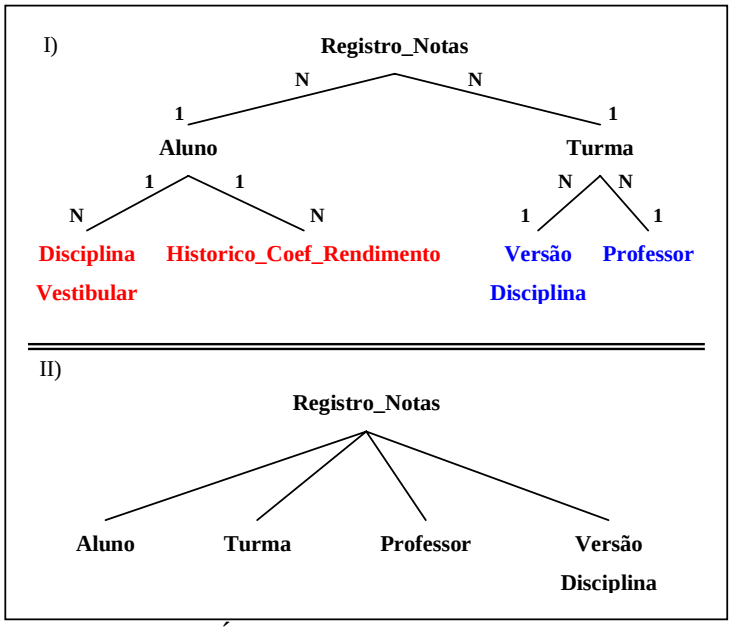

Figura 5.10 - Árvore de *Registro Notas.* (I) Original.
(II) Podada

C.II.b.ii) Realizar a Poda e Enxerto:

Aplicando à árvore o processo de poda das entidades *Versao_Disciplina* e *Professor* e as enxertando na raiz, é obtida a árvore representada na figura 5.10(II).

C.II.b.iii) Criar a Dimensão Tempo

As informações de disciplinas são armazenadas a cada período. Portanto, a granularidade será período/ano.

C.II.c.iii) Elaborar Minidimensões:

Como a entidade *Aluno* foi tratada para a primeira visão dimensional, não precisa ser tratada novamente, sendo desmembrada em suas minidimensões.

C.II.c.iv) Remover Atributos não Relacionados ao Modelo:

A entidade *Turma* apresenta um conjunto de informações que não dizem respeito a análise. São elas: *NUMERO_ALUNOS, INSC_NORMAL_DTR, INSC_TRANC_DTR, INSC_TRANSF_DTR, INSC_IRR_DTR, INSC_AUTCEG_DTR, INS_32CRED_DTR, HORAS_DADAS.*

De uma forma geral, o único atributo em turma que interessa para a visão em questão é o turno em que a disciplina foi cursada. Ao invés de manter a dimensão **TURMA,** uma alternativa nesse caso é criar na tabela de fatos o atributo *TURNO*. Este atributo representará uma dimensão descaracterizada.

C.II.c.v) Estabelecer Hierarquias:

As hierarquias são as mesmas válidas no primeiro modelo, adicionando-se a hierarquia de *Versão_Disciplina:*

CENTRO_DIS → UNIDADE_DIS → CURSO_DIS.

C.II.d.i) Analisar os Atributos com Relação ao Tipo de Fato

O fato é extraído da entidade origem *Historico_disciplina*, sendo, portanto do tipo Linhas de Item. A granularidade é a própria do registro no histórico.

C.II.d.iii) Classificar os Atributos

Chaves de Dimensão: *CH_ALUNO, CH_TEMPO, CH_DISCIPLINA, CH_PROFESSOR, CH_MD_CURSO, CH_DEMOGRAFICA, CH_ESPACIAL*

Valor Aditivo: *CONCEITO_DIS_HIST*

Dimensões Descaracterizadas: *TURNO*

SITUAÇÃO_DIS_HIST

O modelo dimensional resultante do fato CONTROLE_NOTAS é representado na figura 5.11.

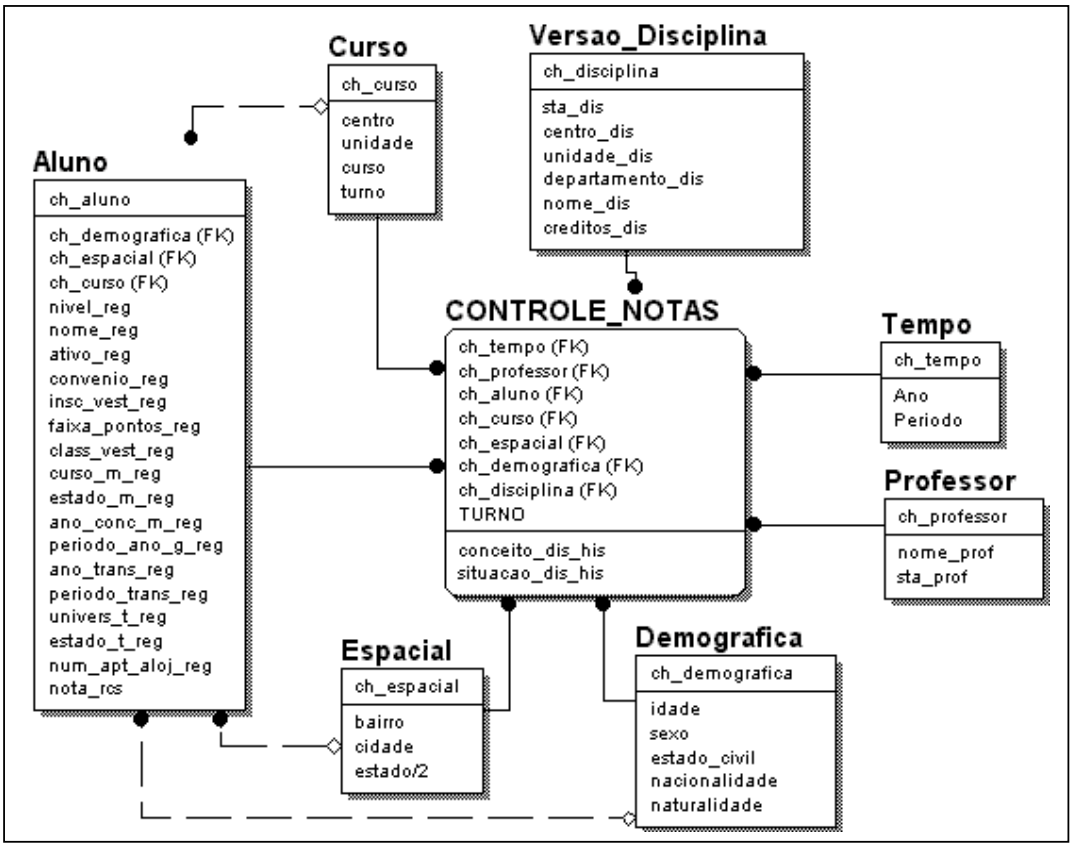

Figura 5. 11 – Modelo Dimensional para CONTROLE_NOTAS

FATO BÁSICO: registro de aluno inscritos, trancados, transferidos e com inscrição irregular em disciplinas por ano/período

C.II.a) Selecionar Entidades Chaves Relacionadas aos Fatos :

A entidade selecionada para servir como fato básico foi *Turma.* A tabela de fatos a ser criada será denominada CONTROLE_TURMA.

C.II.b.i) Construir a árvore:

A árvore representada na figura 5.12(I) foi gerada ao se transformar a entidade *Turma* em raiz. Aplicando o processo de limpeza com base na cardinalidade, a subárvore referente ao nó *Registro_Notas* é removida, gerando a árvore representada na figura 5.12(II).

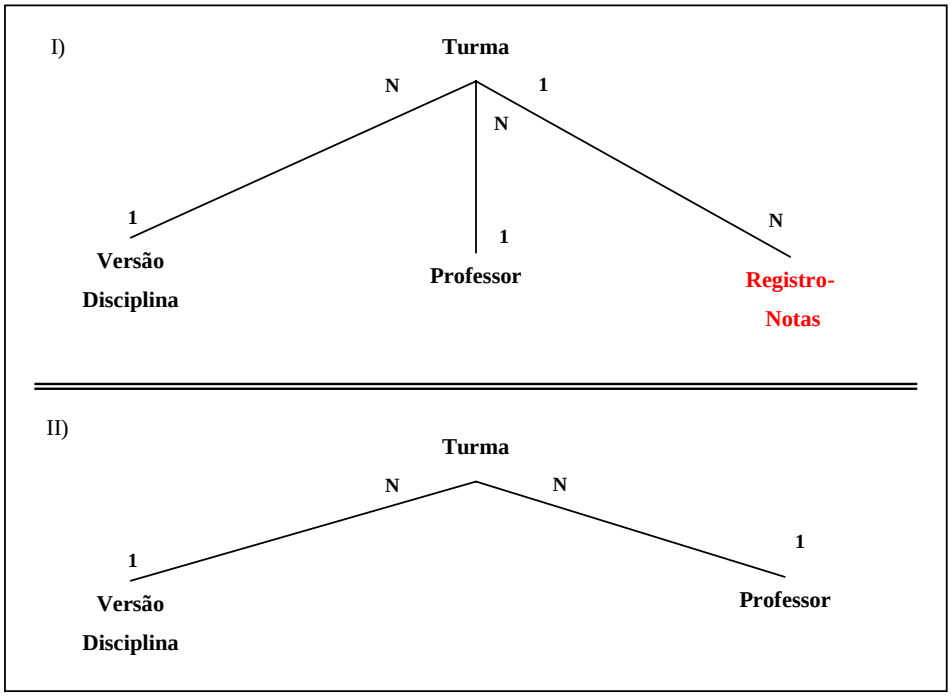

Figura 5.12 - Árvore de *Turma*. (I) Original. (II) Podada.

C.II.b.iii) Criar a Dimensão Tempo

As informações de turma são armazenadas a cada período. Portanto, a granularidade será período/ano.

C.II.c.v) Estabelecer Hierarquias:

As hierarquias são as mesmas definidas para *Versão_Disciplina* no segundo modelo.

C.II.d.i) Analisar os Atributos com Relação ao Tipo de Fato

O fato CONTROLE_TURMA representa um fato do tipo linha de item, apresentando a granularidade desejada (Ano/Período).

C.II.d.iii) Classificar os Atributos da tabela de fatos

Chaves de Dimensão: *CH_TEMPO, CH_DISCIPLINA, CH_PROFESSOR*

Valores Aditivos: *NUMERO_ALUNOS_DTR*

INSC_NORMAL_DTR, INSC_TRANC_DTR, INSC_TRANSF_DTR, INSC_IRR_DTR, INSC_AUTCEG_DTR, INS_32CRED_DTR.

Valores Semi_Aditivos: *HORAS_DADAS.*

Os atributos *NUMERO_TURMA_DTR, TURNO_DTR, CENTRO_DTR, UNIDADE_DTR, CURSO_DTR* na tabela de fatos se referem a entidade *Turma* original. Caso essas informações sejam importantes para a análise é necessário criar a dimensão **TURMA** contendo esses atributos. Se os atributos não fossem importantes eles seriam removidos. No estudo de caso, foi criada a dimensão **TURMA.**

O modelo dimensional resultante para o fato CONTROLE_DISCIPLINA está representado na figura 5.13.

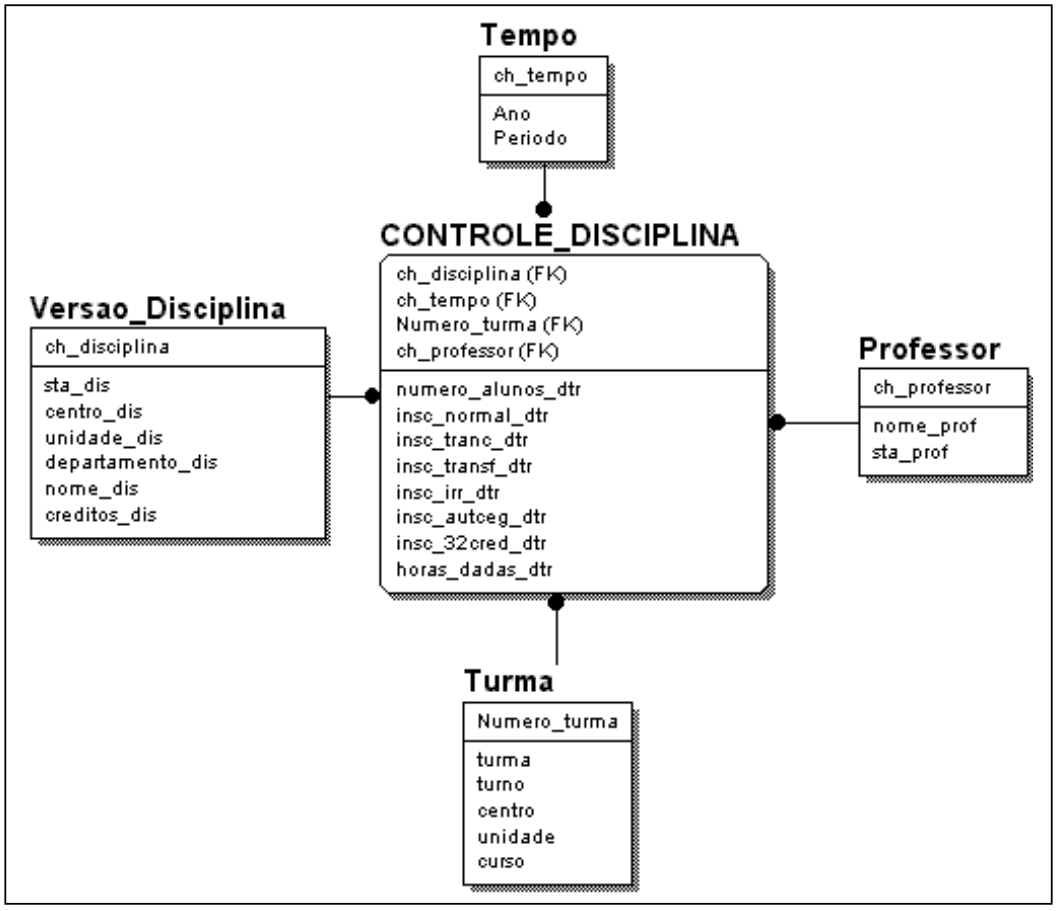

Figura 5.13 – Modelo Dimensional CONTROLE_TURMA

FASE D – INTEGRAR O DM AO DW

Por ser o primeiro DM a ser estabelecido, o modelo do DM será o próprio modelo do DW. Desse modo, na visão Kimball, o modelo do DW será a união dos modelos dimensionais que compõem o DM. A figura 5.14 apresenta o modelo do DW na visão dimensional.

Na visão Inmon, o modelo de DW é integrado com o pré-modelo. Como não existe nenhum modelo para realizar a integração, o pré-modelo automaticamente se transforma no modelo do DW. A figura 5.7 representa o modelo do DW na visão de Inmon.

Figura 5.14 – Modelo Dimensional para o DW

5.2 DM Vestibular

FASE A – ESTUDAR OS MODELOS EXISTENTES

O modelo a ser utilizado no desenvolvimento desse DM é o DER referente ao Sistema Vestibular (SV). A versão simplificada desse DER é apresentada na figura 5.15. Pela simplicidade do DER, ele será empregado na próxima fase.

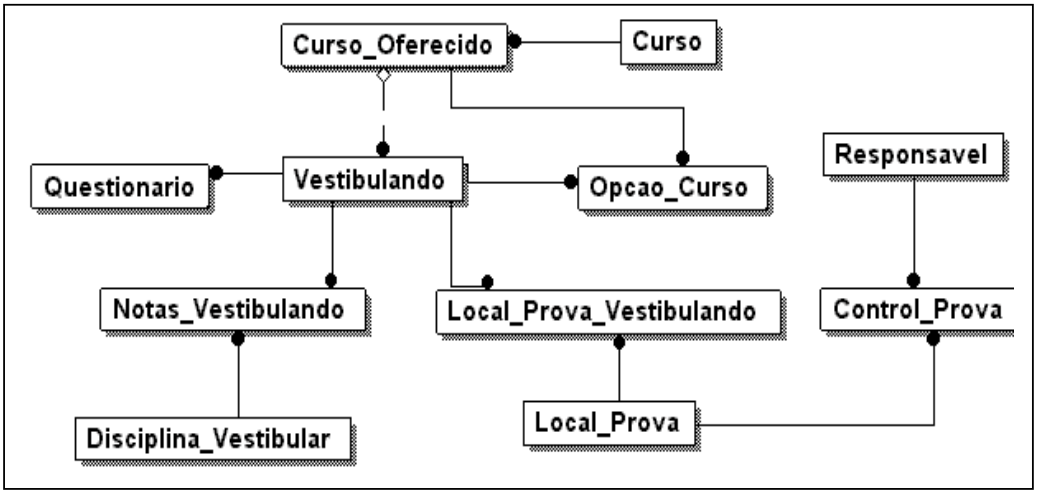

Figura 5.15- DER simplificado do Sistema Vestibular

FASE B – ELABORAR O PRÉ-MODELO

Entrada: DER resultante da fase anterior.

Saída : Pré-modelo.

O DER do SV contendo as entidades e relacionamentos selecionados está representado na figura 5.16.

Figura 5.16 - DER resultante da FASE A

B.I) Limpar e Transformar Modelos:

Entrada: DER resultante da fase anterior.

Saída: DER limpo e transformado.

B.I.a) Excluir Informações Desnecessárias:

Algumas entidades no modelo mantém informações operativas ou que não interessam a análise. Neste modelo as entidades com informações que não interessam à análise são aquelas empregadas para controlar os responsáveis e os locais de prova. Dessa forma, são removidas as seguintes entidades:

- *Responsavel:* Informações operativas de controle dos responsáveis em aplicar as provas;
- *Controle_Prova, Local_Prova* e *Local_Prova_Vestibulando:* Informações operativas de controle do local das provas.

Para as entidades que permanecem no modelo é necessária uma análise no nível dos atributos. A relação dos atributos (por entidade) excluídos por representarem informações desnecessárias à análise encontra-se no Anexo 4, item A4-8- Relação de Atributos excluídos no DER do SV.

B.I.b) Desnormalizar Relações:

No DER é possível observar uma desnormalização para as entidades: *Vestibulando* e *Questionario.*

Analisando a validade da desnormalização para estas entidades é possível observar que compartilham a chave primária, apresentam os dados normalmente juntos e possuem um padrão de inserção semelhante. Portanto, podem ser desnormalizadas. A entidade gerada pela desnormalização é nomeada como *Registro_Vestibulando.*

B.I.c) Definir Categorias:

A entidade *Registro_Vestibulando* apresenta as seguintes informações a serem categorizadas: data de nascimento, pontuação vestibular e classificação no vestibular.

Os atributos *IDADE, FAIXA_PONTOS_VEST e FAIXA_CLASSIFICACAO* serão criados contendo as informações categorizadas e substituindo os atributos *DATA_NASCIMENTO, PONTOS_VESTIBULAR* e *CLASSIFICACAO_VESTIBULAR.*

As regras para estabelecer as categorias estão definidas no anexo 4, item A4-9- Regras de categorias apresentadas no DER do SV.

O modelo resultante da limpeza e tratamento do DER do SV é apresentado na figura 5.17.

5.17 – DER resultante da Subfase Limpeza e Transformação

B.III.a) Criar Novas Chaves:

O SV não armazena histórico de informações. As informações sobre vestibulandos e cursos oferecidos são armazenados em fitas magnéticas para posteriores análises estatísticas. Entretanto, segundo os ABD e usuários do SCV, a matricula do vestibulando é gerada compondo o ano do vestibular a um número seqüencial. Desta forma não existe a possibilidade de reutilização de matrícula. O mesmo não acontece com a entidade *Curso_Oferecido*. O atributo *COD_CURSO_OFERECIDO* é formado por código do curso e semestre em que o curso é fornecido. Neste caso, será realizada a inserção da chave tempo na entidade (B.III.c).

B.III.b) Analisar Periodicidade de Atualização das Informações das Entidades:

O processo de carga dos dados é executado após a realização do concurso vestibular. A partir da carga, os dados não são mais alterados. Deste modo, as atualizações existentes são acertos e inclusões de novos registros a cada concurso.

B.III.c) Inserir a Chave Tempo:

Para tratar as informações de cursos oferecidos ao longo dos anos, o atributo *ANO_VESTIBULAR* será inserido na entidade *Curso_Oferecido.*

B.III.d e B.III.e) Estabelecer Padrões, Valores "Default" e Regras de Conversão para Substituir Código e Abreviaturas dos Atributos do Modelo:

Neste estudo de caso as entidades *Registro_Vestibulando* e *Curso* necessitam de tratamento para alguns de seus atributos. A relação destes atributos com seus respectivos tratamentos estão no anexo 4, item A4-10- Definição de padrão, valores "default" e regras de conversão para os atributos no DER do SV.

B.III.f) Criar Atributos Derivados:

Para atender as consultas do usuário final serão criados na entidade *Curso_Oferecido* os seguintes atributos: *TOTAL_INSCRITOS_1OPCAO, TOTAL_INSCRITOS_2OPCAO, TOTAL_INSCRITOS_3OPCAO* e *TOTAL_CLASSIFICADOS.* A definição das regras para estes atributos está no anexo 4, item A4-11- Regras para cálculo de atributos derivados no DER do SV.

A figura 5.18 apresenta o pré-modelo gerado na fase B a partir do DER do SV.

5.18 – Pré-Modelo resultante da Fase B

FASE C – ELABORAR O MODELO DIMENSIONAL

C.I) – Realizar Levantamento de Fatos Básicos e Visões Dimensionais:

Entrada: Pré-Modelo.

Saída: Pré-Modelo e lista de fatos básicos.

A partir do pré-modelo gerado na FASE B, são identificados os seguintes fatos básicos:

- Registro de pontos nas disciplinas do vestibular por candidato;
- Registro de vagas fornecidas, inscritos e classificados por curso/período ao longo dos anos; e
- Registro de opções de curso por vestibulando ao longo dos anos.

Esses fatos básicos permitem estabelecer as seguintes visões:

- Acompanhamento dos vestibulandos através das médias e nota em disciplinas;
- Acompanhamento das médias por perfil de vestibulando;
- Acompanhamento das notas em disciplinas por perfil de vestibulando;
- Análise de perfil dos vestibulandos não classificados e dos classificados;
- Análise dos cursos mais procurados e
- Avaliação das disciplinas com menores e maiores notas nos últimos anos.

C.II) Derivar os Modelos Dimensionais

Entrada: Pré-modelo e lista de fatos básicos resultantes da subfase anterior.

Saída: Modelos Dimensionais do DM.

Essa subfase será realizada para cada fato básico identificado na subfase anterior.

FATO BÁSICO: registro de pontos nas disciplinas do vestibular por aluno.

C.II.a) Selecionar Entidade Chave Relacionada ao Fato Básico:

A entidade selecionada para servir como fato básico foi *Notas_Vestibulando.* A tabela de fatos a ser criada será denominada CONTROLE_NOTAS_VESTIBULAR.

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.19(I) foi gerada ao se transformar a entidade *Notas_Vestibulando* em raiz. Aplicando o processo de limpeza com base na cardinalidade, as subárvores referentes aos nós *Opcao_Curso*, *Curso_Oferecido* e *Curso* são removidos, gerando a árvore representada na figura 5.19(II).

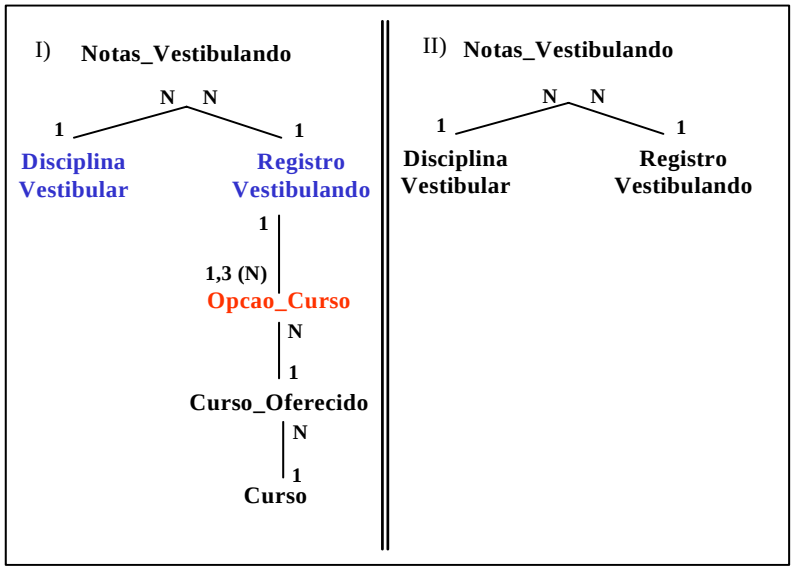

Figura 5.19 - Árvore referente a *Notas_Vestibulando.* (I) Original. (II) Tratada.

C.II.b.iii) – Criar a Dimensão Tempo

A entidade *Notas_Vestibulando* registra as notas dos vestibulandos nas disciplinas para um vestibular. Portanto, as informações estão registradas por ano. A dimensão tempo será criada com o atributo *ANO*.

C.II.c) – Refinar Dimensões:

Essa etapa recebe o esboço do modelo dimensional, apresentado as dimensões: **REGISTRO_VESTIBULANDO, DISCIPLINA_VESTIBULAR** e **TEMPO**.

C.II.c.iii) - Elaborar Minidimensões:

A dimensão **REGISTRO_VESTIBULANDO** permite a criação de minidimensões constituídas por grupos de informações relacionadas. Do mesmo modo, com na dimensão **ALUNO,** sua criação garante um melhor desempenho nas consultas e uma melhor visualização da dimensão. São criadas quatro minidimensões: a primeira

com atributos referentes a ensino, a segunda com as informações relacionadas ao vestibular na UFRJ, a terceira com informações sócio-econômicas e a quarta com informações referentes a classificação no vestibular. Estas minidimensões serão denominadas respectivamente de **ENSINO, VESTIBULAR, SOCIO_ECONOMICO** e **CLASSIFICACAO**. A relação das minidimensões geradas e seus respectivos atributos está no anexo 4, item A4-12- Definição de minidimensões para a dimensão **REGISTRO_VESTIBULANDO** no DER do SV

A dimensão **REGISTRO_VESTIBULANDO** e suas minidimensões estão representadas na figura 5.20.

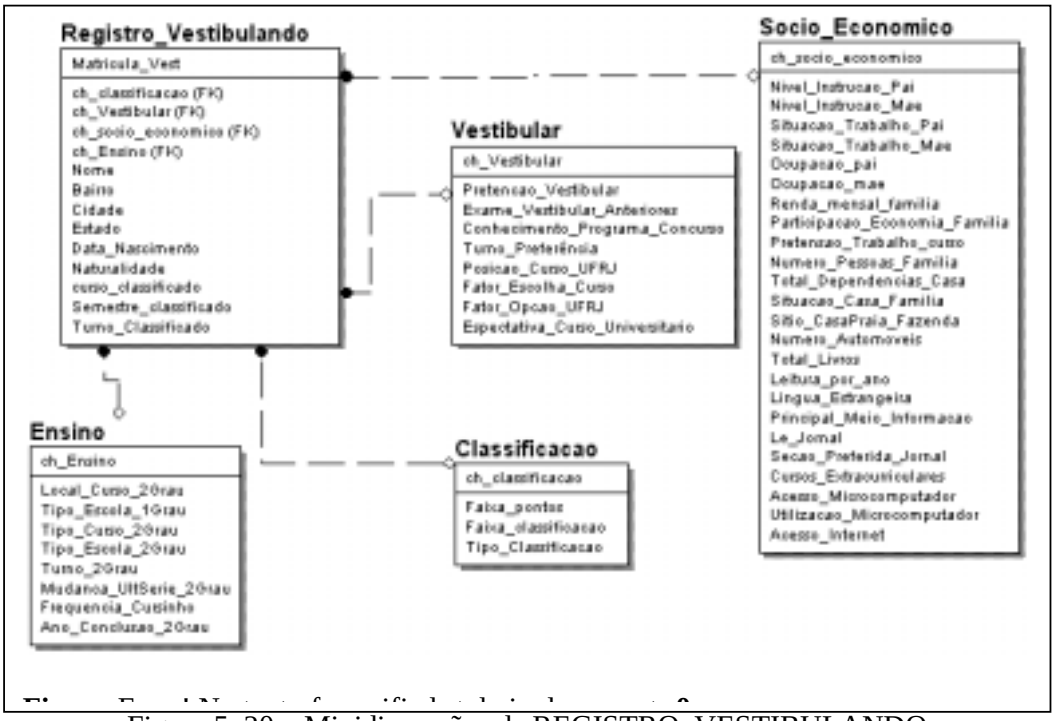

Figura 5. 20 – Minidimensões de REGISTRO_VESTIBULANDO

C.II.c.iv) Estabelecer Hierarquias:

Para este modelo são obtidas as seguintes hierarquias:

A dimensão **REGISTRO_VESTIBU-LANDO** apresenta a hierarquia representada na figura 5.21.

A minidimensão **CLASSIFICACAO** apresenta a seguinte hierarquia: *TIPO_CLASSIFICACAO → FAIXA_CLASSIFICACAO → FAIXA_PONTOS.*

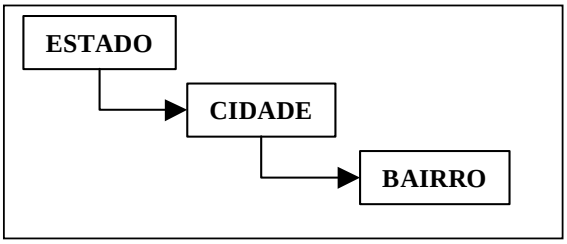

Figura 5.21 – Hierarquia

C.II.d.i) Tratar Atributos de Acordo com Tipo de Fato

O fato CONTROLE_NOTAS_VESTIBULAR representa um fato do tipo linha de item, apresentando a granularidade desejada (Ano).

C.II.d.iii) Classificar os Atributos da Tabela de Fato

Chaves de dimensões: MATRICULA_VEST, CH_TEMPO,
CH_DISCIPLINA_VESTIBULAR,CH_ENSINO,
CH_VESTIBULAR, CH_SOCIO_ECONOMICO,
CH_OPCOES.

Valor aditivo: NOTA_DISCIPLINA (Média)

O modelo dimensional para o fato CONTROLE_NOTAS_VESTIBULAR está representado na figura 5.22.

Figura 5.22 – Modelo Dimensional Controle Vestibular

FATO BÁSICO: registro de vagas fornecidas por curso/período ao longo dos anos.

C.II.a) Selecionar Entidade Chave Relacionada ao Fato Básico:

A entidade selecionada para servir como fato básico foi *Curso_Oferecido.* A tabela de fatos a ser criada será denominada CONTROLE_CURSO.

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.23(I) foi gerada ao se transformar a entidade *Curso_Oferecido* em raiz. Aplicando o processo de limpeza com base na cardinalidade, as subárvores referentes aos nós *Opcao_Curso, Registro_Vestibulando, Notas_Vestibular e Disciplina_Vestibular* são removidos, gerando a árvore representada na figura 5.23(II).

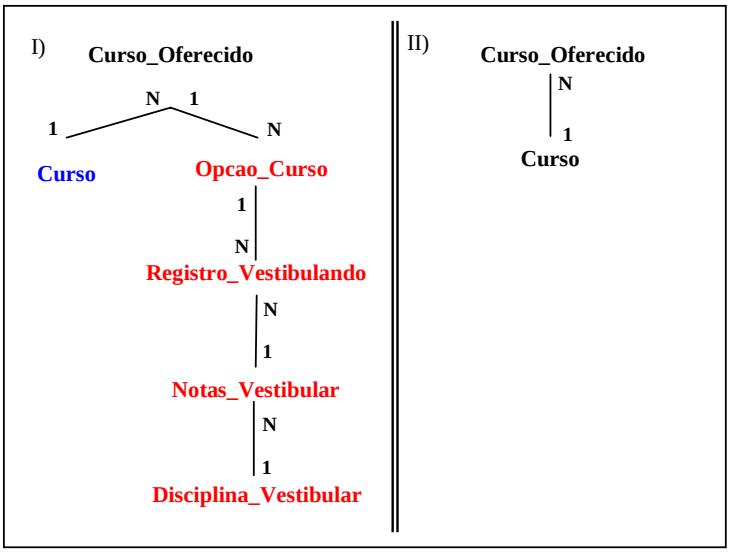

Figura 5.23-Arvore referente a *Curso_Oferecido.*
(I) Original e (II) Tratada

C.II.b.iii) – Criar a Dimensão Tempo

A entidade *Curso_Oferecido* armazena as informações por ano. Portanto, a dimensão tempo empregada, apresenta o atributo *ANO.*

C.II.c) – Refinar Dimensões:

O esboço do modelo dimensional gerado apresenta as dimensões **CURSO** e **TEMPO** e a tabela de fatos **CONTROLE_CURSO.** Esse modelo passa a ser refinado.

C.II.c.iv) Estabelecer Hierarquias:

A dimensão **CURSO** apresenta a seguinte hierarquia: *UNIDADE → CENTRO.*

C.II.d.i) Tratar Atributos de Acordo com Tipo de Fato

O fato CONTROLE_CURSO representa um fato do tipo linha de item, apresentando a granularidade desejada, armazenando informações por Ano.

C.II.d.iii) Classificar os Atributos da Tabela de Fatos

Chaves de dimensões: CH_TEMPO, CH_CURSO.

Valor aditivo: TOTAL_VAGAS, TOTAL_INSCRITOS_OPCAO1, TOTAL_INSCRITOS_OPCAO2, TOTAL_INSCRITOS_OPCAO3, e TOTAL_CLASSIFICADOS.

Ao final da classificação sobram os atributos *COD_CURSO_OFERECIDO, SEMESTRE* e *TURNO.* Por conter a chave para ano e semestre, o atributo *COD_CURSO_OFERECIDO* será removido. Pelas características é possível observar que o atributo *SEMESTRE* é uma informação da dimensão **TEMPO** e o atributo *TURNO* da dimensão **CURSO.** Portanto, os atributos em questão passarão a compor as respectivas dimensões.

O modelo dimensional para o fato CONTROLE_CURSO está representado na figura 5.24.

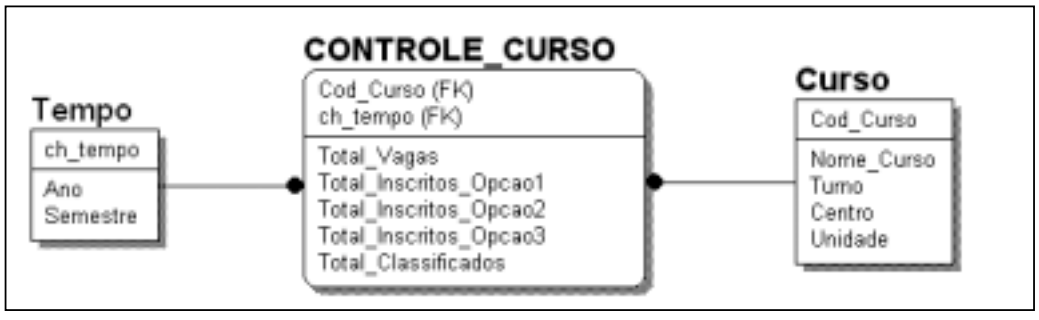

Figura 5.24 – Modelo Dimensional CONTROLE_CURSO

FATO BÁSICO: registro de opções de curso por vestibulando ao longo dos anos.

C.II.a) Selecionar Entidade Chave Relacionada ao Fato Básico:

A entidade selecionada para servir como fato básico foi *Opcao_Curso.* A tabela de fatos a ser criada será denominada CONTROLE_OPCAO.

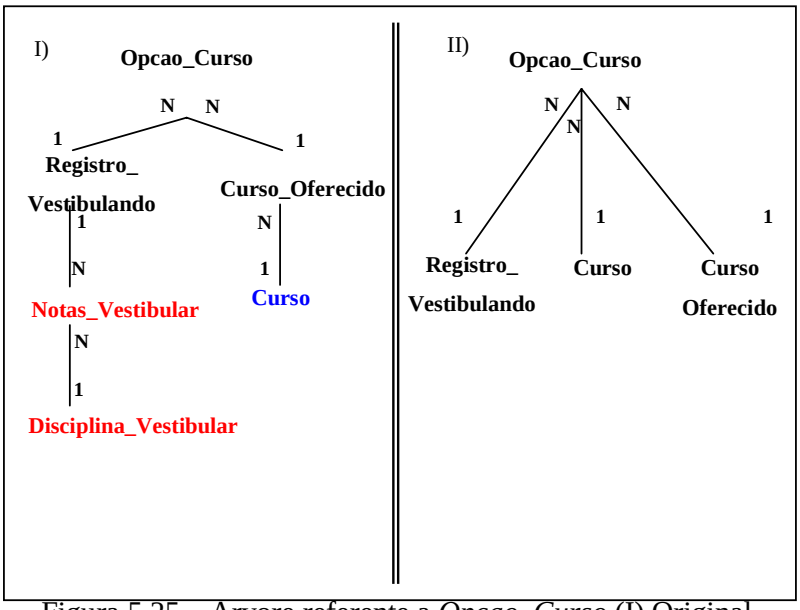

Figura 5.25 – Arvore referente a *Opcao_Curso.*(I) Original
(II) Podada/Enxertada

C.II.b.i) Construir a Árvore:

A árvore representada na figura 5.25(I) foi gerada ao se transformar a entidade *Opcao_curso* em raiz. Aplicando o processo de limpeza com base na cardinalidade, a subárvore contendo os nós *Notas_Vestibular e Disciplina_Vestibular* é removida do modelo.

C.II.b.ii) Realizar a Poda e Enxerto:

Aplicando à árvore o processo de poda e enxerto, o nó *Curso* é enxertado na raiz, gerando a árvore representada na figura 5.25(II).

C.II.b.iii) – Criar a Dimensão Tempo:

A entidade *Opcao_Curso* registra as opções dos vestibulandos para cada ano de vestibular. Portanto, a dimensão tempo apresentará o atributo *Ano.*

C.II.c) – Refinar Dimensões:

Esta etapa recebe o esboço do modelo dimensional, apresentando as dimensões: **REGISTRO_VESTIBULANDO, CURSO_OFERECIDO, CURSO** e **TEMPO**. As dimensões **REGISTRO_VESTIBULANDO** e **CURSO** são dimensões que foram tratadas e serão substituídas no modelo.

C.II.c.iv) Remover Atributos Não Relacionados ao Modelo:

A dimensão **CURSO_OFERECIDO** apresenta os seguintes atributos para remoção: *TOTAL_VAGAS, TOTAL_INSCRITOS_1OPCAO, TOTAL_INSCRITOS_ 2OPCAO , TOTAL_INSCRITOS_3OPCAO, TOTAL_CLASSIFICADOS*. O atributo *TURNO* está representado na dimensão **CURSO** assim como o *COD_CURSO*, podendo ser removidos. Da mesma forma o atributo *ANO_VESTIBULAR* está presente através da chave tempo, não sendo necessário. Resta o atributo *SEMESTRE*. Por ser um atributo único, seria inserido na tabela de fatos como dimensão descaracterizada. Porém, analisando o seu conteúdo, percebe-se que esse atributo garante informação de tempo, sendo, dessa forma, inserido na dimensão **TEMPO**. A dimensão em questão será excluída do modelo.

C.II.c.v) Estabelecer Hierarquias:

A dimensão **CURSO**, como apresentado, possui a hierarquia Unidade → Centro.

C.II.d.i) Tratar Atributos de Acordo com Tipo de Fato

O fato CONTROLE_OPCAO representa um EVENTO. Seu armazenamento anual, garante a granularidade desejada (Ano). Não é necessário nenhum tratamento de atributo.

C.II.d.iii) Classificar os Atributos da Tabela de Fatos

Chaves de dimensões: CH_TEMPO, COD_CURSO, MATRICULA_VEST.

O modelo dimensional para o fato CONTROLE_OPCAO está representado na figura 5.26.

O modelo dimensional para o DM, integrando os modelos gerados anteriormente, está representado na figura 5.27.

Figura 5.26 – Modelo Dimensional **CONTROLE_OPCAO**

Figura 5.27 – Modelo Dimensional **DM VESTIBULAR**

FASE D – INTEGRAR O DM AO DW

Por ser o segundo DM a ser estabelecido, um rigoroso controle é realizado no momento da integração do modelo de dados do DM ao modelo do DW. A seguir serão apresentadas as duas formas de integração para o DW.

D.I.a) Integrar o DM a um DW Dimensional - VISÃO KIMBALL

Na visão Kimball, o modelo dimensional resultante para o DM é integrado com o modelo do DW.

Integrando o modelo apresentado na figura 5.27, ao modelo do DW representado pela figura 5.14 (Modelo DW dimensional) é possível inserir sem conflitos as seguintes tabelas de dimensões e de fatos : **CONTROLE_CURSO, CURSO, CONTROLE_OPCAO.**

A dimensão **REGISTRO_VESTIBULANDO** pode ser representada como uma minidimensão de aluno, substituindo-se o atributo *INSC_VEST_REG* na dimensão **ALUNO** pela chave dessa dimensão.

As minidimensões de **REGISTRO_VESTIBULANDO** podem ser incluídas sem nenhum problema de conflito ou redundância.

As seguintes dimensões são idênticas: **TEMPO_CURSO** e **TEMPO** e **DISCIPLINA_VESTIBULAR** e seu homônimo no modelo existente**.** No segundo caso, se observa nomes diferentes para atributos, porém as informações são as mesmas. Tanto para o primeiro como para o segundo caso, as entidades existentes permanecem.

As tabelas de fatos **CONTROLE_VESTIBULAR** no DW e **CONTROLE_NOTAS_VESTIBULAR** no DM Vestibular são equivalentes. Entretanto, a tabela de fatos do DM Vestibular considera todos os vestibulandos que realizaram o concurso. As minidimensões relacionadas a essa tabela de fatos permitem uma maior combinação de consultas. Dessa forma esta tabela de fatos substituirá a existente no DW. A consulta a essas informações pela dimensão **ALUNO** é obtida através da chave *MATRICULA_VEST*.

O modelo final resultante da integração do DM Vestibular ao DW está representado na figura 5.28.

Figura 5.28 – Modelo Dimensional do DW Universidade

D.I.b) Integrar o DM a um DW relacional tratado - VISÃO Inmon

Na visão Inmon, o modelo de DW é integrado com o pré-modelo.

Integrando o pré-modelo apresentado na figura 5.18, ao modelo do DW representado pela figura 5.4 (pré-modelo do DM Graduação) é possível inserir sem conflitos as entidades: *Opcao_Curso, Curso_Oferecido, Curso* e *Registro_Vestibulando.*

A entidade *Registro_Vestibulando* possui uma relação com a entidade *Aluno* já existente. Essa relação é possível pelo atributo *INSC_VEST_REG* na entidade *Aluno* que contém a matrícula dos alunos no vestibular.

A entidade *Disciplina_Vestibular* apresenta conflito com a entidade de mesmo nome existente. A entidade *Disciplina_Vestibular* pertencente ao pré-modelo (figura 5.18) substituirá a entidade existente no DW. Com essa substituição, para se obter as notas das disciplina no vestibular dos alunos da universidade é necessário acessar a entidade *Registro_Vestibulando* e partir desta acessar a entidade *Notas_Vestibulando*. A entidade *Notas_Vestibulando* do DM Vestibular contém as informações referentes às notas dos vestibulandos por disciplina. Consequentemente, esta entidade conterá as notas dos alunos.

O modelo do DW resultante da integração do pré-modelo do vestibular está representado na figura 5.29.

Com a integração é possível estabelecer consultas a nível de DW como, por exemplo, analisar o desempenho dos alunos com bom desempenho no vestibular de acordo com a situação socio-econômica ou pretensões para o vestibular. Estas novas consultas podem ser sugeridas aos usuários finais do DM graduação como uma nova versão.

Com a substituição de entidades e novos relacionamentos, uma revisão das consultas do DM Graduação é necessária.

Figura 5.29 – Modelo Relacional do DW Universidade

CAPÍTULO 6

ANÁLISE DA PROPOSTA

Como apresentado nos capítulos anteriores, a abordagem "bottom-up" representa uma das melhores alternativas para o desenvolvimento de um ADW. Entretanto, o que se observa são soluções pontuais para o desenvolvimento de DW ou DM e técnicas que abordam o desenvolvimento do ADW aplicando a abordagem "top-down" (SILVERSTON, INMON, GRAZIANO, 1997) (MEYER, CANNON, 1997). Uma característica comum às soluções e técnicas existentes é a superficialidade com que são tratadas.

Este capítulo tem o propósito de analisar as diretrizes, propostas neste trabalho, para a modelagem incremental dos dados em um ADW.

6.1 Técnicas e Recursos Empregados

Ao longo do desenvolvimento das diretrizes propostas no capítulo 4, recursos e técnicas foram adotados para auxiliar o processo de modelagem. Os recursos e técnicas utilizados na proposta podem ser divididos em três grupos:

- recursos e técnicas existentes e aplicados sem alteração;
- recursos e técnicas existentes que foram tratados para se moldarem às necessidades da proposta; e
- recursos e técnicas que foram desenvolvidos para que os objetivos da proposta fossem atingidos.

Estes grupos serão apresentados a seguir.

6.1.1 Recursos e Técnicas Existentes Utilizados

A seguir são apresentados os recursos e técnicas já existentes e que foram utilizados nas diretrizes propostas:

a) Criação de artefatos: a criação de artefatos empregada na fase de elaboração do pré-modelo apresenta os mesmos critérios definidos por Silverston (SILVERSTON,

INMON, GRAZIANO, 1997).

b) Criação de novas chaves: A necessidade de criação de novas chaves para as dimensões em um ADW seguem as características estabelecidas por Inmon (INMON, 1997).

c) Tratamento de atributos de acordo com a periodicidade de atualização: A análise da periodicidade de atualização é aplicada, no conjunto de diretrizes, para a elaboração do pré-modelo e não para o desenvolvimento de um DW. Entretanto, são empregados os tratamentos referenciados na literatura para gerenciar a necessidade de desmembramento de uma entidade. Este tratamento é referenciado em Inmon (INMON, 1997), Kimball (KIMBALL, 1996) (KIMBALL *et al*, 1998) e McGuff (MCGUFF, 1998).

d) Adição da chave tempo: As regras para a inserção da chave tempo em uma dimensão são as mesmas definidas por Silverston (SILVERSTON, INMON, GRAZIANO, 1997).

e) Estabelecimento de padrão e valor "default" de atributos no ADW: Este tratamento surge em virtude do emprego de fontes diversas para a modelagem do DM (INMON, 1997). A definição de um padrão e valor "default" são fundamentais no processo de integração, permitindo uma comparação entre os atributos existentes no ambiente operativo, no DW e nos DM.

f) Estabelecimento de regras de conversão para substituir códigos e abreviações: A preocupação em substituir códigos e abreviaturas é tratada por Inmon (INMON, 1997). Neste trabalho, estas regras são estabelecidas durante o desenvolvimento do pré-modelo e posteriormente, na fase de integração, centralizadas no DW. Esta centralização permite um melhor gerenciamento e controle das regras.

g) Criação de atributos derivados: A criação de atributos derivados é realizada durante a elaboração do pré-modelo. Os critérios estabelecidos seguem os definidos por Silverston (SILVERSTON, INMON, GRAZIANO, 1997).

h) Criação de minidimensões: A criação de minidimensões para permitir o rastreamento de alterações, melhor desempenho nas consultas e melhor visualização de grandes dimensões utiliza as técnicas de Kimball (KIMBALL, 1996) (KIMBALL *et al*, 1998).

i) Tratamento da necessidade de rastreamento de atributos de dimensão: As técnicas

empregadas para tratar a necessidade de rastreamento de atualizações dos atributos em uma dimensão são apresentadas por Kimball (KIMBALL, 1997) (KIMBALL *et al*, 1998) e McGuff (MCGUFF, 1998).

j) Tratamento de dimensões com produtos heterogêneos: Aplica a abordagem de Kimball para detectar a necessidade de criação de dimensões específicas (KIMBALL *et al*, 1998).

k) Nomeação de atributos do modelo com termos de negócio: Operação proposta por Inmon (INMON, 1997) e Silverston (SILVERSTON, INMON, GRAZIANO, 1997) para tornar o modelo mais acessível ao usuário final.

6.1.2 Recursos e Técnicas Adaptados para o Trabalho

A seguir são apresentados os recursos e técnicas já existentes e que foram adaptados para serem utilizados nas diretrizes propostas:

a) Desenvolvimento do pré-modelo:

O pré-modelo adotado neste trabalho tem por base o modelo proposto por Silverston (SILVERSTON, INMON, GRAZIANO, 1997) para a modelagem do DW. Em sua abordagem, Silverston propõe um conjunto de transformações no modelo corporativo da empresa para a elaboração do modelo de dados do DW. O modelo dimensional do DM é derivado a partir do modelo de dados do DW. Entretanto, a modelagem dos fatos e dimensões é realizada de acordo com o problema a ser solucionado, uma prática comum na abordagem que emprega técnicas dimensionais.

As transformações empregadas por Silverston foram adaptadas neste trabalho para elaborar o pré-modelo, ao invés de gerar o modelo de DW. A seguir são apresentadas as alterações empregadas na abordagem proposta por Silverston:

- Exclusão de informações desnecessárias: Na proposta original, Silverston exclui do modelo as informações de caráter operativo. No conjunto de diretrizes, esta operação se tornou mais genérica, abrangendo, além das informações de caráter operativo, aquelas que não apresentam interesse para a análise. Uma outra adaptação se refere ao processo de seleção. Na abordagem original, a análise é realizada sobre atributos. No conjunto de diretrizes, a análise é realizada primeiro a nível de entidades e depois para os atributos das entidades restantes. Desta forma, o processo de remoção de informações que não interessam à análise é acelerado.

- Desnormalização de entidades: Além das sugestões estabelecidas por Silverston para identificar as entidades a serem desnormalizadas, foram incluídas ressalvas com o propósito de impedir desnormalizações indevidas. Um exemplo de ressalva é a prioridade de desnormalização para entidades que apresentem relação de dependência de existência.
- Definição de categorias: A definição de categorias foi estendida para permitir a criação de hierarquias segundo critérios necessários ao DM. Dessa forma, a definição de categorias passa a ser um recurso para tratar a esparsidade no momento da análise através de hierarquias definidas para o DM.
- Casamento de entidades: Na abordagem de Silverston, o casamento de entidades é utilizado para a desnormalização de entidades. No conjunto de diretrizes, entretanto, este casamento se relaciona à necessidade de unir entidades semelhantes de diferentes DER. Essa é uma prática necessária para a elaboração do pré-modelo quando existe mais de um DER sendo analisado.

b) Derivação do modelo dimensional

Dentre as abordagens que mencionam a derivação do modelo dimensional a partir do modelo corporativo, estão as propostas de Golfarelli (GOLFARELLI, MAIO, RIZZI, 1998) e de Silverston (SILVERSTON, INMON, GRAZIANO, 1997).

Na abordagem de Golfarelli, um modelo conceitual para o DW é gerado a partir do modelo corporativo. Os modelos lógico e físico são derivados a partir deste modelo conceitual, seguindo os mesmos padrões empregados para a modelagem do ambiente operativo. A derivação do modelo corporativo para o modelo conceitual do DW emprega o processo de poda e enxerto. Entretanto, a árvore gerada em sua abordagem apresenta duas desvantagens:

- baseia-se no modelo corporativo, sem realizar nenhum tratamento para reduzir o conjunto de informações a ser analisado; e
- representa uma árvore de atributos que pode ficar extremamente difícil de tratar, dependendo do modelo em uso.

Na abordagem de Silverston, o processo de derivação do DM é realizado com base nos dados do DW. Contudo, a criação de dimensões e tabelas de fatos permanece a critério do projetista. O tratamento das dimensões e dos fatos é superficialmente

abordado sem a definição de qualquer sistemática a ser seguida.

A proposta de derivação apresentada no capítulo 4 constrói uma árvore de entidades/relacionamento a partir da qual é realizada a extração do esboço de um modelo dimensional. Este processo é aplicado para cada fato básico identificado no pré-modelo. As operações como construção da árvore, poda e enxerto foram desenvolvidas neste trabalho para propiciar uma maior agilidade à derivação do modelo dimensional. Estas operações estão apresentadas na item 6.1.3 – Recursos e Técnicas Desenvolvidos para o Trabalho.

Ao final do processo, é incluída a dimensão tempo, fundamental neste ambiente, estabelecendo o esboço de modelo dimensional. Este modelo deve ser refinado de modo a apresentar todas as características da modelagem dimensional.

c) Relacionamento M:N entre dimensão e fato

Na abordagem de Kimball (KIMBALL *et al*, 1998), após a definição do modelo dimensional é necessária uma análise de cada dimensão existente, verificando-se a cardinalidade de seu relacionamento com a tabela de fatos. O propósito desta análise é identificar as dimensões que apresentem um relacionamento M:N com a tabela de fatos. Para estas dimensões serão criadas "dimensões ponte", conforme apresentado na seção 3.4.1 -Tratamento de Dimensões e Fatos com Cardinalidade M:N. Esta cardinalidade pode não estar clara no negócio, podendo exigir consultas a ABD e usuários finais.

Na proposta deste trabalho, a derivação a partir do modelo corporativo pode gerar um modelo dimensional que, eventualmente, apresente uma dimensão com a cardinalidade M:N com a tabela de fatos. Portanto, localizar a cardinalidade M:N é um processo fácil. Esta dimensão representa a própria "dimensão ponte" à qual Kimball se refere (KIMBALL *et al*, 1998). Ao localizar esta dimensão, o projetista deve criar uma nova dimensão, mais simples, que permita o desenvolvimento de consultas a partir dela para a tabela de fatos.

d) Criação de sub-fatos ou fatos específicos

A necessidade de fatos específicos é uma característica dos ambientes que operam com diversos produtos e serviços. Na abordagem de Kimball, a modelagem dimensional estabelece uma tabela de fatos gerais, permitindo uma visão genérica. A

partir da tabela de fatos gerais serão definidas as tabelas de fatos específicos, permitindo consultas direcionadas.

Na proposta deste trabalho, as tabelas de fatos específicos e a tabela de fatos gerais podem ser definidas no momento em que o projetista estabelecer os fatos básicos. A preocupação na fase do refinamento das tabelas de fatos consiste em garantir que as chaves dos fatos específicos sejam incluídas na tabela de fatos gerais, proporcionando consultas mais rápidas.

6.1.3 Recursos e Técnicas Desenvolvidos para o Trabalho

A seguir serão analisadas algumas técnicas empregadas na proposta para a modelagem do ADW. Estas técnicas foram desenvolvidas com o propósito de refinar o modelo dimensional gerado a partir de modelos de dados existentes.

a) Levantamento de fatos básicos e visões dimensionais: esta operação é realizada a partir do momento em que o pré-modelo é gerado. O propósito é a identificação dos fatos básicos existentes no pré-modelo e o levantamento das possíveis visões dimensionais que possam ser extraídas a partir deles. Através das visões dimensionais o usuário final terá uma idéia mais clara dos resultados que poderá obter a partir do DM. Os fatos básicos, por sua vez, permitem ao projetista identificar os possíveis modelos dimensionais que comporão o DM.

b) Construção da árvore de entidades/relacionamentos: Na fase de elaboração do modelo dimensional é construída uma árvore de entidades/relacionamento, sobre a qual se realiza um trabalho de limpeza com base nas cardinalidades apresentadas. Estas cardinalidades refletem as regras existentes no ambiente operativo. Através da limpeza, uma árvore simples é obtida. É sobre esta árvore que o projetista deve realizar uma análise, aplicando as técnicas de poda e enxerto de acordo com a relevância da informação dos nós existentes.

c) Aplicação de operações de poda e enxerto: estas operações foram definidas com o propósito de eliminar os níveis desnecessários de detalhamento. A árvore resultante do processo de limpeza proporciona ao projetista uma visão das possíveis dimensões a serem criadas. As operações de poda e enxerto permitem ao projetista definir o

nível de detalhe necessário para o modelo. Estas operações são aplicadas à árvore para a elaboração do esboço de modelo dimensional.

d) Remoção dos atributos de dimensão não relacionados ao modelo: Esta operação foi desenvolvida para o refinamento das dimensões do esboço do modelo gerado. Ela consiste em analisar os atributos das dimensões criadas em um modelo, eliminando os atributos que não apresentem interesse. O processo de remoção dos atributos pode gerar uma dimensão degenerada na tabela de fatos.

e) Estabelecimento de hierarquias: Esta operação consiste em identificar e registrar as hierarquias existentes nas dimensões. Este registro facilita a criação de tabelas sumarizadas no nível físico e apresenta para o usuário final os níveis de visualização possível para as consultas.

f) Classificação dos atributos da tabela de fatos: A classificação dos atributos de uma tabela de fatos permite identificar a existência de atributos que pertençam a uma dimensão. A existência destes atributos é decorrente do processo de transformação de uma entidade em uma tabela de fatos. Ao detectar a existência de atributos descritivos e identificadores, o projetista deve avaliar a sua importância para o modelo. Caso os atributos não interessem à análise, eles simplesmente serão removidos. Caso contrário, será construída uma dimensão com uma nova chave para atender ao modelo.

g) Integração do modelo do DM ao DW: A integração do modelo do DM ao DW pode ser realizada empregando o modelo dimensional do DM, quando o DW emprega a modelagem dimensional, e empregando o pré-modelo, quando o DW emprega a modelagem relacional tratada. No conjunto de diretrizes são discutidas as preocupações que o projetista deve ter no momento de efetuar as integrações. Este tratamento é essencial para evitar a duplicidade de informações no DW que possa gerar inconsistência, promovendo a perda da confiabilidade do ambiente.

6.2 Análise das Diretrizes Propostas

De uma forma geral, o conjunto de diretrizes propostas neste trabalho apresenta as seguintes características:

- Permite a derivação de modelos dimensionais a partir de um modelo corporativo ou dos DER dos sistemas existentes no ambiente operativo e fontes externas;
- Utiliza um pré-modelo que concentra as informações de interesse, reduzindo o escopo a ser trabalhado no processo de derivação;
- Emprega uma abordagem em árvore para realizar a derivação do modelo dimensional a partir do pré-modelo;
- Aplica técnicas de poda e enxerto para definir as dimensões a partir de entidades/relacionamentos. A aplicação do processo de poda e enxerto garante um processo menos intuitivo para a determinação das dimensões;
- Aplica técnicas de modelagem dimensional para refinar as dimensões e tabelas de fatos; e
- Realiza a integração do modelo de dados do DM ao DW.

O conjunto de diretrizes apresentado neste trabalho tem o propósito de desenvolver um ADW empregando a abordagem "bottom-up" para a sua modelagem. Dessa forma, primeiramente é modelado um DM, com enfoque em uma área de interesse. O modelo dimensional do DM é posteriormente integrado ao modelo de dados do DW. A integração é realizada de forma criteriosa, centralizando as regras e expressões no DW. Desse modo, é possível evitar o desenvolvimento de DM independentes e garantir que os dados existentes nos DM sejam gerados a partir de um mesmo conjunto de regras.

Por ser desenvolvido por DM, o processo de implementação tende a ser rápido, permitindo ao usuário final validar e julgar o valor do ADW.

Para realizar a modelagem do DM as diretrizes propostas empregam os modelos de dados dos sistemas de interesse. Quando ocorrer das informações de interesse de um DM terem uma origem externa, os dados relacionados a esta fonte deverão ser modelados através de um DER. Este DER será tratado juntamente com os DER dos sistemas existentes no ambiente operativo para a geração do pré-modelo. A construção

do pré-modelo é fundamental, porque facilita a conversão de um modelo orientado a processo (modelo corporativo), para um modelo orientado ao negócio (modelo dimensional). O pré-modelo apresenta o conjunto de informações que realmente interessam à análise. A partir de um pré-modelo, podem ser elaborados mais de um modelo dimensional. Estes modelos comporão o modelo dimensional do DM.

Para realizar a elaboração do modelo dimensional, o conjunto de diretrizes emprega a construção de uma árvore, cuja raiz é representada pela entidade selecionada como tabela de fato. As demais entidades/relacionamentos do pré-modelo constituem os nós das subarvores. Um processo de limpeza é aplicado na árvore gerada, com base nas cardinalidades entre os nós. O projetista realiza, então, as operações de poda e enxerto, definindo as entidades que serão transformadas em dimensões. É obtido, então, o primeiro esboço de modelo dimensional. O esboço do modelo dimensional é tratado através de técnicas de modelagem, de forma a atender às características de um modelo dimensional. Estas técnicas englobam o refinamento de dimensões e o refinamento da tabela de fatos.

O modelo dimensional final não é simplesmente um modelo criado a partir da necessidades/requisitos do usuário final, mas um modelo que tem sua base nas informações disponíveis para a empresa. A sua derivação a partir dos modelos de dados de sistemas existentes e de modelos de dados relacionados a fontes externas permite a visualização de informações de interesse que possam não ter sido observadas pelo usuário final, até mesmo por desconhecimento. Além disso, a derivação permite uma definição mais ágil das dimensões e reduz a subjetividade envolvida no processo.

Após o desenvolvimento do DM, é realizada a integração do modelo de dados ao DW. Ao contrário das propostas de transformações do modelo corporativo para um DW relacional tratado, as diretrizes propostas permitem ao projetista liberdade de escolha do modelo a ser empregado no DW.

Pela sua característica incremental, gerando o DM e integrando-o ao DW, as diretrizes propostas permitem que o usuário possa avaliar o ADW a cada novo DM gerado. Esta avaliação permite uma baixa taxa de risco, pois o usuário pode julgar as vantagens e desvantagens do ADW a cada novo DM. Além disso, é possível justificar o investimento realizado e atrair a atenção do usuário final, apresentando resultados.

O desenvolvimento incremental permite que a equipe de projetistas ganhe, de

forma gradativa, mais experiência no desenvolvimento do ADW.

O emprego dos modelos de dados existentes no ambiente operativo e de modelos de dados referentes às fontes externas, para a construção do DM, facilita o mapeamento dos dados no momento da elaboração do DM. Este mapeamento é estendido ao DW através da fase de integração. No momento da integração existe a preocupação em garantir o mapeamento entre o modelo de dados do DW e o modelo de dados do DM. Dessa forma, ao final da integração, é possível um mapeamento entre o ambiente operativo, o DW e o DM. As diretrizes propostas exigem um controle rigoroso no momento da integração do modelo de dados do DM ao modelo de dados do DW, para evitar duplicidade de informações. Este controle permite centralizar as regras e a gerência do mapeamento de dados no nível do DW.

Como é possível observar, o conjunto de diretrizes permite uma abordagem mais sistemática para o processo de modelagem de um ambiente de DW. Entretanto, as diretrizes consistem de uma proposta, tendo sido aplicadas apenas em estudos de casos.

Como resultado da análise realizada, obtém-se as seguintes vantagens e desvantagens:

a) Vantagens:

- Modelagem derivada de modelos existentes no ambiente operativo;
- Modelagem com enfoque em uma área de interesse (por DM);
- Facilidade em mapear a origem das informações. O processo de derivação a partir de modelos existentes auxilia o mapeamento dos dados;
- Facilidade para identificar os fatos básicos. O pré-modelo apresenta um conjunto restrito de informações relacionados ao assunto de interesse. Desta forma, a identificação dos fatos básicos se torna fácil e é realizada sobre informações disponíveis na empresa.
- Baixa taxa de risco;
- Implementação rápida (por DM);
- Evita o desenvolvimento de DM independentes (integração);
- Definição de dimensões e tabela de fatos a partir das informações existentes no ambiente operativo;
- Controle e centralização de regras;

- Controle do crescimento do DW (por DM); e
- Permite empregar tanto o modelo dimensional como o relacional para o DW.

b) Desvantagens:

- Necessidade do desenvolvimento de DER para informações de fontes externas. As fontes externas são tratadas como sistemas existentes na empresa;
- Exige o modelo corporativo ou DER dos sistemas de interesse;
- Tempo gasto na elaboração do pré-modelo; e
- Exige um controle rigoroso no momento da integração do modelo de dados do DM ao modelo de dados do DW, para evitar duplicidade de informações.

CAPÍTULO 7

CONCLUSÃO E TRABALHOS FUTUROS

7.1 Considerações Gerais

Como apresentado neste estudo, diversos são os fatores que vêm influenciando a absorção desta nova tecnologia pelas empresas. Um dos pontos considerados fundamentais é a busca da vantagem competitiva das empresas. Esta vantagem, atualmente, reside na capacidade de tomar decisões estratégicas e táticas de forma ágil, com base nas informações disponíveis para os tomadores de decisões das empresas.

O fato de ser reconhecidamente uma tecnologia nova não afasta os investimentos, pelo contrário, representa, atualmente, um campo crescente de desenvolvimento para pesquisadores e empresas dispostas a investir.

Entretanto, a falta de maturidade gera preocupações quanto ao processo de desenvolvimento do ADW. Da mesma forma como se notificam os sucessos de desenvolvimentos deste ambiente, também se notificam os fracassos. A busca por metodologias para o desenvolvimento deste ambiente é contínua. Discussões sobre modelagem e escolha de abordagens são comentadas e apresentadas em diversos artigos e congressos.

Este estudo se preocupou em propor um conjunto de diretrizes que permitissem uma modelagem para o ADW, de uma forma até o momento não abordada pela literatura.

7.2 Contribuições

As diretrizes propostas neste estudo realizam a modelagem dos dados no ADW de forma incremental, seguindo a abordagem recomendada para o desenvolvimento deste ambiente. A modelagem proposta reforça o emprego dos modelos existentes no ambiente operativo para a modelagem do ADW. Esta linha de ação vem ganhando cada vez mais adeptos (MEYER, 1998) (SILVERSTON, 1997) (INMON, 1998).

Conforme apresentado, através do emprego dos modelos de dados existentes no ambiente operativo é possível realizar um trabalho de transformação, reduzindo o

caráter operativo dos modelos. A partir do modelo tratado (pré-modelo) é realizada, então, a derivação de modelos dimensionais.

De acordo com as análises apresentadas no capítulo 6 é possível observar que o emprego do conceito de árvore e cardinalidade, aliada às técnicas de poda e enxerto agilizam o desenvolvimento de modelos dimensionais. Por sua origem, estes esboços refletem muito mais o negócio da empresa, reduzindo significativamente a subjetividade normalmente empregada pelos projetistas na definição de modelos dimensionais.

O tratamento de cardinalidade, aplicado à árvore gerada a partir do pré-modelo, resulta em árvores orientadas ao fato em questão. Desta forma, a árvore final apresenta o escopo de interesse para o projetista. Os modelos dimensionais derivados apresentam uma base sólida, ao invés de dependerem apenas das idéias e requisitos definidos pelo usuário final, como propõem algumas técnicas de modelagem dimensional existentes (KIMBALL, 1996) (KIMBALL *et al*, 1998) (MCGUFF, 1998).

A integração dos modelos dos DM ao modelo do DW permite a centralização de regras e a gerência dos DM criados, evitando o surgimento de DM independentes. A centralização das regras garante a confiabilidade dos dados a serem extraídos a partir do DW.

7.3 Sugestões e Trabalhos Futuros

Para dar continuidade a este trabalho, são sugeridos os seguintes estudos:

a) Ferramenta CASE:

O desenvolvimento de uma ferramenta CASE para atender às diretrizes apresentadas permitiria uma maior agilidade ao processo de modelagem.

Uma ferramenta CASE facilitaria o trabalho de integração dos DER dos modelos existentes. Além disso, poderia agilizar a criação dos esboços dos modelos dimensionais, gerando as árvores (etapa C.II.b) de modo automático a partir de uma entidade selecionada como fato básico pelo usuário. Através das propriedades das ferramentas CASE o mapeamento das origens dos atributos seria facilitado.

b) Gerenciador de METADADOS:

Ao longo deste trabalho foi apresentada a importância dos metadados para o ADW. O desenvolvimento de um gerenciador de METADADOS permitiria um melhor

controle das mudanças estruturais, comuns neste tipo de ambiente. Este gerenciador poderia apresentar facilidades, como por exemplo, relatório de DM afetados pela alteração de uma dimensão.

c) Ordenação para Refinamento de Dimensões e Tabela de Fatos:

Neste trabalho foram apresentados os passos a serem aplicados para realizar o refinamento das dimensões e tabelas de fatos sem, entretanto, uma ordem exata.

A aplicação do conjunto de diretrizes em novos estudos de casos permitirá avaliar seqüências de passos apropriados para os diferentes tipos de fatos e de dimensões.

d) Modelo Físico:

O tema deste trabalho teve como enfoque a modelagem a nível conceitual e lógico do ADW. Um novo trabalho seria o desenvolvimento de técnicas para, a partir do modelo lógico gerado para o DW e DM, derivar o modelo físico.

Estas técnicas poderiam, dentre outras:

- sugerir possíveis agregados com base nas hierarquias definidas;
- avaliar o emprego de agregados idênticos em vários DM; e
- sugerir a criação no DW dos agregados empregados por mais de um DM.

e) Extratores Automáticos:

Este trabalho se preocupou em garantir o mapeamento dos dados entre ambiente operativo, DW e os DM. Este mapeamento poderá ser empregado para o desenvolvimento de uma ferramenta que gere, a partir dos modelos de dados dos DM e do DW, programas responsáveis pela extração dos dados. Estes programas poderiam ser no nível DW→ DM, permitindo a atualização dos DM a partir do DW e no nível ambiente operativo → DW preparando a carga dos dados.

ANEXOS

ANEXO 1

HISTÓRICO DA EVOLUÇÃO DO AMBIENTE DE APOIO A DECISÃO

O AAD é constituído pelos Sistemas de Suporte a Decisão (SSD) e de Informações Estratégicas (SIE). Estes sistemas basicamente promovem o apoio a decisão, empregando ferramentas e métodos de análise sobre os dados coletados na empresa. São desenvolvidos com o propósito de dar suporte aos tomadores de decisão, permitindo uma visualização de vários aspectos do problema a ser analisado (Adelman, 1992).

Diversos fatores levaram à necessidade da evolução do AAD. Dentre eles, merecem destaque:

- A proliferação de informações dentro da empresa, decorrente das técnicas de reengenharia e downsizing;
- A necessidade das empresas consolidarem e integrarem as informações distribuídas por diversos sistemas com características de hardware e software completamente diferentes entre si;
- A Internet e Intranet, movimentando o mercado de informações com dados, até então considerados de difícil acesso;
- Disponibilização de informações de fontes externas;
- Maior divulgação de bancos de dados multidimensionais (BDM). Os BDM têm sido empregados como bancos de dados proprietários em ferramentas que permitem, aos usuários finais, um acesso às informações mais intuitivo, por meio da visão multidimensional;
- Investimento por parte dos desenvolvedores em ferramentas mais poderosas baseadas em redes neurais, algoritmos genéticos e lógica nebulosa, dentre outros, permitindo novas abordagens de análise, como projeções, avaliações de parâmetros e padrões de comportamento; e
- Necessidade de se desenvolver consultas *ad hoc*, não limitando o usuário final apenas a consultas pré-definidas. As consultas *ad hoc* exigem o desenvolvimento de bases de dados intuitivas e de fácil acesso.

A seguir são apresentados os estágios da evolução do AAD.

ESTÁGIO 1: SURGIMENTO DO AMBIENTE DE APOIO A DECISÃO.

A evolução da informática na década de 80, permitiu a informatização das principais atividades das empresas. Sistemas como "Controle de Estoque" e "Controle de Pedido", foram desenvolvidos para auxiliar a empresa em suas operações rotineiras. Por serem destinados a tratar as operações do dia a dia da empresa, esses sistemas são denominados sistemas operativos (Orr, 1997). Os sistemas operativos, também conhecidos como sistemas OLTP (On-line Transaction Processing) ou sistemas legados, normalmente, são empregados como fontes de dados para os SSD.

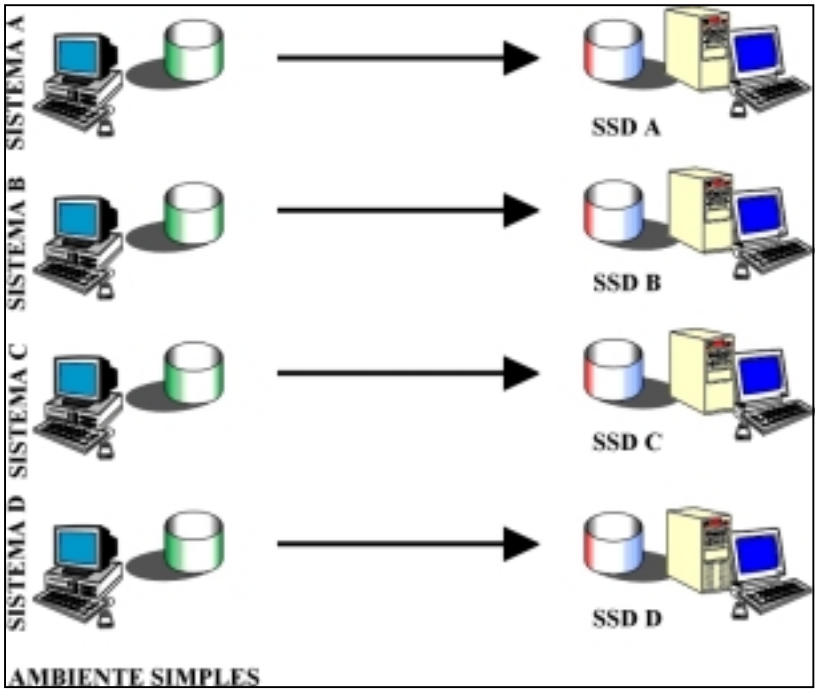

Figura A.1.1 – Ambiente Simples de Apoio a Decisão

A proliferação desses sistemas permitiu melhorar o controle dos processos relacionados à empresa e promoveu um grande acumulo de informações. Essas informações entretanto, eram armazenadas por tempo limitado e não se apresentavam em uma estrutura que facilitasse consultas informativas e analíticas. Com o objetivo de atender as necessidades da empresa quanto a essas consultas, surgem os SSD, também conhecidos como Sistemas de Apoio a Decisão (SAD), permitindo a visualização de vários aspectos do problema a ser decidido (Adelman,

1992).

O crescimento dos SSD nas empresas, para atender a demanda de análises estratégicas, estabeleceu um ambiente, que ficou conhecido como Ambiente de Apoio a Decisão (AAD).

ESTÁGIO 2: A PROLIFERAÇÃO DOS SISTEMAS DE SUPORTE A DECISÃO E O AUMENTO DA COMPLEXIDADE DO AMBIENTE DE APOIO A DECISÃO.

A análise dos negócios da empresa, pelos usuários dos SSD, normalmente exige a combinação de dados de mais de um sistema operativo, provocando um cruzamento de extrações de dados muitas vezes de difícil gerenciamento.

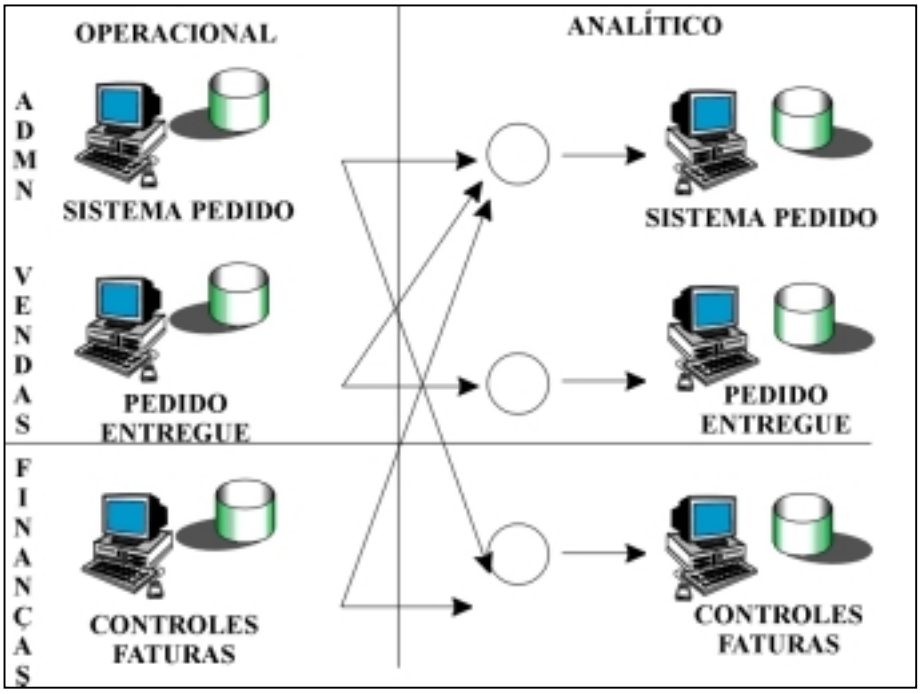

Figura A.1.1 – Complexidade no AAD

A proliferação dos SSD fez crescer a complexidade do AAD, dificultando ainda mais o controle e a gerência das informações acessadas nos sistemas operativos (Devlin,1997). Não era incomum a existência de dois relatórios semelhantes com resultados diferentes, por acessarem diferentes fontes (Inmon, 1997). Dentre os fatores responsáveis pelos problemas nesse ambiente merecem destaque (Inmon, Devlin, 1997):

- Distribuição de dados por diversos bancos de dados que alimentam os SSD;
- Dificuldade na determinação de fontes consistentes de informação. Essa dificuldade está

relacionada a distribuição das informações pelos diversos sistemas da empresa, onde se pode encontrar informações com o mesmo nome, porém com significados diferentes; informações com nomes diferentes, porém com mesmo significado e as mesmas informações porém com periodicidade de atualizações diferentes;

- Dificuldade em manter a coerência dos dados externos, uma vez que cada departamento poderia entrar com seus dados, gerando informações conflitantes, como por exemplo, a utilização de indicadores financeiros extraídos de fontes distintas;
- Necessidade de armazenar dados históricos. Os dados históricos são freqüentemente empregados nos SSD para análises estratégicas que visam, por exemplo, projeções e estudos de comportamento;
- Demora na recuperação dos dados. Em alguns casos, os dados obtidos nem ao menos atendem às expectativas dos usuários;
- Duplicidade de código de extração, para evitar alterar códigos existentes. Como cada novo SSD se encarrega de criar sua massa de dados, é necessário avaliar a existência de extratores que trabalham com as informações de interesse ou então desenvolver um novo. Nos AAD é comum o desenvolvimento de uma nova aplicação de extração, mesmo existindo uma similar, para evitar alterações que possam comprometer os sistemas em funcionamento;
- Impacto das mudanças nas fontes de dados. Uma mudança em uma fonte de dados implica em uma série de alterações nos programas de extração, e cada programa necessita de análise e alterações independentes;
- Dificuldade de manutenção da documentação de cópias dos programas. Essa falta de atualização ou mesmo falta de uma documentação pode provocar sérias conseqüências quanto à consistência e à integridade das informações após uma mudança em uma fonte;
- Crescimento do volume de informações duplicadas de forma quase incontrolável, uma vez que uma mesma informação pode ser utilizada por diversos SSD, e consequentemente estar carregada em cada uma de suas bases de dados; e
- Necessidade de uma administração de dados complexa.

Analisando os fatores apresentados, é possível observar que o foco dos problemas está na manutenção e no gerenciamento dos dados, para garantir a confiabilidade e a consistência das informações processadas.

ESTÁGIO 3: A TECNOLOGIA DE DATA WAREHOUSING.

A proliferação dos SSD resultou em dificuldades no controle e gerenciamento de dados. Nesse contexto surge a tecnologia de Data Warehousing. Essa tecnologia é uma combinação de tecnologias destinadas ao suporte a decisão, com o propósito de habilitar o usuário final a tomar decisões melhores e mais rápidas (Golfarelli, 1998). Essa tecnologia se preocupou em:

- Introduzir o DW, garantindo a integração e a consistência dos dados distribuídos por diversas fontes;
- Padronizar e controlar as fontes do ambiente, sejam elas internas ou externas. Com uma única base de dados, se torna mais fácil o gerenciamento e controle dos metadados; e
- Disponibilizar um novo conjunto de ferramentas voltadas ao ambiente de suporte a decisão, permitindo consultas e visões multidimensionais.

O passo fundamental desse estágio foi a introdução do DW, que passou a funcionar como fonte de dados para os SSD. As preocupações sobre onde buscar as informações e quando carregá-las passam ao DW, permitindo aos SSD se preocuparem apenas com questões sobre o processamento de consultas .

ESTÁGIO 4: SURGEM OS DATA MARTS (DM)

O surgimento dos DM está associado a dois aspectos importantes: desempenho e desenvolvimento.

Com o crescimento do DW, tornou-se interessante, para os departamentos e divisões das empresas, manter as informações de seu interesse, em uma base de dados própria. Essa base representaria um sub-conjunto de informações do DW. Dessa forma, o acesso à informação não seria compartilhado com outros departamentos, propiciando um melhor desempenho nas consultas. O outro ponto interessante, nessa separação, está no fato do DM não conter, necessariamente, a mesma granularidade de dados apresentada no DW, podendo, entretanto, fazer consultas mais específicas a ele.

No aspecto desenvolvimento, os DM surgem como uma resposta rápida ao problema de se projetar um DW para uma empresa. Em grandes empresas, com diversos sistemas, o desenvolvimento de um DW pode levar anos. O mercado de informática, entretanto, está sempre

em busca de soluções rápidas e eficazes. Os DM representam uma solução para conciliar esses dois pontos, pois permitem o desenvolvimento do DW de forma iterativa, através de pequenos módulos orientados a áreas específicas, apresentando resultados rápidos (Firestone, 1997).

Como pode ser observado, uma abordagem não invalida a outra. Na verdade, o que se observa é o desenvolvimento incremental do DW, através de pequenos módulos que representam os DM. Os DM, por sua vez, otimizam a consulta por apresentarem dados restritos às suas necessidades.

ANEXO 2

PECULIARIDADES DO MODELO FÍSICO DO AMBIENTE DE DATA WAREHOUSE

Como mencionado, este trabalho não tem o propósito de atingir o modelo físico, restringindo-se a modelagem conceitual e lógica do ADW. Entretanto, a seguir são apresentadas algumas considerações sobre a modelagem física de interesse aos projetistas.

A2.1 - REAVALIAR DIMENSÕES POPULOSAS

Normalmente, recomenda-se o emprego de uma única tabela para preservar o desempenho da navegação e a simplicidade de interface com o usuário. Entretanto, as dimensões que apresentam muitas tuplas precisam de um tratamento especial. Dentre as soluções propostas merece destaque a aplicação de uma indexação especial, como por exemplo a utilização de índices "bitmap". O emprego desse tipo de indexação vem sendo considerado um sucesso.

Nem todos os atributos de uma dimensão precisam ser indexados. Uma análise deve ser realizada sobre a dimensão, com o propósito de selecionar os atributos amplamente utilizados em consultas, e então indexá-los.

A2.2 – EMPREGO DE MINIDIMENSÕES PARA MELHORAR PERFORMANCE

Atributos muito empregados em cruzamentos, podem ser separados em um ou mais grupos, através da criação de minidimensões, indexadas separadamente com o propósito de melhorar a performance dos movimentos de "drill up/down" . Como exemplo, os atributos "estado civil", "idade", "sexo", "renda", considerados atributos DEMOGRÁFICOS de uma dimensão Cliente.

Uma minidimensão é, normalmente, composta por 5 (cinco) ou 6 (seis) campos. Apenas as informações existentes são armazenadas, não sendo necessário armazenar todas as combinações possíveis. Os campos com valores contínuos, que comporão a minidimensão, devem ser transformados em intervalos, reduzindo o total de

combinações. Para preservar a alta performance é recomendável que as minidimensões apresentem no máximo 100.000 combinações (Kimball, 1996).

A referência à minidimensão pode ser direta a partir da tabela de fatos, e/ou através da dimensão originária, que representa a origem dos atributos. Quando os atributos "idade, sexo, renda e estado civil" são excluídos da dimensão originária, é importante criar uma referência a dimensão geográfica, permitindo que a partir de um cliente, seja possível obter suas características demográficas.

A criação de uma minidimensão não exige que os atributos tenham algum relacionamento lógico, sendo possível misturar atributos geográficos com atributos demográficos em uma única minidimensão (Kimball, 1996). Entretanto, a mistura de atributos, sem um critério, pode ocasionar problemas para futuros projetistas e desenvolvedores.

A2.3 - CRIAÇÃO E GERÊNCIA DE AGREGADOS

O emprego de agregações representa uma das mais poderosas ferramentas disponíveis no ADW, em termos de controle de desempenho. Uma agregação nada mais é do que uma tabela de fatos que sumariza outros fatos de nível mais baixo.

As questões de performance são mais críticas nas aplicações de DW. O fato de que um DW é um sistema de suporte a decisões de nível gerencial, faz surgir a necessidade de um DBA mais atento à satisfação do usuário final. Aspectos como por exemplo o custo da hora de trabalho de um executivo e o custo de armazenamento e carga dos dados, devem ser analisados e monitorados constantemente, com o objetivo de reavaliar a necessidade de agregações e índices.

a) Implementação das Agregações

A agregação de informações pode ser implementada através de duas técnicas:

- criação de novas tabelas de fatos agregadas; e
- a criação de novos campos de indicação que indicam o nível de agregação nas tabelas dimensões.

No primeiro caso, exige-se a criação de chaves artificiais em cada uma das dimensões que está sendo agregada. No caso da criação de campos de indicação do nível de agregação, os fatos agregados serão incluídos na própria tabela de fatos original. Esta segunda opção não é a mais recomendada por gerar problemas de dupla contagem.

b) Controle das Agregações

As agregações podem causar uma explosão do tamanho do DW, no caso de processos comerciais. Uma das formas eficientes de controlar esta explosão é garantir que cada agregação sumariza mais de 10 ou 20 itens.

c) Balancear os Índices das Agregações:

A combinação do uso Índices em agregados é bastante interessante, pois tira da tabela de fatos o peso de carregar todos os índices, distribuídos por entre as tabelas de fatos agregadas.

ANEXO 3

ÁRVORES

Árvores são estruturas de dados que caracterizam uma relação entre os dados que a compõem. Cada um desses dados é denominado "nó". Entre os nós de uma árvore é apresentada a relação de hierarquia ou composição. Dessa forma, um conjunto de dados é hierarquicamente subordinado a outro.

Esquematicamente, uma árvore pode ser representada como apresentado na figura D.1.

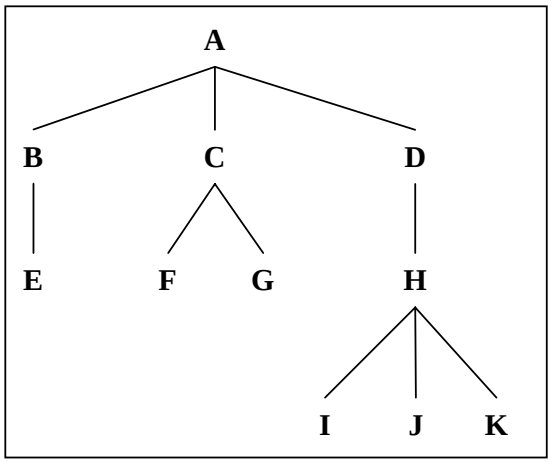

Figura A.3.1 Exemplo de árvore

TERMINOLOGIA

- Cada nó de uma árvore é raiz de uma sub-árvore.
- GRAU: o número de sub-árvores de um nó é o grau daquele nó.
- FOLHA: um nó com grau zero é denominado folha ou grau terminal.
- NÍVEL DE UM NÓ: é o comprimento do caminho que vai da raiz até este nó. Representa o número de linhas do caminho entre a raiz e o nó.
- ALTURA: representa o nível mais alto da árvore.
- ÁRVORE ORDENADA: uma árvore é considerada ordenada quando a ordem das sub-árvores é significativa. O exemplo apresentado na figura D.2 mostra duas árvores diferentes.
- ÁRVORE ORIENTADA: uma árvore é orientada quando apenas a orientação dos nós é importante, sendo as árvores da figura D.2

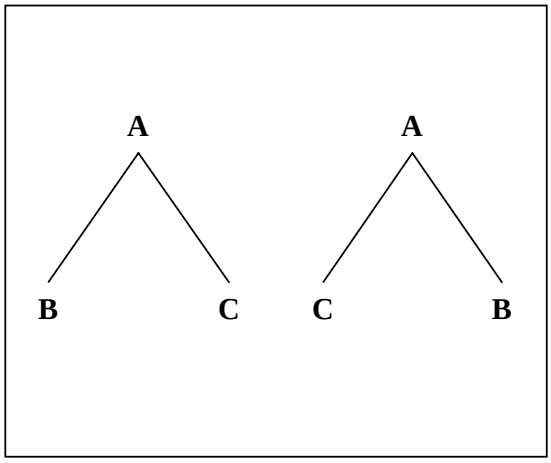

Figura A.3.2 Árvores Ordenadas

orientadas, elas seriam iguais.

APLICAÇÕES

A estrutura de árvore é empregada nos casos onde os dados ou objetos a serem empregados possuem relações hierárquicas entre si. A figura D.3 apresenta um exemplo de representação hierárquica em uma Universidade.

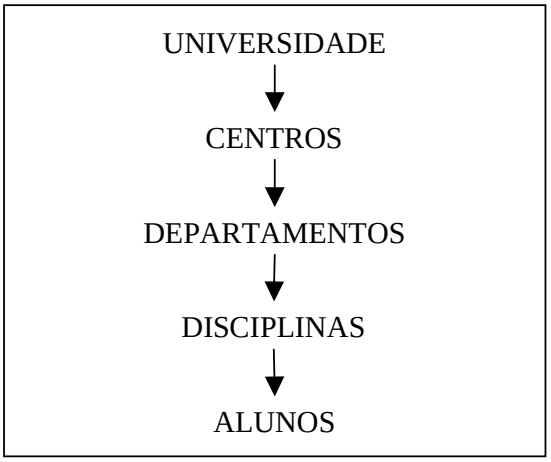

Figura A.3.4 Exemplo de relação hierárquica

ANEXO 4

APOIO AO ESTUDO DE CASO DA UNIVERSIDADE

A4.1) Relação de Atributos Excluídos no DER do SCG:

ENTIDADE: *ALUNO*

pai_reg	**- nome do pai do aluno**
mae_reg	**- nome da mãe do aluno**
rua_reg	**- rua do endereço do aluno**
cep_reg	**- c.e.p. do endereço do aluno**
fone_reg	**- telefone**
numero_ident	**- número da identidade**
expedidor_reg	**- órgão expedidor da identidade**
estado_ident_reg	**- sigla do estado do órgão expedidor da identidade**
num_tit_novo_reg	**- número do título eleitoral**
dv_tit_novo_reg	**- dígito do título eleitoral**
num_cm_reg	**- número do certificado militar ou branco**
serie_cm_reg	**- série do certificado militar ou branco**
estado_cm_reg	**- estado de emissão do certificado militar ou branco**
num_disp_reg	**- número do certificado de dispença de incorporação**
orgao_disp_reg	**- órgão emissor do certificado de dispença de incorporação**
ano_r_reg	**- ano da última rematrícula**
periodo_r_reg	**- período da última rematrícula**
ano_t_reg	**- ano do último trancamento**
periodo_t_reg	**- período do último trancamento**
ano_c_reg	**- ano do cancelamento**
periodo_c_reg	**- período do cancelamento**
num_hist_reg	**- número de períodos do histórico do aluno (não contam os períodos zero)**

curso_pret_reg	**- código da opção (curso/turno/semestre) conquistada no vestibular**
dia_d_reg	**- dia da emissão do diploma**
mes_d_reg	**- mês da emissão do diploma**
ano_d_reg	**- ano da emissão do diploma**
cred_obt_a_reg	**- créditos obtidos acumulados até a última atualização**
num_pag_hist_reg	**- número de páginas a serem impressas no currículo relativas ao período anterior**
cont_tes_reg	**- controle de defesa de tese (" " ou "X")**
ver_curric_reg	**- versão do currículo (ano/período)**
orientador_reg	**- professor orientador do RCS**
dv_orientador	**- digito verificador do professor orientador do RCS**
dia_u_reg	**- dia da última atualização**
mes_u_reg	**- mês da última atualização**
ano_u_reg	**- ano da última atualização**
ativo_futuro_reg	**- futuro valor a ser atribuido a "ativo_reg"**
ano_futuro_reg	**- ano da futura alteração de "ativo_reg"**
per_futuro_reg	**- período da futura alteração de "ativo_reg"**
data_futuro_reg	**- data da futura alteração de "ativo_reg"**

ENTIDADE : *HISTORICO*

byte_calc_cra_his	**- controle para calcular o coeficiente de rendimento acumulado**
ocorr_msg_his	**- quantidade de mensagens neste histórico (1..10)**
ocorr_dis_his	**- quantidade de disciplinas neste histórico (1..90)**
dia_u_his	**- dia da última atualização**
mes_u_his	**- mês da última atualização**
ano_u_his	**- ano da última atualização**

ENTIDADE : *HISTORICO_DISCIPLINA*

mes_atu_his	**- mês da última atualização ("A" = janeiro, até "L" = dezembro)**

ENTIDADE : *RCS*

ano_inicio_rcs	**- ano de início do r.c.s.**
per_inicio_rcs	**- período de início do r.c.s.**
local_rcs	**- local de realização do r.c.s.**
descricao_rcs	**- descrição**
dia_ult_a	**- dia da última alteração**
mes_ult_a	**- mês da última alteração**
ano_ult_a	**- dia da última alteração**

ENTIDADE: *TURMA*

estado_turma_dtr	**- "A" se ativa, "D" se desativada e "P" se pendente**

ENTIDADE: *VERSAO_DISCIPLINA*

estado_dis	**- "A" se ativa, "D" se desativada**
carga_teorica_dis	**- carga de aulas teóricas por semana em horas**
carga_pratica_dis	**- carga de aulas práticas por semana em horas**
cht_teorica_dis	**- carga de aulas teóricas no período todo**
cht_pratica_dis	**- carga de aulas práticas no período todo**
byte_regant	**- "X" se a disciplina tem versão anterior**
campos_preenc_dis	**- número de campo da disciplina preenchida**
confere_grau_cred	**- "X" se a disciplina confere grau**
duracao_maxima	**- quantos períodos no máximo a disciplina pode durar (só RCS) ou branco**

A4.2) Regras de Categorias Apresentadas no DER do SCG.

IDADE:

Cálculo da idade a partir dos atributos *DIA_NASC, MES_NASC* e *ANO_NASC*.

FAIXA_PONTOS_VEST:

PONTUAÇÃO INFERIOR A 5000 : *PONTOS_VEST*< 5000;

PONTUAÇÃO ENTRE 5000 e 7000: *PONTOS_VEST* >= 5000 e < 7000;

PONTUAÇÃO ENTRE 7000 e 8000: *PONTOS_VEST* >= 7000 e < 8000; e

PONTUAÇÃO ACIMA 8000 : *PONTOS_VEST* >= 8000.

PERIODO_ANO_GRAD:

Cálculo do período e ano a partir dos atributos *MÊS_GRAD* e *ANO_GRAD.*

Se *MES_G_REG* >= 7 então

PERIODO_ANO_GRAD = 2PERIODO + *ANO_GRAD*

Senão *PERIODO_ANO_GRAD* = 1PERIODO + *ANO_GRAD*

A4.3) Regras de Definição de Artefatos Apresentados no DER dO SCG.

FREQUENTA_ALOMENTO:

Se *Aluno.NUM_ALOJ_REG* > 0 **Sim**

senão **Não**

RECEBE_AJUDA:

Se existe *Ajuda_Custo* para aluno então **Sim**

senão **Não**

MONITOR:

Se existe *Monitor* para aluno então **Sim**

senão **Não**

NOTA_RCS_REG:

Recebe o conteúdo de *RCS_Concluido*.NOTA_RCS.

A4.4) Regras de definição de novas Chaves de Entidades no DER do SCG.

Aluno.CH_ALUNO:

Aluno.NUMERO_REG +"-"+ Aluno.DV_REG;

Professor.CH_PROFESSOR:

Professor.NUM_REG_PROF +""+ Professor.DV_REG_PROF; e

Versao_Disciplina.CH_DISCIPLINA:

Versao_Disciplina.PARTE_ALFA_DIS + "" +

Versao_Disciplina.PARTE_NUM_DIS.

A4.5) Análise da Periodicidade de Atualização das Entidades no DER do SGC.

	Requer Mapeamento	**Não requer mapeamento**
Raramente Mudam	------	*ATIVO_REG, CONVENIO_REG, NOTA_RCS FREQUENTA_ALOJ_REG*
Mudam Pouco	------	------
Mudam frequentemente	*COEF_REND_A_REG*	------

Tabela A4.1 – Análise da Periodicidade Atualização Entidade *Aluno*

Atributo frequentemente alterado e que requer mapeamento: *COEF_REND_A_REG*

Nova Entidade: *Historico_Coef_Rend* para receber *Aluno.COEF_REND_A_REG.*

	Requer Mapeamento	**Não requer mapeamento**
Raramente Mudam	------	*STA_PROF*
Mudam Pouco	------	------
Mudam fr equentemente	------	------

Tabela A4.2 – Análise da Periodicidade Atualização Entidade *Professor*

	Requer Mapeamento	**Não requer mapeamento**
Raramente Mudam	------	*STA_DIS*
Mudam Pouco	------	------
Mudam frequentemente	------	------

Tabela A4.3 – Análise da Periodicidade Atualização Entidade *Versao_Disciplina*

A4.6- Definição de Padrão, Valores "Default" e Regras de Conversão para os Atributos no DER do SCG.

Atributo	Formato	"default"	Regras para Conversão
nivel_reg	char(15)	"Graduação"	**1,2,3 → graduação,** **4 → extensão,** **5 → aperfeiçoamento,** **6 → especialização,** **7 → mestrado,** **8 →doutorado,** **9→ pós-doutorado**
centro_reg	char(15)	Nulo	**Substituir código pela descrição**
unidade_reg	char(15)	Nulo	**Substituir código pela descrição**
curso_reg	char(15)	Nulo	**Substituir código pela descrição**
sexo_reg	char(1)	"M"	**"0" → "M"** **"1" → "F"**
nacionalidade_reg	char(15)	Nulo	**1 → brasileiro,** **2 → Naturalizado,** **3 → Estrangeiro**
naturalidade_reg	char(15)	Nulo	**Substituir código pela descrição**
ativo_reg	char(10)	"Ativo"	**"A" → ativa,** **"T" → trancada,** **"C"→ cancelada**
curso_m_reg	char(20)	Nulo	**Substituir código pela descrição**
univers_t_reg	char(20)	Nulo	**Substituir código pela descrição**
Nota_RCS	char(02)	Nulo	-----

Tabela A4.4 – Padrões, valores "default" e Regras de conversão para *Aluno* no DER do SCG.

Atributo	Formato	"default"	Regras para a Conversão
centro_dis	char(15)	Nulo	**Substituir código pela descrição**
unidade_dis	char(15)	Nulo	**Substituir código pela descrição**
curso_dis	char(15)	Nulo	**Substituir código pela descrição**

Tabela A4.5 – Padrões, valores "default" e Regras de conversão para *Versao_Disciplina* no DER do SCG.

Atributo	Formato	"default"	Regras para a Conversão
centro_dtr	char(15)	Nulo	**Substituir código pela descrição**
unidade_dtr	char(15)	Nulo	**Substituir código pela descrição**
curso_dtr	char(15)	Nulo	**Substituir código pela descrição**

Tabela A4.6 – Padrões, valores "default" e Regras de conversão para *Turma* no DER do SCG.

Entidade *Professor*			
Atributo	**Formato**	**"default"**	**Regras para a Conversão**
sta_prof	char(15)	"Ativa"	**"A" → ativa,** **"T" → trancada,** **"C"→ cancelada**

Tabela A4.7 – Padrões, valores "default" e Regras de conversão para *Professor* no DER do SCG.

A4.7- Definição de Minidimensões para a Dimensão Aluno no DER do SCG

A tabela A4.8 apresenta as minidimensões da dimensão **ALUNO**, com seus respectivos atributos.

Minidimensão	Atributo
CURSO	Centro Unidade Curso Turno
DEMOGRAFICA	Idade Sexo Estado_Civil Nacionalidade Naturalidade
ESPACIAL	Bairro Cidade Estado

Tabela A4.8 Minidimensões da Dimensão ALUNO para SCG

A4.8- Relação de Atributos Excluídos no DER do SV.

ENTIDADE: *Vestibulando*

filiacao_pai	- **nome do pai do vestibulando**
filiacao_mae	**- nome da mãe do vestibulando**
endereco	- **endereço do vestibulando**
telefone	- **telefone do vestibulando**
e-mail	- **endereço eletrônico**
RGI	**- número da identidade**
expedidor_RGI	- **órgão expedidor da identidade**
estado_RGI	- **sigla do estado do órgão expedidor da identidade**

A4.9- Regras de categorias apresentadas no DER do SV.

IDADE:

Cálculo da idade a partir do atributo *DATA_NASCIMENTO*.

FAIXA_PONTOS:

PONTUAÇÃO INFERIOR A 5000 : *PONTOS_VESTIBULAR*< 5000;

PONTUAÇÃO ENTRE 5000 e 7000: *PONTOS_VESTIBULAR* >= 5000 e < 7000;

PONTUAÇÃO ENTRE 7000 e 8000: *PONTOS_VESTIBULAT* >= 7000 e < 8000; e

PONTUAÇÃO ACIMA 8000 : *PONTOS_VESTIBULAR* >= 8000.

FAIXA_CLASSIFICACAO:

DEZ PRIMEIROS: *CLASSIFICACAO_VESTIBULAR* <= 10;

ENTRE DECIMO E VIGESIMO: *CLASSIFICACAO_VESTIBULAR* > 10 e <= 20; e

ACIMA DO VIGESIMO: *CLASSIFICACAO_VESTIBULAR* <= 10

A4.10- Definição de padrão, valores "default" e regras de conversão para os atributos no DER do SV.

Atributo	Formato	"default"	Regras para a Conversão
curso_classificado	char(15)	Nulo	**Substituir código pela descrição**
tipo_classificacao	char(15)	NÃO CLAS-SIFICADO	0 → NÃO CLASSIFICADO 1 → 1ª CLASSIFICAÇÃO 2 → 1ª RECLASSIFICAÇÃO 2 → 2ª RECLASSIFICAÇÃO

Tabela A4.9 – Padrões, valores "default" e Regras de conversão para *Registro_Vestibulando* no DER do SV.

A seguir são apresentados os formatos e regras para conversão das informações referentes ao questionário. Como valor "default" será assumido a alternativa "A".

Atributo	Formato	Regras para a Conversão
local_curso_2Grau	Char(20)	A → Cidade do Rio de Janeiro B → Outra Cidade do RJ C → Região SE – RJ D → Região S E → Região N e CO F → Região NE G → Exterior
tipo_escola_1grau	Char(20)	A → Escola Pública B → Escola Particular C → Maior Parte Pública D → Maior Parte Particular
tipo_curso_2grau	Char(20)	A → Atual 2º Grau B → Técnico C → Magistério – Antigo Normal D → Supletivo E → Outro
tipo_escola_2grau	Char(20)	A → Escola Pública B → Escola Particular C → Maior Parte Pública D → Maior Parte Particular

turno_2grau	Char(10)	A → Manhã B → Tarde C → Noite D → Integral
mudanca_ultser_2grau	Char(20)	A → Não B → Sim. Escola Conceituada C → Sim. Localização D → Sim. Financeira E → Sim. Outras
frequencia_cursinho	Char(20)	A → Não B → Sim. Semestre C → Sim. Ano D → Sim. Mais de um Ano
ano_conclusao_2grau	Char(20)	A → Ano atual B → Ano anterior C → Dois anos antes D → Três anos antes E → Quatro ou mais
pretensao_vestibular	Char(20)	A → Único B → Outros com mesma opção C → Outros com outras opções
exame_vestibulares_anterio-res	Char(30)	A → Não B → Sim. Sem classificação curso desejado C → Sim. Com classificação / sem instituição D → Sim. Mudou idéia do curso E → Sim. Problemas financeiros F → Sim. Outros
conhecimento_programa_ concurso	Char(20)	A → Desconhece programas B → Ouviu falar C → Conhece, mas não emprega D → Estuda por eles
turno_preferencia	Char(20)	A → Noite B → Diurno, aceita Noite C → Diurno, não aceita Noite
posicao_curso_UFRJ	Char(40)	A → Curso inscrito B → Outros cursos da mesma área C → Qualquer curso da mesma área D → Determinado curso outra área E → Qualquer Curso / qualquer área
fator_escolha_curso	Char(40)	A → Mercado de Trabalho B → Prestígio social da profissão C → Adequação aptidões pessoais D → Baixa concorrência vagas E → Amplas possibilidades salariais

fator_opcao_UFRJ	Char(30)	A → Única com o curso B → Melhor curso C → Melhor horário D → Pouco procurada – fácil classificação E → Mais fácil acesso F → Seguir opção amigos
expectativa_curso_ universitario	Char(40)	A → Cultura geral ampla B → Voltado ao Mercado de Trabalho C → Voltado para pesquisa D → Formação acadêmica – ativ. prática E → Compreender melhor o mundo F → Melhorar nível de instrução
nivel_instrucao_pai	Char(30)	A → Nenhum B → Menos que 4ª série 1º Grau C → 4ª série 1º Grau D → Mais 4ª série 1º Grau e Menos que 8ª série E → 1º Grau completo F → 2º Grau incompleto G → 2º Grau completo H → Superior incompleto I → Superior completo
nivel_instrucao_mae	Char(30)	A → Nenhum B → Menos que 4ª série 1º Grau C → 4ª série 1º Grau D → Mais 4ª série 1º Grau e Menos que 8ª série E → 1º Grau completo F → 2º Grau incompleto G → 2º Grau completo H → Superior incompleto I → Superior completo
situacao_trabalho_pai	Char(30)	A → Empregado B → Desempregado C → Aposentado D → Vive de renda E → Falecido F → Não tem informação
situacao_trabalho_mae	Char(30)	A → Empregada B → Desempregada C → Aposentada D → Vive de renda E → Falecida F → Não tem informação

ocupacao_pai	Char(30)	A → Empresário B → Proprietário C → Executivo D → Ocupação de nível superior E → Ocupação de nível médio F → Ocupação manual G → Trabalhador rural
ocupacao_mae	Char(30)	A → Empresária B → Proprietária C → Executiva D → Ocupação de nível superior E → Ocupação de nível médio F → Ocupação manual G → Trabalhadora rural
renda_mensal_familia	Char(30)	A → Até 1 SM B → Maior que 1 SM a 3 SM C → Maior que 3 SM a 5 SM D → Maior que 5 SM a 10 SM E → Maior que 10 SM a 20 SM F → Maior que 20 SM a 30 SM G → Maior que 30 SM
participacao_economia_ familia	Char(30)	A → Gastos financiados família B → Trabalha e auxílio financeiro C → Trabalha e auto-sustenta D → Trabalha e contribui p/ família E → Trabalha e responsável p/ família
pretensao_trabalho_curso	Char(20)	A→ Não B → Sim. Estágio C → Sim. Últimos anos D → Sim. Em tempo parcial E → Sim. Em tempo integral
numero_pessoas_familia	Char(10)	A → Vive só B → Duas C → Três D → Quatro E → Cinco F → Seis ou mais
total_dependencias_casa	Char(10)	A → 1 ou 2 B → 3 C → 4 D → 5 E → Mais de cinco
situacao_casa_familia	Char(30)	A → Própria e quitada B → Própria e não quitada C → Alugada D → Outra forma de ocupação
sitio_casapraia_fazenda	Char(10)	A → Sim B → Não

numero_automoveis	Char(20)	A → Não tem B → Tem 1 C → Tem 2 D → Tem mais de 2
total_livros	Char(30)	A → Nenhum B → Até 20 C → de 21 a 50 D → de 51 a 100 E → de 101 a 200 F → de 201 a 500 G → mais de 500
leitura_por_ano	Char(30)	A → Nenhuma B → 1 a 2 C → 3 a 5 D → 6 a 10 F → 11 ou mais
lingua_estrangeira	Char(30)	A → Sim. Fluentemente B → Sim. Razoavelmente C → Não. Gostaria de aprender D → Não. Sem necessidade
principal_meio_informacao	Char(20)	A → Jornal B → Televisão C → Internet D → Rádio E → Revista F → Outras pessoas G → Não se atualiza
le_jornal	Char(30)	A → Não B → Sim. Ocasionalmente C → Sim. Todos os domingos D → Sim. Diariamente
secao_preferida_jornal	Char(30)	A → Política B → Economia C → Esporte D → Cultura E → Notícias internacionais F → Notícias locais G → Quadrinho H → Ciência I → Informática J → Outros assuntos
cursos_extracurriculares	Char(30)	A → Não B → Sim. Línguas estrangeiras C → Sim. Ginástica/balé/esportes D → Sim. Música/arte E → Sim. Outros

acesso_microcomputador	Char(20)	A → Não B → Sim. Em casa C → Sim. Outros locais
utilizacao_microcomputador	Char(20)	A → Não usa B → Trabalhos escolares C → Jogos D → Fins profissionais E → Outros
acesso_internet	Char(20)	A → Não B → Sim. Em casa C → Sim. Outros locais D → Desconhece Internet

Tabela A4.10 – Padrões, valores "default" e Regras de conversão para Questionario do *Vestibulando* no DER do SV.

Atributo	**Formato**	**"default"**	**Regras para a Conversão**
centro	char(15)	Nulo	**Substituir código pela descrição**
unidade	char(15)	Nulo	**Substituir código pela descrição**

Tabela A4.11 – Padrões, valores "default" e Regras de conversão para *Curso* no DER do SV.

A4.11- Regras para cálculo de atributos derivados no DER do SV.

Entidade : *Curso_Oferecido.*

TOTAL_INSCRITOS_1OPCAO =

Sum (*Opcao_Curso* onde

Opcao_Curso.COD_OPCAO_CURSO= Curso_Oferecido.COD_CURSO

e *Opcao_Curso.PRIORIDADE_OPCAO = 1*);

TOTAL_INSCRITOS_2OPCAO =

Sum (*Opcao_Curso* onde

Opcao_Curso.COD_OPCAO_CURSO= Curso_Oferecido.COD_CURSO

e *Opcao_Curso.PRIORIDADE_OPCAO = 2*);

TOTAL_INSCRITOS_3OPCAO =

Sum (*Opcao_Curso* onde

Opcao_Curso.COD_OPCAO_CURSO= Curso_Oferecido.COD_CURSO

e *Opcao_Curso.PRIORIDADE_OPCAO = 3*);

TOTAL_CLASSIFICADOS =

Sum (*Registro_Vestibulando* onde

Registro_Vestibulando.CURSO_CLASSIFICADO =

Curso_Oferecido.COD_CURSO e

Registro_Vestibulando.SEMESTRE_CLASSIFICADO =

Curso_Oferecido.SEMESTRE);

A4.12- Definição de Minidimensões para a Dimensão referente a informações de Vestibulando no DER do SV

Minidimensão	Atributo
ENSINO	local_curso_2Grau tipo_escola_1grau tipo_curso_2grau tipo_escola_2grau turno_2grau mudanca_ultser_2grau frequencia_cursinho ano_conclusao_2grau
VESTIBULAR	pretensao_vestibular exame_vestibulares_anteriores conhecimento_programa_concurso turno_preferencia posicao_curso_UFRJ fator_escolha_curso fator_opcao_UFRJ expectativa_curso_universitario
SOCIO_ECONOMICO	nivel_instrucao_pai nivel_instrucao_mae situacao_trabalho_pai situacao_trabalho_mae ocupacao_pai ocupacao_mae renda_mensal_familia participacao_economia_familia

Minidimensão	Atributo
	pretensao_trabalho_curso numero_pessoas_familia total_dependencias_casa situacao_casa_familia sitio_casapraia_fazenda numero_automoveis total_livros leitura_por_ano lingua_estrangeira principal_meio_informacao le_jornal secao_preferida_jornal cursos_extracurriculares acesso_microcomputador utilizacao_microcomputador acesso_internet
CLASSIFICACAO	faixa_pontos faixa_classificacao tipo_classificacao

Tabela A4.12 Minidimensões da Dimensão

REGISTRO_VESTIBULANDO no SV

ANEXO 5

DICIONÁRIO DE DADOS REFERENTE AO ESTUDO DE CASO UNIVERSIDADE

A5.1) Dicionário de Dados do Sistema Graduação:

ENTIDADE: Ajuda_Custo "**ajuda de custo**"

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
ano_ajuda	char(2)	**número de vias do histórico ou boletim solicitado (index1 = 1)**
entid_ajuda	char(5)	**CPF se index1 = 2, ou senha na pré-inscrição se index1 = 3**
valor_ajuda	char(4)	**continuação do campo anterior**
prz_utl_ajuda	char(2)	**continuação do campo anterior**
prz_con_ajuda	char(2)	**não utilizado**
tipo_ajuda_reg	char(1)	**não utilizado**

ENTIDADE : Alojamento **"alojamento"**

numero_aloj	char(8)	**número do registro do aluno**
dv_aloj	char(1)	**dígito verificador do registro do aluno**
ala_aloj	char(1)	**ala do alojamento**
apt_aloj	char(4)	**apartamento do alojamento**

ENTIDADE : Aluno **"aluno"**

nivel_reg	char(15)	**nível do curso do aluno graduação, extensão, aperfeiçoamento, especialização, mestrado, doutorado, pós-doutorado**
centro_reg	char(15)	**centro do curso no qual o aluno ingressou**

unidade_reg	char(15)	**unidade do curso no qual o aluno ingressou**
curso_reg	char(15)	**curso no qual o aluno ingressou**
sexo_reg	char(1)	**"M" – Masculino e "F" Feminino**
nacionalidade_reg	char(1)	**"1" brasileiro, "2" Naturalizado , "3" – Estrangeiro**
naturalidade_reg	char(15)	**naturalidade**
ativo_reg	char(1)	**"A"- matrícula ativa, "T" trancada, "C"- cancelada**
ano_r_reg	char(2)	**ano da última rematrícula**
periodo_r_reg	char(1)	**período da última rematrícula**
ano_t_reg	char(2)	**ano do último trancamento**
periodo_t_reg	char(1)	**período do último trancamento**
ano_c_reg	char(2)	**ano do cancelamento**
periodo_c_reg	char(1)	**período do cancelamento**
orientador_reg	char(7)	**matrícula do professor orientador do projeto final**
convenio_reg	char(1)	**forma de ingresso do aluno na universidade ("2" se convênio, "3" se cortesia)**
num_hist_reg	char(2)	**número de períodos do histórico do aluno (não contam os períodos zero)**
insc_vest_reg	char(6)	**número de inscrição no vestibular**
pontos_vest_reg	char(5)	**pontos obtidos no vestibular**
class_vest_reg	char(4)	**classificação no vestibular**
curso_pret_reg	char(2)	**código da opção (curso/turno/semestre) conquistada no vestibular**
curso_m_reg	char(2)	**código do curso médio**
estado_m_reg	char(2)	**sigla do estado do curso médio**
ano_conc_m_reg	char(2)	**ano de conclusão do curso médio**
dia_g_reg	char(2)	**dia da colação de grau**
mes_g_reg	char(2)	**mês da colação de grau**
ano_g_reg	char(2)	**ano da colação de grau**
dia_d_reg	char(2)	**dia da emissão do diploma**

mes_d_reg	char(2)	**mês da emissão do diploma**
ano_d_reg	char(2)	**ano da emissão do diploma**
ano_trans_reg	char(2)	**ano da transferência**
periodo_trans_reg	char(1)	**período da transferência**
univers_t_reg	char(2)	**código da universidade de origem**
estado_t_reg	char(2)	**sigla do estado da universidade de origem**
cred_obt_a_reg	smallint	**créditos obtidos acumulados até a última atualização**
coef_rend_a_reg	char(3)	**coeficiente de rendimento acumulado até o período em curso do histórico do aluno**
num_pag_hist_reg	char(1)	**número de páginas a serem impressas no currículo relativas ao período anterior**
cont_tes_reg	char(1)	**controle de defesa de tese (" " ou "X")**
ver_curric_reg	char(3)	**versão do currículo (ano/período)**
dia_u_reg	char(2)	**dia da última atualização**
mes_u_reg	char(2)	**mês da última atualização**
ano_u_reg	char(2)	**ano da última atualização**
ativo_futuro_reg	char(1)	**futuro valor a ser atribuido a "ativo_reg"**
ano_futuro_reg	char(2)	**ano da futura alteração de "ativo_reg"**
per_futuro_reg	char(1)	**período da futura alteração de "ativo_reg"**
data_futuro_reg	char(6)	**data da futura alteração de "ativo_reg"**

ENTIDADE : Disciplina_Vestibular "**disciplina**"

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
nome_dis_vest_reg	char(10)	**nome da disciplina do vestibular**
pont_dis_vest_reg	char(5)	**pontos na disciplina**

ENTIDADE : Equivalencia **"relações de equivalência entre versões de disciplinas"**

eqv_cod_curso	char(5)	**código do curso onde vale a equivalência**
eqv_ano	char(2)	**ano do currículo**
eqv_per	char(1)	**período do currículo**

parte_alfa_dis	char(3)	**parte alfabética do código da disciplina ("MAB")**
parte_num_dis	char(3)	**parte "numérica" do código da disciplina ("123")**
eqv_equacao	smallint	**número da equação**
eqv_lado	tinyint	**lado da equação ("1" se direito, "2" se esquerdo)**

ENTIDADE : Historico **"histórico"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
ano_his	char(2)	**ano do histórico**
periodo_his	char(1)	**período do histórico (ano e período devem ser maior ou igual aos 3 primeiros dígitos do registro do aluno, ou o período é zero e o ano é maior ou igual aos 2 primeiros dígitos)**
nivel_his	char(1)	**igual ao nível da ENTIDADE "aluno"**
centro_his	char(1)	**centro do curso do aluno no ano/período**
unidade_his	char(2)	**unidade do curso do aluno no ano/período**
curso_his	char(2)	**código do curso atual do aluno no ano/período**
byte_calc_cra_his	char(1)	**controle para calcular o coeficiente de rendimento acumulado**
ativo_his	char(1)	**situação do aluno no ano/período**
ocorr_msg_his	tinyint	**quantidade de mensagens neste histórico (1..10)**
ocorr_dis_his	tinyint	**quantidade de disciplinas neste histórico (1..90)**
dia_u_his	char(2)	**dia da última atualização**
mes_u_his	char(2)	**mês da última atualização**
ano_u_his	char(2)	**ano da última atualização**

ENTIDADE : Historico_Disciplina **"disciplinas do histórico"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
ano_his	char(2)	**ano do histórico**
periodo_his	char(1)	**período do histórico**
part_alfa_dis_his	char(3)	**parte alfabética do código da disciplina**
parte_num_dis_his	char(3)	**parte numérica do código da disciplina (o primeiro dígito é alfabético, se for r.c.s)**
turma_dtr	char(3)	**código da turma da disciplina**
conceito_dis_his	char(3)	**de 0,0 a 10,0**
situacao_dis_his	char(1)	**situação na disciplina: aprovado, reprovado por falta, por nota**
mes_atu_his	char(1)	**mês da última atualização ("A" = janeiro, até "L" = dezembro)**

ENTIDADE: Historico_Mensagem **"mensagens do histórico"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
ano_his	char(2)	**ano do histórico**
periodo_his	char(1)	**período do histórico**
part_alfa_msg_his	char(1)	**parte alfabética da chave do tipo de mensagem do histórico**
parte_num_msg_his	char(3)	**parte numérica da chave do tipo de mensagem do histórico**
parte_var_msg_his	char(9)	**dados variáveis de cada mensagem (data etc)**

ENTIDADE : Historico_Obs **"observações no histórico"**

obs_registro	char(9)	**número de registro do aluno**
obs_texto	text	**texto da observação**
ano_his	char(2)	**ano do histórico**
periodo_his	char(1)	**período do histórico**

ENTIDADE: Horário **"horários da turma"**

turno	char(1)	**"D", "V" e "I"**
ini_dtr	char(6)	**hora/minuto/segundo do início da aula**
fim_dtr	char(6)	**hora/minuto/segundo do início da aula**
cod_hor	char(12)	**chave de Horario**

ENTIDADE : Inscricao_Periodo **"inscrição em disciplina"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do registro do aluno**
parte_alfa_dri	char(3)	**parte alfabética do código da disciplina**
parte_num_dri	char(3)	**parte numérica do código da disciplina**
turma_dri	char(3)	**código da turma**
sit_inscr_dri	char(1)	**situação da inscrição (ativo, trancado, transferido, excedido, autorizado pelo C.E.G.)**
sit_inscr_ant_dri	char(1)	**situação da inscrição anterior**
num_per_rcs_dri	char(1)	**se for r.c.s., número do período (1, 2, 3 ...)**

ENTIDADE: LocalDisciplina **"locais de aula"**

parte_alfa_dtr	char(3)	**parte alfabética do código da disciplina**
parte_num_dtr	char(3)	**parte "numérica" do código da disciplina**
turma_dtr	char(3)	**identificador da turma de uma disciplina**
cod_hor	char(12)	**chave de Horario**

ENTIDADE : Monitor **"monitores"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do número de registro do aluno**
mon_disciplina	char(6)	**disciplina da monitoria**
mon_dia_incl	char(2)	**dia da inclusão**
mon_mes_incl	char(2)	**mês da inclusão**
mon_ano_incl	char(2)	**ano da inclusão**

mon_status	tinyint	**ativa ou desativa**

ENTIDADE : Professor **"professores"**

num_reg_prof	char(8)	**número de registro do professor**
dv_reg_prof	char(1)	**dígito verificador do número de registro do professor**
sta_prof	char(1)	**situação da matricula ("A" Ativa, "I" – Inativa)**

ENTIDADE : RCS_Concluido **"r.c.s requisito curricular suplementar concluído"**

numero_reg	char(8)	**número de registro do aluno**
dv_reg	char(1)	**dígito verificador do número de registro do aluno**
ano_concl_rcs	char(2)	**ano de conclusão do r.c.s.**
per_concl_rcs	char(1)	**período de conclusão do r.c.s.**
codigo_rcs	char(6)	**código da disciplina (do r.c.s.)**
ano_inicio_rcs	char(2)	**ano de início do r.c.s.**
per_inicio_rcs	char(1)	**período de início do r.c.s.**
nota_rcs	char(3)	**nota conferida**
conceito_rcs	char(1)	**conceito (aprovado, reprovado etc)**
local_rcs	char(60)	**local de realização do r.c.s.**
descricao_rcs	char(150)	**descrição**
dia_ult_a	char(2)	**dia da última alteração**
mes_ult_a	char(2)	**mês da última alteração**
ano_ult_a	char(2)	**dia da última alteração**
curso_rcs	char(5)	**curso a qual a r.c.s. pertence**

ENTIDADE: Sala **"salas de aula"**

cod_sala_dtr	char(6)	**chave da sala**
bloco_dtr	char(3)	**bloco**
andar_dtr	char(2)	**número do andar**
unidade_dtr	char(2)	**unidade do curso do aluno no ano/período**

ENTIDADE : Turma **"turma"**

centro_dtr	char(1)	**centro a quem é oferecida a turma**
unidad_dtr	char(2)	**unidade a quem é oferecida a turma**
curso_dtr	char(2)	**curso a quem é oferecida a turma**
parte_alfa_dtr	char(3)	**parte alfabética do código da disciplina**
parte_num_dtr	char(3)	**parte "numérica" do código da disciplina**
turma_dtr	char(3)	**identificador da turma de uma disciplina**
numero_alunos_dtr	char(3)	**previsão máxima de alunos**
numero_turma_dtr	smallint,	**número seqüêncial da turma (de 0001 a 9999)**
num_reg_prof	char(8)	**número de registro do professor**
dv_reg_prof	char(1)	**dígito verificador do número de registro do professor**
turno_dtr	char(1)	**turno ("M", manhã, "T", tarde, "N", noite, "I" se a turma não em prioridade de quem irá se inscrever)**
insc_normal_dtr	smallint	**quantidade de alunos com inscrição normal**
insc_tranc_dtr	smallint	**quantidade de alunos que trancaram a inscrição na turma**
insc_transf_dtr	smallint	**quantidade de alunos que se transferiram para a turma**
insc_irr_dtr	smallint	**quantidade de alunos com a inscrição irregular**
insc_autceg_dtr	smallint,	**quantidade de alunos autorizados pelo CEG**
insc_32cred_dtr	smallint,	**quantidade de alunos com excesso de créditos**
horas_dadas_dtr	char(3)	**carga horária da disciplina ou branco (opcional)**
estado_turma_dtr	char(1)	**"A" se ativa, "D" se desativada e "P" se pendente**

ENTIDADE : Versao_Disciplina **"versões das disciplinas"**

parte_alfa_dis	char(3)	**parte alfabética do código da disciplina ("MAB")**
parte_num_dis	char(3)	**parte "numérica" do código da disciplina ("123")**
centro_dis	char(1)	**centro da unidade**
unidade_dis	char(2)	**unidade que oferece a disciplina**
departamento_dis	char(2)	**departamento responsável pela disciplina**
estado_dis	char(1)	**"A" se ativa, "D" se desativada**
nome_dis	char(30)	**nome da disciplina**
carga_teorica_dis	dec(3,1)	**carga de aulas teóricas por semana em horas**
carga_pratica_dis	dec(3,1)	**carga de aulas práticas por semana em horas**
cht_teorica_dis	dec(5,1)	**carga de aulas teóricas no período todo**
cht_pratica_dis	dec(5,1)	**carga de aulas práticas no período todo**
creditos_dis	tinyint	**créditos da disciplina (com um decimal)**
byte_regant	char(1)	**"X" se a disciplina tem versão anterior**
campos_preenc_dis	char(2)	**número de campo da disciplina preenchida**
confere_grau_cred	char(1)	**"X" se a disciplina confere grau**
duracao_maxima	char(1)	**quantos períodos no máximo a disciplina pode durar (só RCS) ou branco**

A5.2) Dicionário de Dados do Sistema Vestibular

ENTIDADE: Curso **"cadastro de cursos da UFRJ"**

Codigo_curso	char(05)	**Código do curso**
nome_curso	char(20)	**descrição do curso**
cod_centro	char(05)	**código do centro**
cod_departamento	char(05)	**código do departamento**

ENTIDADE: Controle_Prova **"mapeamento responsável/sala"**

Cod_local	char(05)	**Código do Local de Prova**
CPF	char(11)	**Controle pessoa física**

data_evento	char(08)	**Data das Provas**

ENTIDADE: Disciplina_Vestibular **"disciplinas avaliadas no vestibular"**

Cod_disciplina	char(08)	**Código disciplina**
descricao	char(20)	**Descrição disciplina**

ENTIDADE: Local_Prova **"local onde as provas serão realizadas"**

cod_departamento	char(05)	**código do departamento**
Bloco	char(02)	**identificação do bloco**
Sala	number	**número da sala**

ENTIDADE: Local_Prova_Vestibulando **"local onde os candidatos farão prova"**

Cod_local	char(05)	**Código do Local de Prova**
Matricula_Vest	char(10)	**Matricula do vestibular**
data_evento	char(08)	**Data das Provas**
hora_inicio	char(06)	**hora início das provas**
hora_fim	char(06)	**hora final das provas**

ENTIDADE: Notas_Vestibulando **"notas dos vestibulandos por disciplina"**

Cod_disciplina	char(08)	**Código disciplina**
Matricula_Vest	char(10)	**Matricula do vestibular**
nota_disciplina	number	**nota na disciplina**

ENTIDADE: Opcao_Curso **"seleção de cursos pelo vestibulando"**

Matricula_Vest	char(10)	**Matricula do vestibular**
Codigo_curso	char(05)	**Código do curso**
Ano_Vestibular	char(04)	**ano do vestibular**
Turno	char(01)	**manha/tarde/integral**
Prioridade_Opcao	char(01)	**primeira/segunda/terceira**

ENTIDADE: Questionário **"questionário biográfico padrão para Candidato"**

Matricula_Vest	char(10)	**Matricula do vestibular**
local_curso_2Grau	char(01)	**local onde fez a maior parte do 2 grau**
tipo_escola_1grau	char(01)	**frequentou o curso de 1 grau**
tipo_curso_2grau	char(01)	**que curso de 2grau fez**
tipo_escola_2grau	char(01)	**frequentou o curso de 2 grau**
turno_2grau	char(01)	**turno em que cursou a maior parte do 2 grau**
mudanca_ultser_2grau	char(01)	**mudou de colégio na última série do 2 grau**
frequencia_cursinho	char(01)	**frequentou cursinho**
ano_conclusao_2grau	char(01)	**ano de conclusão do 2 grau**
pretensao_vestibular	char(01)	**pretensão a outros vestibulares**
exame_vestibulares_anteriores	char(01)	**prestou algum exame vestibular**
conhecimento_programa_concurso	char(01)	**conhecimento dos programas para as provas do concurso**
turno_preferencia	char(01)	**preferência por turno**
posicao_curso_UFRJ	char(01)	**posição frente aos cursos oferecidos pela UFRJ**
fator_escolha_curso	char(01)	**fator principal para escolha de curso**
fator_opcao_UFRJ	char(01)	**fator que influenciou a opção pela UFRJ**
expectativa_curso_universitario	char(01)	**o que espera de um curso universitário**
nivel_instrucao_pai	char(01)	**nível de instrução do pai**
nivel_instrucao_mae	char(01)	**nível de instrução da mãe**
situacao_trabalho_pai	char(01)	**situação do pai no trabalho**
situacao_trabalho_mae	char(01)	**situação da mãe no trabalho**
ocupacao_pai	char(01)	**ocupação principal do pai**
ocupacao_mae	char(01)	**ocupação principal da mãe**

renda_mensal_familia	char(01)	**renda mensal total da família**
participacao_economia_familia	char(01)	**participação na vida econômica da família**
pretensao_trabalho_curso	char(01)	**pretensão a trabalhar durante o curso superior**
numero_pessoas_familia	char(01)	**total de pessoas na família**
total_dependencias_casa	char(01)	**total de dependências da casa**
situacao_casa_familia	char(01)	**situação da casa em que a família reside**
sitio_casapraia_fazenda	char(01)	**família possui sítio, casa de praia, fazenda (fim de semana e férias)**
numero_automoveis	char(01)	**total de automóveis**
total_livros	char(01)	**total de livros em casa**
leitura_por_ano	char(01)	**total de livros, em média lidos por ano**
lingua_estrangeira	char(01)	**domina uma língua estrangeira**
principal_meio_informacao	char(01)	**principal meio de informação**
le_jornal	char(01)	**realiza leitura de jornal**
secao_preferida_jornal	char(01)	**qual a seção preferida do jornal**
cursos_extracurriculares	char(01)	**frequenta cursos extracurriculares**
acesso_microcomputador	char(01)	**tem acesso a microcomputador**
utilizacao_microcomputador	char(01)	**utiliza microcomputador para que fins**
acesso_internet	char(01)	**tem acesso a internet**
media_vestibular	number	**média obtida no vestibular**
classificacao_vestibular	char(03)	**classificação geral**
forma_classificacao	char(01)	**forma da classificação**

ENTIDADE: Responsavel **"cadastro de responsáveis por acompanhar as provas"**

CPF	char(11)	**Controle pessoa física**
nome	char(40)	**Nome do responsável**

endereço	char(30)	**endereço domiciliar**
bairro	char(15)	**bairro domiciliar**
cidade	char(15)	**cidade domiciliar**
estado	char(02)	**estado domiciliar**

ENTIDADE: Vagas_Oferecidas **"vagas por curso para o vestibular"**

Codigo_curso	char(05)	**Código do curso**
Ano_Vestibular	char(04)	**ano do vestibular**
Turno	char(01)	**manha/tarde/integral**
Total_Vagas	number(03)	**total de vagas de curso por turno**

ENTIDADE: Vestibulando **"alunos inscritos no vestibular"**

Matrícula_Vest	char(10)	**Matricula do vestibular**
RGI	char(09)	**Registro geral de identificação**
CPF	char(11)	**Controle pessoa física**
nome	char(40)	**Nome do candidato**
endereço	char(30)	**endereço domiciliar**
bairro	char(15)	**bairro domiciliar**
cidade	char(15)	**cidade domiciliar**
estado	char(02)	**estado domiciliar**
telefone	char(12)	**ddd + telefone**
e_mail	char(20)	**endereço eletrônico**
filiacao_pai	char(40)	**nome do pai**
filiacao_mae	char(40)	**nome da mãe**
data_nascimento	char(08)	**data de nascimento**
naturalidade	char(15)	**naturalidade**

REFERÊNCIAS BIBLIOGRÁFICAS

ADELMAN, Leonard. **Evaluating Decision Suport and Expert Systems.** Nova York: John Wiley & Sons, Inc., 1992. 232 p.

BATINI, Carlo, CERI, Stefano, NAVATHE, Shamkant B. **Conceptual Database Design: An Entity-Relationship Approach.** Califórnia: The Benjamin/Cummings Publishing Company, Inc., 1992. 470 p.

CAMPOS, Maria Luiza, FILHO, A.V. Rocha. **Data Warehouse.** In: XVII Congresso da Sociedade Brasileira de Computação – XVI Jornada de Atualização em Informática. Brasília, 1997, pg 221-261.

CHAUDHURI, Surajit, DAYAL, Umeshwar. **An Overview of Data Warehousing and OLAP Technology.** Disponível na INTERNET via http://bunny.cs.uiuc.edu/sigmod/sigmod_record/9703/chaudhuri.ps. Arquivo consultado em 1997.

COREY, Michael J., ABBEY, Michael. **Oracle Data Warehousing**. Califórnia: Oracle Press™ Edition, 1996. 384p.

COUGO, Paulo Sérgio. **Modelagem Conceitual e Projeto de Banco de Dados.** Rio de Janeiro: Editora Campus, 1997. 284 p.

DEVLIN, Barry. **Data Warehouse from Architecture to Implementation**. Massachusetts: Addison-Wesley, 1997. 432 p.

DYCHÉ, Jill. **Scoping Your Data Mart Implementation**. Disponível na INTERNET via http://www.dbmsmag.com/9808d13.html. Arquivo consultado em 1998.

FERREIRA, Aurélio Buarque de H. **Novo Dicionário da Língua Portuguesa – 2ª Edição Revisada e Aumentada.** Rio de Janeiro: Editora Nova Fronteira, 1986, p. 1656.

FIRESTONE, Joseph M. . **Data Warehouse and Data Marts: A Dynamic View .** Disponível na INTERNET via http://www.dkms.com/DWDMDV.html. Arquivo consultado em 1997.

_____ . **Dimensional Modeling and E-R Modeling in the Data Warehouse.** Disponível na INTERNET via http://www.dkms.com/AI/DMERDW.html. Arquivo consultado em 1998 (a).

_____ . **Architectural Evolution in Data Warehousing and Distribuited Knowledge Management Architecture**. Disponível na INTERNET via http://www.dkms.com/ARCHEV.html. Arquivo consultado em 1998 (b).

GOLFARELLI, Matteo, MAIO, Dario , RIZZI, Stefano. **Conceptual Design of Data Warehouse from E/R Schemes.** Disponível na INTERNET via http://www.csr.unib.it/~golfare/dw.html . Arquivo consultado em 1998.

GUPTA, Vivek R. **An Introdution to Data Warehousing.** Disponível na INTERNET via http://www.system-services.com/dwintro.html . Arquivo consultado em 1997.

HACKNEY, Douglas. ***Data Warehouse Delivery: Who Are You? – Part I.*** DM Review Magazine, Vol. 8, Nº 2, Fevereiro, 1998.

INMOM, W.H.. **Data Mart Does Not Equal Data Warehouse.** Disponível na INTERNET via http://www.dmreview.com/issues/1998/may/articles/may98_38.htm. Artigo consultado em 1998.

_____, W.H. .**Como construir o Data Warehouse.** Rio de Janeiro: Editora Campus, 1997. 388p.

INMOM, W.H, HACKATHORN, Richard D. **Como Usar o Data Warehouse.** Rio de Janeiro: Editora Infobook, 1997. 277p.

KIMBALL, Ralph. **Is ER Modeling Hazerdous to Dss?** Disponível na INTERNET via http://www.dbmsmag.com/9510d15.html. Arquivo consultado em 1995.

_____. **Dangerours Preconceptions.** Disponível na INTERNET via http://www.dbmsmag.com/9608d15.html. Arquivo consultado em 1996-a.

_____. **Data Warehouse Toolkit.** São Paulo: Makron Books, 1996-b. 388 p.

_____. **A Dimensional Modeling Manifesto – Drawing the Line Between Dimensional Modeling and ER Modeling**. Disponível na INTERNET via http://www.dbmsmag.com/9708d15.html. Arquivo consultado em 1997.

KIMBALL, Ralph, REEVES, Laura, ROSS, Margy, THORNTHWAITE, Warren. **The Data Warehouse Lifecycle Toolkit – Expert Methods for Designing, Developing and Deploying Data Warehouses**. New York: John Wiley & Sons, Inc., 1998. 771 p.

MACHADO, Felipe Nery R., ABREU, Maurício. **Projeto de Banco de Dados, Uma Visão Prática.** São Paulo: Editora Érica, 1996.

MCGUFF, Frank. **Designing the Perfect Data Warehouse - Data Modeling for Data Warehouse.** Disponível na INTERNET via http://member.aol.com/fmcguff/dwmodel/perfect.htm. Arquivo consultado em 1998.

MELO, Rubens Nascimento. **Data Warehousing (Tutorial).** In: XIII SBBD - Simpósio Brasileiro de Banco de Dados, Salvador, 1997.

MEYER, Don, CANNON, Casey. **Building a Better Data Warehouse.** New Jersey: Prentice-Hall PTR, 1998. 227 p.

POE, Vidette, KLAUER, Patricia, BROBST, Stephen. **Building a Data Warehouse for Decision Support.** New Jersey. Prentice-Hall, Inc, 1998. 285 p.

RADEN, Neil. **Modeling the Data Warehouse.** Disponível na INTERNET via http://members.aol.com/nraden/iw0196_1.htm. Arquivo consultado em 1996.

RUBINI, Eduardo. **Definindo Modelos "Star Schema" para Projetos de Data Warehouse**. In: Developers & Object Forum'97 , 1997.

SILVERSTON, Len, INMON, W.H., GRAZIANO, Kent. **The Data Model Resource Book - A Library of Logical Data Models and Data Warehouse Designs**. Nova York: Wiley Computer Publishing, 1997. 335 p.

TANLER, Rick. **Intranet Data Warehouse.** Rio de Janeiro: Livraria e Editora Infobook S.A., 1998. 394 p.

THOMSEN, Erik. **OLAP Solutions - Building Multidimensional Information Systems**. New York: John Wiley & Sons,Inc , 1997. 576 p.

WU, Ming_Chuan, BUCHMANN, Alejandro P. **Research Issues in Data Warehousing.** Disponível na INTERNET via http://www.informatik.thdarmstadt.de/DVS1/staff/wu/btwllncs.ps .Arquivo consultado em 1997.