UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

CLÁUDIO ARRUDA WAGNER

ESTUDO PARA IMPLANTAÇÃO DE UM DATA WAREHOUSE EM UM AMBIENTE EMPRESARIAL

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação
Orientador: Prof. Vitório Bruno Mazzola

FLORIANÓPOLIS
2003

Esta dissertação é dedicada,
com muito amor, especialmente, à
minha família (Anamáris, minha
esposa, e meus filhos, Lucas e Marina),
pelo carinho, paciência e apoio
constante que tiveram durante o
período deste mestrado.

AGRADECIMENTOS

A Deus Todo Poderoso, pela vida e por tudo que conquistei.
A Lauro e Ângela, pelo incentivo e apoio financeiro.
A Washington Luiz de Lira e Juano Del Prado, pela oportunidade na utilização do *Data Warehouse*.
Aos meus eternos avós Joaquim e Ceci, que contribuíram direta ou indiretamente nos planos de estudos.
Aos meus pais, Alcione e Angelita.
Em especial, agradeço ao meu orientador Dr. Vitório Bruno Mazzola, pela sua valiosa orientação.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Para tomar decisões de negócios rápidos, precisos e valiosos, muitas organizações decidem criar um *Data Warehouse*, que permite às empresas conhecerem os fatores que afetam o seu negócio. *Data Warehouse* não é um produto, é um processo evolutivo; não é um *software* que pode ser comprado e instalado em todos os computadores da empresa em algumas horas. Na realidade, sua implantação exige a integração de vários produtos e processos. Este trabalho propõe apresentar as etapas de implantação do *Data Warehouse* em um ambiente empresarial, descrevendo-as conforme aconteceram em uma Cooperativa Médica. Como metodologia para construção de um *Data Warehouse*, foi adotada a proposta por Kimball com uma série de modificações. Nesta metodologia são descritas as fases que podem ser seguidas para construir-se um *Data Warehouse*.

Palavras-chave: *Data Warehouse;* Data Mart; Data Mining.

ABSTRACT

In order to take decisions regarding quick, precise and valuable businesses many organizations decide to create a Data Warehouse, which allows the companies to know the factors that influence their business. In the stage of implantation the Data Warehouse may not succeed if there is a lack of vision and understanding of the business."Data Warehouse isn't a product, but an evolutionary process". The Data Warehouse isn't also a software which could be bought and installed in every computer of the company but in a few hours. In reality, its installement requires the integration of a number of products and processes. This work proposes to present the processes to implant the Data Warehouse in a business setting, showing the stages of the Data Warehouse process at one Medical Association. As a methodology for the construction of a Data Warehouse, the Kimball methodology was used with several modifications. On this methodology the steps that can be followed for building a data warehouse are described.

Key-words: Data Warehouse; Data Mart; Data Mining.

1 INTRODUÇÃO

As organizações procuram ativamente tornarem-se mais competitivas e rentáveis. Para obter vantagem competitiva, as companhias precisam acelerar o processo de tomada de decisão, devendo, para isso, reagir rapidamente às modificações de ambientes, normalmente através da análise, planejamento e execução de ações táticas ou estratégicas adequadas.

Um dos pontos-chave para acelerar a tomada de decisão é ter informações corretas no momento oportuno e facilmente acessíveis. O conceito de *Data Warehouse* engloba todo o conjunto de informações com o valor do negócio, não se restringindo a simples extrações de informações, com fins diversos, mas, sim, ao negócio como um todo.

Construir um *Data Warehouse* não é uma tarefa fácil, sobretudo nas grandes corporações, onde as redes de informações, muitas vezes, baseiam-se em múltiplos sistemas operacionais e em enorme quantidade de aplicativos espalhados pelos diversos departamentos. A integração dos dados reunidos em diferentes áreas pode demandar meses e meses de trabalho, sem falar nos custos de implantação dessa solução. Mas essa é uma dificuldade que merece ser enfrentada, até porque uma boa implementação de projetos envolvendo o conceito de Customer Relationship Management (CRM) e a extração de relatórios de Bussiness Intelligence (BI) dependem, em larga escala, da estrutura desses enormes armazéns de dados.

O presente trabalho tem por objetivo apresentar os processos para implantação de um *Data Warehouse*, utilizando as abordagens desenvolvidas por Raph Kimball. Para tanto, está estruturado da seguinte forma:

- no capítulo II, são abordados a evolução dos sistemas de apoio à decisão, os principais conceitos de *Data Warehouse*, diferenças entre dados operacionais e informacionais, ambiente, arquitetura, modelagem, ferramentas, etapas para implantação do *Data Warehouse* na empresa e metodologia, segundo Raph Kimball e Willan H. Inmon.

- no capitulo III, é apresentado um estudo de caso: "*Data Warehouse em uma Cooperativa Médica"*.
- no capitulo IV, é relacionada, sucintamente, a seqüência de ações realizadas neste trabalho e as futuras atividades de pesquisa que poderão ser desenvolvidas.

2 ESTUDO DE UM DATA WAREHOUSE

Neste capítulo, serão abordados os principais conceitos, arquitetura, modelagem, ferramentas, metodologia, segundo Raph Kimball e Willan H. Inmon e as etapas para implantação de um *Data Warehouse* em um ambiente empresarial.

2.1 Evolução dos Sistemas de Apoio à Decisão

A evolução dos *Sistemas de Apoio à Decisão* (SAD) pode ser dividida em cinco fases entre 1960 e 1980. No início da década de 1960, o mundo da computação consistia na criação de aplicações individuais que eram executadas sobre arquivos mestres, caracterizadas por programas e relatórios.

Aproximadamente em 1965, o crescimento dos arquivos mestres e das fitas magnéticas explodiu, surgindo problemas como: a complexidade de manutenção dos programas e do desenvolvimento de novos programas, a quantidade de *hardware* para manter todos os arquivos mestres e a necessidade de sincronizar dados a serem atualizados.

Por volta de 1970, surgiu a tecnologia de *Dispositivos de Armazenamento de Acesso Direto (DASD)*, substituindo as fitas magnéticas pelo armazenamento em disco. Com o DASD, surgiu um novo tipo de *software* conhecido como *Sistema de Gerenciamento de Banco de Dados (SGBD)* que tinha o objetivo de tornar o armazenamento e o acesso a dados, no DASD, mais fáceis para o programador. Com o SGBD, surgiu a idéia de um "*Banco de Dados*" que foi definido como: "***uma única fonte de dados para todo o processamento***".

Aproximadamente em 1975, surgiu o processamento de transações *on-line* de alta performance, que permitiu o uso do computador para tarefas que antes não eram viáveis como: controlar sistemas de reservas, sistemas de caixas bancários, sistemas de controle de produção e outros.

Até o início da década de 1980, novas tecnologias, como os PC's e as L4G's (linguagens de quarta geração) surgiram no contexto da informática. O usuário final passou a controlar diretamente os sistemas e os dados, descobrindo que era possível utilizar as informações para outros objetivos além de atender ao processamento de transações on-line de alta performance. Foi nesse período, também, que se tornou viável a construção dos MIS (*Management Information Systems*), hoje conhecido como *Sistemas de Apoio à Decisão (SAD),* que consistia em processamento utilizado para subsidiar decisões gerenciais, conforme Inmon (1997).

2.2 A Diferença entre Dados Operacionais e Informacionais

Os *Sistemas de Suporte à Decisão (SSD)* e *Sistemas de Informações Executivas (SIE)* possuem funcionalidade e desempenho diferentes dos sistemas de produção da empresa. Estes sistemas recuperam e atualizam um registro por vez, geralmente atendendo a muitos usuários de forma concorrente, exigindo um tempo de resposta imediato; aqueles normalmente lidam com poucos usuários por vez e os requisitos em termos de tempo de resposta podem não ser críticos. No entanto, freqüentemente lidam com consultas complexas, não antecipadas ou previstas, envolvendo grande quantidade de registros básicos referentes aos processos operacionais da empresa.

Os aplicativos SSD e SIE necessitam de dados consistentes, normalmente originários de mais de um sistema de produção, organizados de forma que favoreçam o trabalho por ferramentas de análise de dados.

Os bancos de dados que oferecem recursos para suporte a SSD e SIE devem ser capazes de oferecer um bom tempo de resposta para consultas que recuperam grandes conjuntos de dados agregados e históricos.

SSD e SIE usualmente lidam com tendências e não com um único instante de tempo: "*cada elemento de dado é acompanhado do correspondente período de tempo a que se refere*". A importância em separar dados que dão suporte aos sistemas de caráter operacional da empresa daqueles que subsidiam os processos gerenciais e de

suporte à decisão está no fato de que cada tipo de aplicação pode se concentrar naquilo que faz melhor, oferecendo melhor funcionalidade e desempenho para seu caso específico.

2.3 Conceitos de *Data Warehouse*

Não há uma definição precisa sobre o que é e o que constitui um *Data Warehouse*, segundo afirma Sen (1998). Observa-se, dentre os vários autores que escrevem sobre o assunto, uma grande quantidade e diversidade de definições: DATE (1986), INMON (1992), BABCOCK (1995), KIMBALL (1996), INMON (1997), CORBA (1997), KIMBALL (1998), GRAY (1998), GARDNER (1998), SEN, (1998), POE (1998).

Date (1986) define um banco de dados tradicional como sendo composto de uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de uma empresa.

Esta afirmação leva-nos a refletir sobre o que são realmente estes dados operacionais e o que eles têm em comum com os dados utilizados nos sistemas de *Data Warehouse*.

Todas as empresas que possuem algum nível de informatização armazenam os dados de "produção", ou seja, as informações geradas a partir das operações do seu dia-a-dia. Porém estes dados não formam necessariamente os dados operacionais da empresa. Em Date (1986), temos uma classificação que distingue os dados em “de entrada”, “de saída” e ainda outros.

Os dados de entrada são todas as informações que estão num meio externo ao sistema e, de alguma forma, passam a fazer parte do mesmo, como a digitação num terminal. Uma vez digitadas, estas informações podem até mesmo modificar os dados operacionais, apesar de não fazer parte do banco de dados.

Os dados de saída são mensagens e relatórios gerados no sistema cujas informações podem não fazer parte dos dados operacionais, mas serem o resultado de aplicações de fórmulas e totalizações sobre os mesmos. De acordo com Inmon (1992), são uma coleção de dados orientada ao assunto, integrada, não-volátil e variável em relação ao tempo, que tem por objetivo dar apoio aos processos de tomada de decisão. Segundo Babcock (1995), constituem um repositório de dados sumarizados ou agregados de forma simplificada, provenientes de sistemas operacionais. Os usuários obtêm os dados para suporte a decisão a partir de ferramentas geradas de relatório ou de acesso a dados". O autor definiu, ainda, que *Data Warehouse* é informacional e não operacional; é orientado a análises e decisões e não a processamento de transações; usualmente é cliente/servidor e não baseado em sistemas legados.

Em Inmon (1997) encontramos o conceito de que um *Data Warehouse* (que pode ser traduzido como armazém de dados) é um banco de dados que armazena dados sobre as operações da empresa (vendas, compras, etc) extraídos de uma fonte única ou múltipla e, transforma-os em informações úteis, oferecendo um enfoque histórico, para permitir um suporte efetivo à decisão. Uma filosofia de *Data Warehouse* pode prover múltiplas visões da informação para um espectro de usuários. O poder deste conceito é que possibilita aos usuários acesso a dados de fontes não relacionadas para a procura de respostas a questões de negócios, ou seja, o *Data Warehouse* permite que os usuários prevejam informações relevantes de dados antes independentes.

Corey e Abbey (1997) definem como uma coleção de informação corporativa, derivada diretamente de sistemas operacionais e algumas fontes externas. Tem o propósito específico de suportar decisões de negócios e não operações de negócios.

Kimball (1996) afirma que *Data Warehouse* é o lugar onde as pessoas podem acessar seus dados. Em 1998, define-o como a fonte de dados de consulta do empreendimento, onde as pessoas podem acessar as informações que lhe são necessárias.

Seguindo Gray e Watson (1998), podemos afirmar que um *Data Warehouse* é tipicamente um sistema de banco de dados dedicado, que é separado dos sistemas OLAP da organização.

Já Gardner (1998), diz que *Data Warehouse* é um processo, e não um produto, para a montagem e administração de dados provenientes de várias fontes com o propósito de obter uma visão simples e detalhada de parte de todo o negócio.

Consultando Sen (1998), encontramos a afirmação de que os *Data Warehouse* são construídos no interesse de suporte à decisão de negócios e contêm dados históricos sumarizados e consolidados provenientes de registros individuais de bancos de dados operacionais. Poe (1998) o define como uma base de dados analítica que dá apoio a processos decisórios mais recursos de acesso intuitivos.

2.4 Utilização de um *Data Warehouse*

A informação é o bem mais valioso para uma empresa. Decisões precisam ser tomadas rápida e corretamente, usando todo o dado disponível. Os sistemas convencionais de informática não são projetados para gerar e armazenar as informações estratégicas, o que torna os dados vagos e sem valor para apoio ao processo de tomada de decisões das organizações que, normalmente, são tomadas com base na experiência dos administradores, quando poderiam, também, ser baseadas em fatos históricos que foram armazenados pelos diversos sistemas de informação utilizados pelas organizações. No *Data Warehouse* possuímos *uma fonte única de informações com dados consolidados*.

2.5 Tipos e Aplicações de um *Data Warehouse*

Podemos destacar os seguintes tipos de *Data Warehouse*:

- *Marketing*: Avalia a performance comercial de um produto ou serviço a partir de diversas perspectivas diferentes.
- *Financeiro*: Monitora a performance comercial em termos financeiros.

- *Comportamental*: Contem informações individuais a respeito de cada cliente e seus comportamentos.

As aplicações típicas do *Data Warehouse* em uma empresa podem ser classificadas em dois grandes conjuntos:

- *Aplicações do Negócio*: constituem as aplicações que dão suporte ao dia-a-dia do negócio da empresa, que garantem a sua operação, também chamadas de sistemas de produção;
- *Aplicações sobre o Negócio*: são as aplicações que analisam o negócio, ajudando a interpretar o que ocorreu e a decidir sobre estratégias futuras para a empresa - compreendem os *Sistemas de Suporte a Decisão* (SSD) e *Sistemas de Informações Executivas* (SIE).

As arquiteturas de dados adequadas para dar suporte a estes dois tipos de aplicações devem estar baseadas, analogamente, em dois ambientes de bancos de dados: os operacionais (para dar suporte às *aplicações do negócio*) e os bancos de dados para suporte à decisão (que embasam as *aplicações sobre o negócio*).

2.6 Vantagens de um *Data Warehouse*

☛ **Simplicidade** ☚ Facilita a administração da empresa porque fornece uma imagem simples da realidade, com integração de vários dados de sistemas diferentes. Permite que os sistemas operacionais continuem em uso, transformando os dados inconsistentes desses sistemas em um conjunto de dados coerentes, que se constituem em informações vitais para as empresas. As operações atuais podem ser monitoradas e comparadas com as operações passadas; futuras operações podem ser previstas racionalmente; novos processos podem ser inventados; os sistemas operacionais podem ser alterados para suportar estes processos. O *Data Warehouse* também pode armazenar um grande número de dados históricos que auxiliam as empresas na tomada de decisões. Oferece o benefício de ser único, com dados centralizados, mas que mantém uma estrutura de cliente/servidor. Além disso, *Data Warehouse* é um sistema para empresas médias e grandes, o que melhora a distribuição das informações internamente.

☛ **Qualidade dos dados** ☚ Proporciona consulta em dados de maior qualidade o que garante a consistência, precisão e documentação, além de aumentar a produtividade dos usuários através de utilização de ferramentas OLAP e de *Data Mining*.

☛ **Acesso rápido** ☚ **P**ermite aos usuários recuperar, rapidamente, os dados necessários para suas consultas, eliminando o trabalho de busca em vários sistemas operacionais, uma vez que todas as informações estão em um único local.

☛ **Facilidade de uso** ☚ A maioria das ferramentas de consulta facilita o acesso aos dados pois trabalha com interfaces gráficas e comandos predefinidos, o que torna a análise das informações armazenadas no *Data Warehouse* uma tarefa intuitiva para os usuários finais.

☛ **Separa as operações de decisão das operações de produção** ☚ Como os dados do *Data Warehouse* ficam separados dos dados dos sistemas operacionais, mesmo sendo continuamente atualizados com informações sobre as operações realizadas, os gerentes e analistas de negócios podem fazer estudos nestes dados sem sobrecarregar os sistemas operacionais.

☛ **Vantagem competitiva** ☚ Auxilia o administrador a gerenciar melhor utilizando o conhecimento incorporado que possibilita à empresa ser mais competitiva, entendendo as necessidades dos clientes e detectando mais rapidamente as demandas de mercado. Esta vantagem pode compensar o grande custo de implantação de um *Data Warehouse.*

☛ **Custo de operações** ☚ Oferece uma boa base para o desenvolvimento de novos sistemas operacionais, além de eliminar o uso de arquivos baseados em papéis e, uma vez coberto o investimento inicial, o grupo de tecnologia da informação da empresa, normalmente, consome menos recursos do que antes da implantação do *Data Warehouse*, já que as informações ficam centralizadas e podem ser acessadas facilmente pelos usuários finais.

☛ **Administração do fluxo de informação** ☚ Recebe uma grande quantidade de dados de várias fontes operacionais e os envia para várias aplicações *Front End*. Para se adaptarem às mudanças nas regras de negócio das empresas, os sistemas operacionais e as estruturas dos dados são constantemente modificados. No *Data Warehouse*, isto dificilmente ocorre, pois os metadados auxiliam na configuração dos dados para que eles atendam aos novos requisitos da empresa.

☛ **Habilitação do processamento paralelo** ☚ O processamento paralelo ajuda os usuários a realizar consultas no *Data Warehouse* mais rapidamente, pois suporta grandes demandas em ambientes cliente/servidor, onde os usuários podem fazer perguntas ou consultas simultâneas que exijam um processamento intensivo. Com o processamento paralelo, o *Data Warehouse* oferece uma melhor relação de preço/performance.

☛ **Infra-estrutura computacional** ☚ Ajuda as organizações a montar uma infra-estrutura que pode suportar mudanças nos sistemas operacionais e na estrutura dos seus negócios.

☛ **Valores quantitativos** ☚ Pode mostrar um retrospecto realista da evolução da empresa, pois possui medidas quantitativas que permitem a comparação e a análise em períodos de vários anos.

☛ **Segurança** ☚ O fato dos usuários do *Data Warehouse* não acessarem diretamente as base de dados dos sistemas operacionais aumenta a segurança destas informações, além de diminuir o número de acessos aos mesmos.

2.7 Desvantagens de um Data Warehouse

☛ **Complexidade no desenvolvimento** ☚ Uma empresa não pode simplesmente comprar um *Data Warehouse*. É necessário construir um ambiente composto de hardware e software como bancos de dados, ferramentas de extração de dados, ferramentas de recuperação de dados etc. Um *Data Warehouse* deve atender às necessidades específicas de uma empresa na construção deste ambiente individualizado, sendo fundamental ter muito conhecimento das necessidades predefinidas para a construção da estrutura, definições e fluxo dos dados, assim como na escolha do *hardware* e *software* necessários. O desenvolvimento do *Data Warehouse* requer um senso de antecipação sobre as necessidades futuras dos usuários, assim como a previsão de possíveis alterações nas regras de negócio da empresa. Definir como aumentar o *Data Warehouse* por causa da demanda de dados, tanto em volume como em complexidade, torna o seu desenvolvimento bastante difícil e requer uma equipe de especialistas.

☛ **Tempo de desenvolvimento** ☚ Como é uma tarefa complexa, é natural que também seja demorada. Estudos indicam que, em média, um ambiente completo de *Data Warehouse* demora de dois a três anos para ficar pronto, o que pode ser muito tempo para uma empresa que necessita de um ambiente de suporte à decisão em um curto espaço de tempo.

☛ **Alto custo de desenvolvimento e administração** ☚ Um *Data Warehouse* pode consumir milhares de reais até que esteja pronto para ser utilizado, e continuará a consumir recursos durante toda sua "vida" útil, pois necessitará de constantes manutenções.

☛ **Treinamento** ☚ **O**s usuários do *Data Warehouse* devem ser constantemente treinados e comunicados das mudanças. Isto se deve ao fato de que é importante que todos estejam aptos a retirar o máximo de informações possíveis que o *Data Warehouse* oferece.

2.8 Erros na Implantação de um *Data Warehouse*

A fase de levantamento do *Data Warehouse*, onde a equipe de consultoria está em contato direto com os usuários tomadores de decisão, está ligada ao sucesso ou ao fracasso do projeto. Alguns pontos simples, como estar conversando com as pessoas erradas, a demora na implantação do projeto ou a falta de flexibilidade do mesmo podem levá-lo a construir um amontoado de dados estáticos e inúteis em vez do *Data Warehouse*.

Outros aspectos também podem prejudicar o projeto *Data Warehouse*:

- Começar o projeto com o tipo errado de "patrocínio".
- Gerar expectativas que não podem ser atendidas, frustrando os executivos quando da utilização do *Data Warehouse.*
- Empregar expressões politicamente ingênuas como: “Isto vai ajudar os gerentes a tomar decisões melhores”.
- Carregar o *Data Warehouse* com informações desnecessárias “só porque estavam disponíveis".
- Escolher um gerente para o *Data Warehouse* que seja voltado para a tecnologia e não para o usuário.
- Focalizar o *Data Warehouse* em dados tradicionais internos orientados a registro e ignorar o valor potencial de dados textuais, imagens, som, vídeo e dados externos.
- Fornecer dados com definições confusas e sobrepostas.
- Acreditar nas promessas de desempenho, capacidade e escalabilidade dos vendedores de produtos para *Data Warehouse.*
- Usar *Data Warehouse* como uma justificativa para modelagem de dados e uso de ferramentas case.

- Não descobrir quem são os verdadeiros conhecedores do negócio.
- Não entender os objetivos e as demandas dos usuários finais.
- Fazer com que tudo fique mais complicado do que o necessário.
- Permitir que o sistema fique extremamente lento.
- Prolongar o projeto por mais de um ano.
- Fazer com que o *Data Warehouse* não se adapte a novos tempos e regras de negócio.

2.9 Questões Críticas em um *Data Warehouse*

Algumas questões representam verdadeiros desafios na implantação de um *Data Warehouse*:

- integração de dados e metadados de várias fontes;
- qualidade dos dados: limpeza e refinamentos;
- sumarização e agregação de dados;
- sincronização das fontes com o *Data Warehouse* para assegurar a atualização;
- problemas de desempenho relacionados ao compartilhamento do mesmo ambiente computacional para abrigar as bases de dados corporativas operacionais e o *Data Warehouse*.

2.10 Abordagens para o Desenvolvimento de um *Data Warehouse*

De acordo com Weldon (1997) o sucesso do desenvolvimento do Data Warehouse depende fundamentalmente de uma escolha correta da estratégia a ser adotada, de forma que seja adequada às características e necessidades específicas do ambiente onde será implementado. Existe uma variedade de abordagens para o

desenvolvimento de Data Warehouse, devendo-se fazer uma escolha fundamentada com pelo menos três dimensões: escopo do *Data Warehouse* (departamental, empresarial, etc), grau de redundância de dados e tipo de usuário alvo.

O escopo do *Data Warehouse* pode ser tão amplo quanto aquele que inclui todo o conjunto de informações de uma empresa ou tão restrito quanto um *Data Warehouse* pessoal de um gerente. Quanto maior o escopo, mais valor o *Data Warehouse* tem para a empresa e mais cara e trabalhosa sua criação e manutenção. Por essa razão, muitas organizações tendem a começar com um ambiente departamental e só depois de obter um retorno de seus usuários, expandem seu escopo.

Existem, essencialmente, três níveis de redundância de dados: o *Data Warehouse* virtual, o *Data Warehouse* centralizado e o *Data Warehouse* distribuído.

O *Data Warehouse* virtual consiste em prover os usuários finais com facilidades adequadas para extração das informações diretamente dos bancos de produção, o que não caracteriza redundância, mas pode sobrecarregar o ambiente operacional.

O *Data Warehouse* central constitui-se em um único banco de dados físico que contém todos as informações para uma área funcional específica, um departamento ou uma empresa, indicado quando existe necessidade comum de informações. Normalmente contém dados oriundos de diversos bancos operacionais, devendo ser carregado e mantido em intervalos regulares.

O *Data Warehouse* distribuído, como o nome indica, possui seus componentes distribuídos por diferentes bancos de dados físicos. Geralmente possui um grau de redundância alto e, por conseqüência,exige procedimentos mais complexos de carga e manutenção.

Os padrões de uso do *Data Warehouse* também constituem um fator importante na escolha de alternativas para o ambiente. Relatórios e consultas pré-estruturadas podem satisfazer o usuário final e geram pouca demanda sobre o SGBD e sobre o ambiente servidor. Análises complexas, por sua vez, típicas de ambientes de suporte à decisão, exigem mais de todo o ambiente. Ambientes dinâmicos, com necessidades em constante mudança, são melhor atendidos por uma arquitetura simples e de fácil alteração (por exemplo, com uma estrutura altamente normalizada), em vez de uma estrutura mais complexa que necessite de reconstrução a cada mudança (por

exemplo, estrutura multidimensional). A freqüência da necessidade de atualização também é determinante: grandes volumes de dados que são atualizados em intervalos regulares favorecem uma arquitetura centralizada.

2.11 Estratégia Evolucionária

Data Warehouse, em geral, são projetados e carregados passo a passo, seguindo, portanto, uma abordagem *evolucionária*. Os custos de uma implementação "por inteiro", em termos de recursos consumidos e impacto no ambiente operacional da empresa, justificam esta estratégia. Muitas empresas iniciam o processo a partir de uma área específica, normalmente carente de informação e cujo trabalho é relevante para os negócios da empresa. Criam os chamados *Data Marts* (um *Data Warehouse* Departamental) para, depois, evoluir aos poucos, seguindo uma estratégia "botton-up" ou assunto-por-assunto.

Data Marts também podem ser criados como subconjuntos de um *Data Warehouse* maior com vista à autonomia, melhor desempenho e simplicidade de compreensão.

A alternativa é selecionar um grupo de usuários, prover ferramentas adequadas, construir um protótipo do *Data Warehouse*, deixar que os usuários experimentem-no com pequenas amostras de dados. Somente após a concordância do grupo quanto aos requisitos e funcionamento, é que o *Data Warehouse* será, de fato, carregado com dados dos sistemas operacionais na empresa e dados externos.

2.12 Etapas do Desenvolvimento de um *Data Warehouse*

Na verdade, é difícil apontar, no momento, uma metodologia consolidada e amplamente aceita para o desenvolvimento do *Data Warehouse*. O que se vê na literatura e nas histórias de sucesso de implementações em empresas são propostas no sentido de construir um modelo dimensional a partir do modelo de dados corporativo ou departamental (base dos bancos de dados operacionais da empresa), de forma incremental.

Inmon (1992) salienta que, de fato, um *Data Warehouse* é construído de uma maneira 'heurística', confirmando a estratégia evolucionária.

De qualquer forma, a metodologia a ser adotada é ainda bastante dependente da abordagem escolhida em termos de ambiente, distribuição, etc.

Segundo Kimball (1996), desenvolver um *Data Warehouse* é uma questão de "casar" as necessidades dos seus usuários com a realidade dos dados disponíveis. O autor aponta um conjunto de nove pontos fundamentais no projeto da estrutura de *Data Warehouse*. São os seguintes os chamados pontos de decisão, que constituem definições a serem feitas e correspondem, de fato, a etapas do projeto:

- os processos, e por conseqüência, a identidade das tabelas de fatos;
- a granularidade de cada tabela de fatos;
- as dimensões de cada tabela de fatos;
- os fatos, incluindo fatos pré-calculados;
- os atributos das dimensões;
- o acompanhamento das mudanças graduais em dimensões;
- as agregações, dimensões heterogêneas, minidimensões e outras decisões de projeto físico;
- a duração histórica do banco de dados;
- a urgência com que se dá a extração e carga para o *Data Warehouse.*

Kimball (1998a) recomenda que estas definições sejam feitas na ordem acima descrita.. Esta metodologia segue a linha *top-down*, pois começa identificando os grandes processos da empresa.

Como exemplo, temos os processos de uma empresa revendedora de produtos: planos de estoque, ordens de compra, inventário, pedidos de clientes, expedição de pedidos, créditos, etc. Quando os processos estiverem identificados, cria-se uma ou mais tabelas de fatos a partir de cada um deles. Neste ponto, é necessário decidir o fato individual naquela tabela (esta é a granularidade da tabela). O próximo passo é definir as dimensões e suas granularidades. Neste exemplo, considera-se a tabela de fatos *vendas acumuladas do produto*. Uma vez determinada a granularidade, as dimensões e suas respectivas granularidades podem ser identificadas. Assim, as dimensões tempo, produto e vendedor são criadas, além de outras dimensões descritivas como local de expedição, local de recebimento, modo de envio. A adição destas dimensões descritivas não altera o número de instâncias na tabela de fatos. A escolha das dimensões é o ponto chave no projeto. O passo seguinte consiste em detalhar todos as medidas que constarão da tabela de fatos e, finalmente, completar as tabelas de dimensões. Neste ponto, a estrutura do projeto lógico está completa.

Passa-se, então, a trabalhar questões relativas ao projeto físico, avaliando mudanças graduais em dimensões e discutindo a inclusão de agregações, minidimensões e dimensões heterogêneas.

2.13 Tipos e as Necessidades dos Usuários de um *Data Warehouse*

💻 **Executivo:**

- Visão abrangente do negócio;
- necessita de informações básicas sobre o andamento da empresa;
- suas necessidades, na maioria das vezes, são conhecidas antes de uma pesquisa;

- pesquisa detalhes sobre itens que não estão de acordo com sua expectativa;
- é usuário de Internet;
- necessita de um caminho rápido para identificar itens de interesse (previstos ou não previstos);
- suporte deve ser específico.

💻 **Gerente:**

- Possui necessidades bem definidas que raramente mudam - sabe o que quer antes de executar uma pesquisa;
- espera bons tempos de resposta;
- submete, quase sempre, as mesmas queries caracterizadas como simples e que retornam um pequeno número de ocorrências;
- interessa-se por um histórico recente;
- cria os relatórios básicos do negócio.

💻 **Operador:**

- Utiliza dados recentes, detalhados, do dia-a-dia, para executar suas análises táticas;
- normalmente executa queries padronizados, estruturados, e necessita de respostas imediatas;
- raramente faz uso de dados históricos;
- requer uma interface simples e fácil de usar;
- pode não ser uma pessoa do grupo, mas uma aplicação.

💻 **Minerador:**

- Procura metodicamente "preciosidades" nos dados da empresa, necessitando das mesmas nos seus maiores detalhes;
- aprova ou não as hipóteses dos exploradores ou suas próprias e aplica uma ferramenta quando não há hipóteses;

- categoriza clientes de acordo com seus comportamentos, possibilitando oferecer serviços com base em suas necessidades;
- desenvolve atividades de classificar, estimar, predizer, agrupar por afinidade, segmentar e descrever;
- utiliza várias técnicas: árvores de decisão, redes neurais, algoritmos genéticos etc.

💻 **Explorador:**

- Usuário fora do padrão - não sabe o que quer antes de realizar uma pesquisa;
- opera com a intuição e observação;
- procura por padrões e relacionamentos, criando hipóteses. Em alguns casos, seus *insights* são muito valiosos, em muitos outros são simplesmente "miragens";
- as análises e procedimentos que utiliza não são estruturados, são heurísticos e tendem a envolver grande volume de dados, incluindo histórico;
- o tempo de resposta pode ser longo.

2.14 Características Básicas de um Data Warehouse

Gray (1998), apresenta um resumo das principais características básicas do *Data Warehouse*, devidamente discriminados na Tabela 1. A maior parte das características apresentadas a seguir também consta em Inmon (1997) e em Campos (1998).

TABELA 1 – CARACTERÍSTICAS DE UM *DATA WAREHOUSE*

Características	*Descrição resumida*
Orientado por assunto/tema/áreas de negócios do empreendimento	Os dados são organizados da forma como os usuários se referem a eles.
Integrado	As inconsistências são removidas em nomenclatura e conflito de informação, isto é os dados são limpos. Os dados fontes de sistemas OLTP são modificados e convertidos para um estado uniforme de modo a permitir a carga no *Data Warehouse.*
Não volátil	Os dados são somente de leitura, não podendo ser atualizados pelos usuários.
Variável em relação ao tempo	Dados armazenados são históricos. O elemento tempo é fundamental.
Sumarizados	Dados operacionais são agregados, quando necessário, em um formato que permita o suporte à decisão.
Grandes volumes – Granularidade	Mantendo dados históricos implica em uma maior quantidade de detalhes e necessidade de espaço de armazenamento. É o nível de detalhes dentro do banco de dados do *Data Warehouse*. Quanto menor a granularidade, maior o nível de detalhes e conseqüentemente, maior o volume de dados armazenados.
Não normalizados	Dados podem ser redundantes
Metadados	Dados a cerca de dados
Entrada	Dados oriundos de fontes internas (sistemas legados) e fontes externas, de acordo com as necessidades.

FONTE: Gray (1998)

☛ **Orientado a um assunto ou tema** ☚ De acordo com Gray (1998), em um *Data Warehouse*, os dados são organizados em torno dos principais assuntos ou temas também chamados de processos de negócio de um empreendimento, em lugar de sistemas operacionais que realizam transações individuais. Para Kimball (1998b), um processo de negócio é uma atividade importante na organização (e.g. inventário de estoque, administração de contas e vendas), normalmente suprido por uma ou mais fontes de dados do empreendimento.

☛ **Integrado** ☚ Poe (1998) considera que a integração de dados é o processo pelo qual as características dos dados-fonte são modificadas para possibilitar a sua carga no *Data Warehouse*, sendo tipicamente realizada quando os dados são extraídos do sistema operacional. Constitui-se de um conjunto de atividades que possibilitam a integração de diferentes tipos de dados, a modificação de códigos e a reconciliação da definição dos dados. De acordo com Gray (1998), a integração refere-se ao conjunto de processos e atividades necessárias para realizar a consistência dos dados, antes destes serem introduzidos no *Data Warehouse*, independentemente da origem dos dados. Chama essa atividade de "limpeza de dados.

Os dados que populam um *Data Warehouse* normalmente são originários de vários sistemas operacionais e, muitas vezes, de fontes de dados externas. Cada sistema operacional possui características específicas como tipos de dados, convenções, atributos de nomes, unidades de medidas, dentre outras. Exemplos:

- os dados provenientes de variados sistemas fontes poderão ter notações diferentes para o gênero masculino e feminino, sendo comum valores como "M" ou "F", "m" ou "f", "1" ou "0"; em outras situações pode-se representar uma informação no banco de dados como "sim" ou "não" e "s" ou "n"; da mesma forma sistemas legados poderiam registrar padrões monetários em formatos diferentes como "R$ 200" ou somente "$200". No *Data Warehouse* esses valores serão padronizados em um formato único. (INMON, 1997); (CAMPOS, 1998).
- Sistemas fontes diferentes podem registrar o nome de um cliente como sendo "Pedro da Silva" ou "Pepe da Silva", ou, ainda de uma empresa como "IBM" ou "I.B.M.". Apesar serem a mesma pessoa e empresa, respectivamente, no *Data Warehouse* são tratados como duas entidades completamente distintas. Esse problema de duplicação, freqüentemente, é referenciado na literatura através dos nomes *deduplicating*, *customer matching* e *householding* (KIMBALL, 1998a). Durante o processo de integração, os dados são convertidos para um estado uniforme. Por exemplo, em um *Data Warehouse* que possua uma dimensão com o atributo "unidade de medida",

> independente da fonte de dados, os dados serão armazenados unicamente na unidade de "metros" (INMON, 1997).

Para Gray (1998), a integração de dados, além da limpeza dos dados, deveria ser composta pelas ações de validação e correta agregação. Por validação, entendam-se os passos necessários a serem seguidos para garantir que os dados existentes na base de dados analítica sejam corretos. Já a correta agregação significa sumarizar (resumir) alguns dados a partir de outros, de acordo com as necessidades dos usuários.

Conforme Poe (1998), podem surgir diversos problemas durante o processo de integração devido à natureza dos elementos de dados que podem ser sinônimos, homônimos e análogos. *Sinônimos* são elementos de dados que têm diferentes nomes, mas o mesmo significado ou representam o mesmo fato do negócio. *Homônimos* são elementos de dados que têm o mesmo nome, mas representam diferentes fatos do negócio. *Análogos* são os elementos de dados que têm significados equivalentes, assemelhando-se a dados sinônimos, mas possuem diferenças sutis que são relevantes para o entendimento do negócio. Caso essas diferenças não sejam entendidas e identificadas no processo de análise e modelagem, os dados-fonte poderão ser integrados incorretamente.

☛ **Não volátil** ☚ Nos sistemas operacionais, continuamente são realizadas operações básicas sobre registros de dados pelos usuários, tais como consulta, inserção, atualização e deleção. Esse conjunto de operações possíveis sobre os bancos de dados operacionais lhes confere a característica de serem eminentemente de leitura e escrita.

No ambiente de *Data Warehouse*, os dados depois de serem integrados, são carregados e armazenados no banco de dados analítico, ficando disponíveis para os usuários realizarem apenas consultas e geração de relatórios que permitam a tomada de decisão. Como os usuários não atualizam dados, os bancos de dados analíticos são também chamados de "somente leitura", o que possibilita a existência de grande volume de dados históricos, uma de suas principais características.

☛ **Variável em relação ao tempo** ☚ Segundo Inmon (1997) e Gray (1998), os *Data Warehouse* geralmente armazenam informações por um período válido de tempo de 5 a 10 anos, enquanto os sistemas operacionais armazenam dados históricos limitados a um horizonte de tempo de 60 a 90 dias.

Os dados operacionais devem sempre apresentar valores atualizados e precisos no momento em que forem acessados pelo usuário. Já os *Data Warehouse* armazenam grande quantidade de dados históricos formados por um conjunto sofisticado de instantâneos de tempo, capturados em momento determinado.

No *Data Warehouse*, sempre haverá uma tabela-dimensão ou fato cuja estrutura registrará especificamente o elemento tempo (e.g. hora, dia, semana, mês, ano, etc.). Nos sistemas operacionais, essa é uma característica opcional.

☛ **Agregação** ☚ Poe (1998) denomina agregação **ao** processo de acumulação de dados das tabelas-fato ao longo de atributos predefinidos. Para Kimball (1998a), são registros sumarizados que estão logicamente redundantes com os dados básicos que já estão no *Data Warehouse*, e são utilizados para aumentar enormemente o desempenho das consultas. Segundo Inmon (1997), um registro de agregação é criado pelo agrupamento de diversos registros detalhados.

Conforme Poe (1998), dentro do contexto de um projeto de banco de dados analítico, deve-se tomar a decisão sobre a criação de agregados durante o processo de integração e carga dos dados pré-calculados dento do *Data Warehouse*.

Para Inmon (1997), a agregação de dados operacionais em um único registro pode assumir várias formas, como por exemplo:

- os valores provenientes dos dados operacionais podem ser resumidos, totalizados ou processados para se obter o ponto mais alto, mais baixo, média, etc;
- dados de predeterminados tipos, que estejam dentro de limites, podem ser medidos;
- dados válidos em relação a um determinado momento podem ser dispostos em um bloco.

De acordo dom Kimball (1998a), normalmente, a maioria dos *Data Warehouse* contem agregados pré-calculados , cujos objetivos básicos são:

- melhorar o tempo de resposta de consultas para o usuário final, uma vez que os dados são organizados de forma compacta e prática;
- reduzir o tempo de processamento e ciclos de máquina, uma vez que as tabelas agregadas possuem, normalmente, uma quantidade significativamente menor de dados;
- reduzir o espaço de armazenamento, tendo em vista o agrupamento de vários dados básicos em um único registro.

Ainda de acordo com o autor acima referenciado, os dados agregados devem ser armazenados em suas próprias tabela-fato, separados dos dados de nível básico. Cada agregação distinta deve ocupar sua própria e única tabela-fato. As tabelas-dimensão que se ligam à tabela-fato do agregado devem ser uma versão reduzida das tabelas associadas à tabela-fato base.

Para construir uma agregação, deve-se levar em conta as necessidades mais comuns dos usuários por dados sumarizados e considerar a distribuição estatística das informações.

O processo de agregação no *Data Warehouse* tem a desvantagem de poder vir a reduzir a capacidade ou funcionalidade deste, ocasionando a perda de detalhes. O grande trabalho do projetista é garantir que os detalhes perdidos não sejam de extrema importância. Para isso, segundo Inmon (1997), pode-se utilizar dois métodos:

- o *primeiro* consiste em criar os agregados iterativamente com o usuário. No início, deve-se disponibilizar uma agregação mais ampla que servirá de base para as iterações seguintes, até se chegar ao nível desejado.
- O *segundo* método consiste em garantir que nenhum detalhe importante será perdido durante a construção do processo de agregação, o que pode ser feito pela criação de níveis duais de granularidade, conforme o exposto na seção 3.1.6 desse capítulo, garantindo-se, assim, que, independente do tempo e custo envolvidos, todo e qualquer dado podem ser recuperado

☛ **Granularidade** ☚ Em Kimball (1998a), Kimball (1998b), Inmon (1997), Poe (1998) e Gray (1998) encontramos a definição de granularidade como sendo o nível de detalhes dentro do banco de dados do *Data Warehouse*. É uma das principais questões a serem levadas em consideração em um projeto de um *Data Warehouse*. Quanto maior for o nível de detalhes que se deseja armazenar, menor será a granularidade. A definição do nível de detalhes ou o resumo dos dados refletirá diretamente o volume de dados armazenado, o tipo de informações que poderão ser obtidas e o esforço computacional necessário para a obtenção das informações desejadas por ocasião da realização de consultas.

Referentes a esta questão, podem ocorrer duas situações extremas:

1ª) nível de granularidade muito alto: Nesse caso, será possível obter considerável economia de espaço de armazenamento no banco de dados analítico. Para resolver esse problema, existem várias técnicas, como por exemplo, a utilização de níveis de agregação e sumarização. Entretanto, como desvantagem, haverá uma redução drástica na capacidade de atender a consultas, uma vez que não se tem acesso aos dados de mais baixo nível, mas somente aos sumarizados.

2ª) Nível de granularidade muito baixo: Nesse caso, será possível responder praticamente a qualquer consulta, entretanto, a desvantagem está no fato da necessidade de grande espaço de armazenamento no banco de dados do *Data Warehouse*, que, dependendo das limitações tecnológicas da organização, poderá ser um fator crítico de sucesso no empreendimento.

A primeira idéia para resolver este problema é realizar um balanceamento entre o espaço de armazenamento e a capacidade de atender a consultas, o que se traduz basicamente na definição do nível da granularidade a ser adotado após cuidadoso estudo realizado no início do projeto e construção do *Data Warehouse*.

Uma outra forma de minimizar esse problema é criar níveis duais de granularidade. Na estrutura, todos os dados provenientes dos sistemas operacionais e fontes externas são inicialmente carregados no nível de detalhes corrente que se constitui o banco de dados analítico do *Data Warehouse*. Desse nível, os dados são resumidos e armazenados no nível de dados levemente resumidos, onde haverá um volume de dados significativamente menor e que visa atender a pré-consultas padronizadas disponibilizadas para usuários finais. Os dados altamente sumarizados são

obtidos a partir do resumo de dados levemente resumidos e têm a finalidade de atender a gerentes e administradores dos mais altos escalões, onde as informações são disponibilizadas de forma compacta e facilmente acessível. Por outro lado, as informações provenientes do sistema operacional também são armazenadas no nível de detalhes antigos quando atingirem uma idade limite, normalmente especificada em projeto e, em decorrência, são movidos para dispositivos de armazenamento de massa (e.g. fita, disco ótico, compact disc, DVD, etc.), tipicamente mais baratos quando comparados a outras tecnologias, como por exemplo, discos rígidos. A realização dessas duas atividades cria dois níveis de granularidade.

Ainda segundo Inmon (1997), 95% ou mais do processamento em sistemas de SAD's é feito sobre o nível de dados levemente resumidos e altamente resumidos, que é mais compacto, de fácil acesso e de menor tempo de resposta a consultas. Apenas 5 % ou menos das consultas normalmente são feitas sobre o nível de dados antigos, no qual o acesso é mais complexo, caro e de elevado tempo de apresentação dos dados consultados.

Dessa forma, o nível "dual" de granularidade possibilita alta flexibilidade ao *Data Warehouse*, sendo possível a obtenção de informações tanto de forma eficiente e com tempo de resposta adequado, quanto através de atividades complexas e custosas. Entretanto, o mais importante é que, em qualquer uma destas situações, obter-se-á sempre as informações desejadas já que os dados existem.

Inmon (1997) e Gray (1998) sugerem a necessidade do estabelecimento de um equilíbrio na escolha dos níveis adequados de granularidade, de modo que haja uma solução de compromisso entre o volume de dados e o nível de detalhes que podem ser consultados.

Poe (1998) afirma que é uma pressuposição incorreta de muitos projetos de *Data Warehouse* armazenar todos os detalhes de dados com pouco armazenamento de agregados

Já Kimball *et al* (1998a) afirmam que há uma nova e crescente justificativa para se armazenar uma extrema granularidade de dados em um *Data Warehouse*, que é possibilitar a realização de mineração de dados (*Data Mining*) e o entendimento do comportamento de clientes. Mineração de dados é muito pouco eficiente quando realizada em dados agregados.

Analisando as diversas abordagens sobre granularidade a colocação de Kimball (1998a) que defende a adoção do mais baixo nível de granularidade possível em um *Data Warehouse* parece ser mais adequada, pois, dessa forma, é possível capturar todos os instantâneos de tempo e submetê-los, por exemplo, à realização de mineração de dados.

Tal ponto de vista é reforçado por Gardner (1998) ao afirmar que as necessidades que os usuários têm hoje não serão obrigatoriamente as mesmas de amanhã. Essa afirmação cresce ainda mais em importância ao se observar o período de tempo de validade de um *Data Warehouse:* 5 a 10 anos. Dessa forma, é fundamental que uma solução de *Data Warehouse* seja flexível e escalável, de modo a atender qualquer exigência do usuário, independente de envolver dias, meses ou anos, o que somente será possível se os dados existirem.

☛ **Metadados** ☚ Apesar de vários autores afirmarem que metadados são dados sobre dados, tais como Inmon (1997), Poe (1998), Gray (1998), Sem (1998) e Gardner (1998), não há um consenso sobre uma definição precisa.

Para Poe (1998), os metadados provêm informações sobre a estrutura de dados e as relações entre estas dentro ou entre bancos de dados. No ambiente de *Data Warehouse*, há dois tipos de metadados. O primeiro é o de integração de dados que associa os do sistema-fonte aos do *Data Warehouse*, podendo incluir o nome original do sistema-fonte, tipos de dados e outras informações. O segundo tipo de metadado é o de transformação, também podendo ser chamado de metadados Sistema de Apoio a Decisão (SAD), cuja finalidade é mapear os dados do *Data Warehouse* para o usuário final através de ferramentas de *Front End.*.

Em Gray (1998), encontramos a definição de metadados como informações mantidas acerca do *Warehouse* em lugar de informações providas pelo *Warehouse*, sendo essenciais tanto para o *staff* quanto aos usuários do *Data Warehouse*. Cada um desses grupos requer diferentes informações. Para o *staff*, os metadados incluem: um diretório de quais dados estão e onde estão no *Data Warehouse*; um guia que mapeia a forma como os dados provenientes das fontes de dados são carregados no *Data Warehouse* e, finalmente, regras usadas para sumarização. Por sua vez, para os usuários, os metadados incluem: os termos do negócio usados para descrever os dados, os nomes

técnicos correspondentes ao termos dos negócios que podem ser usados para acessar os dados; as fontes de dados, as regras usadas para derivá-las e onde foram criadas.

Para Sen (1998), os metadados são uma abstração de dados que é um instrumental na definição de dados brutos. A necessidade por metadados é crítica em um projeto de *Data Warehouse*. Para ele, existem três tipos de metadados: o de *nível operacional*, que define a estrutura de dados nos bancos de dados operacionais; o de *nível warehouse*, que define a forma como os dados transformados são interpretados e o *nível de negócio*, onde os metadados do *Data Warehouse* são mapeados para os conceitos do negócio.

Gardner (1998) afirma que, em bancos de dados relacionais, metadados são a representação de objetos definidos no banco de dados, especificamente as definições de tabelas, colunas, banco de dados, visões e qualquer outro objeto. Em *Data Warehouse*, metadados referem-se a qualquer coisa que defina objetos *Data Warehouse*, tal como tabelas, colunas, consultas, relatórios, regras de negócio ou um algoritmo de transformação. Os metadados deveriam gerenciar rigorosamente tudo, desde o desenvolvimento de programas que extraem dados dos sistemas operacionais-fonte até a coleção de dados que é introduzida no *Data Warehouse*.

Kimball *et al* (1998a) diz, por sua vez, que "metadados são todas as informações do ambiente do *Data Warehouse* que não são seus próprios dados".

2.15 Componentes de um Sistema de um *Data Warehouse*

A implementação de um sistema de *Data Warehouse* faz uso constante de ferramentas que realizam tarefas definidas. Temos a seguir uma lista dos componentes destes sistemas:

- fontes de dados do *Data Warehouse*;
- um conjunto de estruturas de dados analíticos (o *Data Warehouse*);
- um sistema de gerência de banco de dados (SGBD) configurado especialmente para atender aos requisitos analíticos dos sistemas de *Data Warehouse*s;

- um componente *back end*: fazem a extração, limpeza, transformação, integração e carga dos dados das diferentes origens (fontes);
- um componente *front end*: disponibiliza aos usuários finais o acesso aos dados do *Data Warehouse*;
- um repositório para armazenar e gerenciar os metadados.

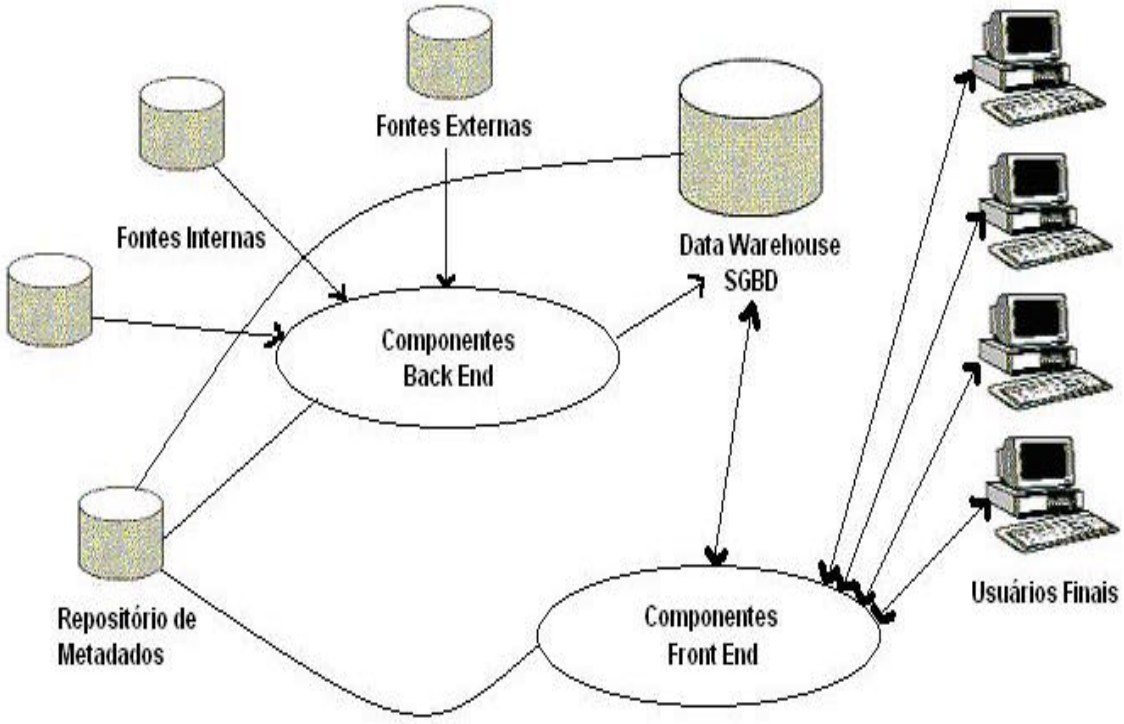

FIGURA 1 – COMPONENTES DE UM SISTEMA *DATA WAREHOUSE* (RELACIONAMENTO ENTRE OS COMPONENTES)

2.16 Arquitetura Genérica de um *Data Warehouse*

A seguir, é descrita uma arquitetura genérica proposta por Kimball (1998a) e ilustrada na FIGURA 2, que procura apenas sistematizar papéis no ambiente de *Data Warehouse*, permitindo que as diferentes abordagens encontradas no mercado, atualmente, possam ser adaptadas a ela. Deve-se considerar que esta arquitetura tem o objetivo de representar a funcionalidade do *Data Warehouse*, sendo que várias camadas propostas podem ser atendidas por um único componente de *software*.

Esta arquitetura é composta pela camada dos dados operacionais e outras fontes de dados que são acessados pela camada de acesso aos dados. As camadas de gerenciamento de processos, transporte e *Data Warehouse* formam o centro da arquitetura e são as responsáveis por manter e distribuir os dados. A camada de acesso à informação é formada por ferramentas que possibilitam aos usuários extrair informações do *Data Warehouse.* Todas as camadas desta arquitetura interagem com o dicionário de dados (metadados) e com o gerenciador de processos.

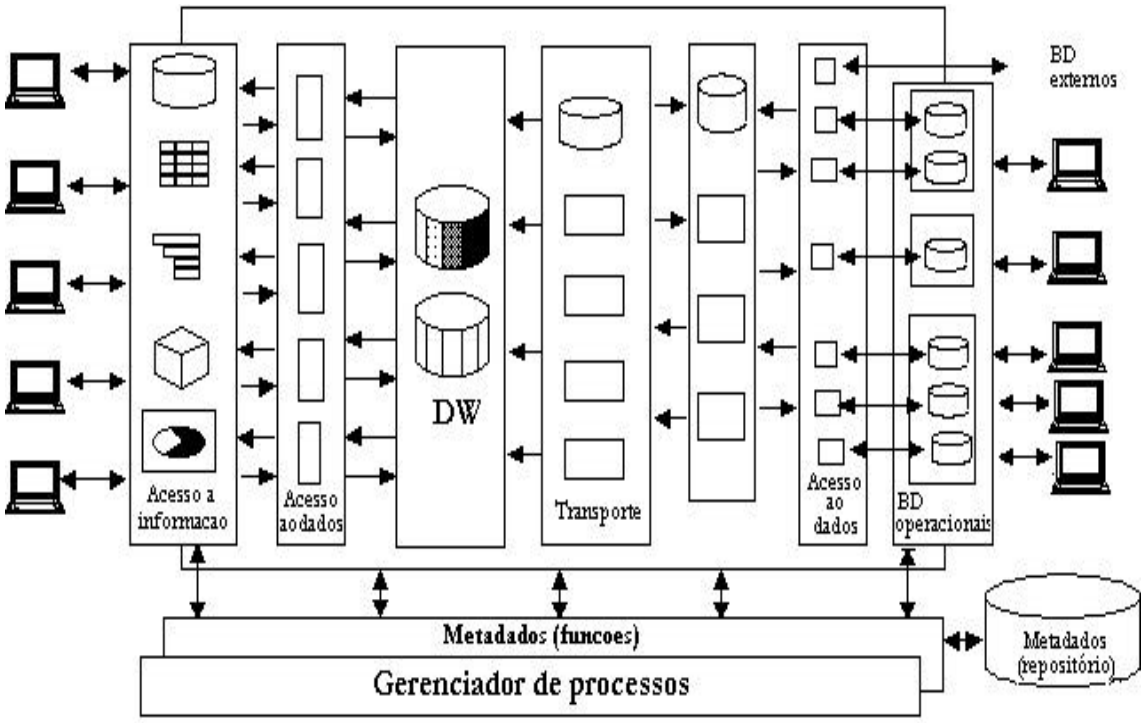

FIGURA 2 - ARQUITETURA GENÉRICA DE UM DATA WAREHOUSE

☛ **Camadas de bancos de dados operacionais e fontes externas** ☚ São compostas pelos dados dos sistemas operacionais das empresas e informações provenientes de fontes externas de onde é realizada a extração de dados para compor o *Data Warehouse.*

☛ **Camada de acesso à informação** ☚ Envolve *o hardware* e o *software* utilizados para obtenção de relatórios, planilhas, gráficos e consultas. É nesta camada que os usuários finais interagem com o *Data Warehouse,* utilizando ferramentas de

manipulação, análise e apresentação dos dados, incluindo-se as ferramentas de *Data Mining* e visualização.

☛ **Camada de acesso aos dados** ☚ Esta camada faz a ligação entre as ferramentas de acesso à informação e os bancos de dados operacionais. Comunica-se com diferentes sistemas de bancos de dados, sistemas de arquivos e fontes sob diferentes protocolos de comunicação, o que é chamado de *acesso universal de dados*.

☛ **Camada de metadados (dicionário de dados)** ☚ Metadados são as informações que descrevem os dados utilizados pela empresa. Envolvem informações como descrições de registros, comandos de criação de tabelas, diagramas Entidade/Relacionamentos (E-R), dados de um dicionário de dados etc. É necessário que exista uma grande variedade de metadados no ambiente de *Data Warehouse* para que ele mantenha sua funcionalidade e os usuários não precisem se preocupar onde residem os dados ou a forma com que estão armazenados.

☛ **Camada de gerenciamento de processos** ☚ É a camada responsável pelo gerenciamento dos processos que contribuem para manter o *Data Warehouse* atualizado e consistente. Está envolvida com o controle das várias tarefas que devem ser realizadas para construir e manter as informações do dicionário de dados e do *Data Warehouse.*

☛ **Camada de transporte** ☚ Gerencia o transporte de informações pelo ambiente de rede. Inclui a coleta de mensagens e transações e se encarrega de entregá-las em locais e tempos determinados. Também é usada para isolar aplicações operacionais ou informacionais, do formato real dos dados nas duas extremidades.

☛ **Camada do *Data Warehouse*** ☚ É o *Data Warehouse* propriamente dito, corresponde aos dados utilizados para obter informações. Às vezes, o *Data Warehouse* pode ser simplesmente uma visão lógica ou virtual dos dados, podendo não envolver o armazenamento dos mesmos ou de dados operacionais e externos para facilitar seu acesso e manuseio.

2.17 Modelagem de Dados em um *Data Warehouse*

Modelar significa, simplificar, converter informações em estrutura de campos que deverão compor o banco de dados, tomando como base o segmento de atuação da empresa e o tipo de relacionamento que se pretende estabelecer com os seus clientes. Com estas informações, é desenhado um modelo que permitirá criar e gerenciar campanhas de relacionamento.

A modelagem correta é fundamental para o sucesso do processo, pois uma vez definida a estrutura da base, torna-se extremamente complexo e custoso alterá-la. A especificação de requisitos do ambiente de suporte à decisão associada a um *Data Warehouse* é fundamentalmente diferente da especificação de requisitos dos sistemas que sustentam os processos usuais do ambiente operacional de uma empresa. Os requisitos dos sistemas do ambiente operacional são claramente identificáveis a partir das funções a serem executadas pelo sistema. Requisitos de *Sistemas de Suporte à Decisão* (SSD) são, por sua vez, indeterminados. O objetivo por trás de *Data Warehouse* é prover dados com qualidade; os requisitos dependem das necessidades de informação individuais de seus usuários. Ao mesmo tempo, os requisitos dos sistemas do ambiente operacional são relativamente estáveis ao longo do tempo, enquanto que os dos *Sistemas de Suporte a Decisão* (SSD) são instáveis, dependem das variações das necessidades de informações daqueles responsáveis pelas tomadas de decisões dentro da empresa.

No entanto, embora as necessidades por informações específicas mudem com freqüência, os dados associados não mudam. Imaginando-se que os processos de negócio de uma empresa permaneçam relativamente constantes, existe apenas um número finito de objetos e eventos com as quais uma organização está envolvida. Por esta razão, um modelo de dados é uma base sólida para identificar requisitos para um *Data Warehouse*.

Inmon (1992) salienta que é um erro pensar que técnicas de projeto que servem para sistemas convencionais serão adequadas para a construção do *Data Warehouse*. Os requisitos para um *Data Warehouse* não podem ser conhecidos até que ele esteja parcialmente carregado e já em uso.

Outra questão interessante, discutida por Kimball (1996), é a adequação do modelo Entidade-Relacionamento ao tipo de transação de sistemas OLTP. O principal objetivo da modelagem, neste caso, é eliminar, ao máximo, a redundância, de tal forma que uma transação que promova mudanças no estado do banco de dados atue o mais pontualmente possível. Com isso, nas metodologias de projeto usuais, os dados são "fragmentados" por diversas tabelas, o que traz uma considerável complexidade à formulação de uma consulta por um usuário final. Por isso, salienta Kimball, esta abordagem não parece ser a mais adequada para o projeto do *Data Warehouse*, onde estruturas mais simples, com menor grau de normalização devem ser buscadas.

Existe um grande número de enfoques sobre modelagem de dados já desenvolvidos por vários autores. A maioria deles pode ser usada para construir um *Data Warehouse*. Dentre estes modelos apenas o multidimensional será apresentado.

2.17.1 Modelagem Dimensional

Modelagem multidimensional é o nome de uma técnica de projeto lógico freqüentemente usada para *Data Warehouse*, cujo principal objetivo é apresentar o dado em uma arquitetura padrão e intuitiva que permita acessos de alta performance. Cada eixo no espaço multidimensional corresponde a um campo ou coluna de uma tabela relacional e cada ponto, um valor correspondente à interseção desses campos ou colunas. Assim, o valor para o campo “vendas”, correspondente ao mês igual a “março” e “filial 8”, é um ponto com coordenada [março, filial 8]. Neste caso,”mês” e “filial” são duas dimensões e”vendas”, uma medida. Em teoria, quaisquer dados podem ser considerados multidimensionais. Entretanto, o termo normalmente se refere a dados que representam objetos ou eventos que podem ser descritos e, portanto, classificados, por dois ou mais de seus atributos.

Dados multidimensionais podem ser armazenados e representados em estruturas relacionais, sendo necessário, para isso, utilizar formas específicas de modelagem como o modelo “ESTRELA (*Star Schema*)” e o modelo “FLOCO DE NEVE (*Snowflake Schema*)”. Contudo, quaisquer que seja o esquema utilizado, existem basicamente dois tipos de tabelas:

- as *Tabelas-Fato*: um fato usualmente é uma coisa sobre a qual não se conhece antecipadamente; é uma observação da realidade. (KIMBALL, 1998a). A tabelas-fato são centrais e, geralmente, armazenam grande quantidade de dados (de gigabytes a terabytes), dependendo diretamente da granularidade adotada; possuem chaves primárias compostas e contêm as medições numéricas do negócio denominados fatos (GRAY, 1998). Os melhores fatos e os mais úteis são numéricos e caracterizam-se por serem continuamente valorados (diferente a cada medida) e aditivos (os valores modificam-se a cada combinação de atributos das tabelas dimensão). (KIMBALL, 1998b).
- As *Tabelas Dimensão*: os componentes da tabela descrevem as características de uma coisa tangível. As tabelas-dimensão são simétricas em relação à tabela-fato. Normalmente possuem uma chave primária simples e campos denominados atributos; armazenam pequena quantidade de dados, quando comparadas à tabela-fato, e contêm os dados descritivos do negócio (KIMBALL, 1998b). Os melhores atributos são textuais e discretos. Os atributos das tabelas-dimensão são a principal fonte de restrições em consultas SQL e, usualmente, estão presentes em cabeçalhos de linhas no conjunto resposta de consultas realizadas pelos usuários (KIMBALL, 1998a). A cada chave primária da tabela dimensão corresponderá exatamente uma chave estrangeira na tabela-fato, permitindo a ligação entre ambas. Apesar de, muitas vezes, conterem campos numéricos, a diferença em relação aos fatos é que, nas dimensões, estes não variam continuamente a cada nova amostra (são constantes), por exemplo: "número_da_loja", "peso_produto", etc. Segundo KIMBALL (1998a), modelos dimensionais reais no mundo dos negócios contêm entre 4 e 15 dimensões, sendo raros modelos com 2 ou 3 dimensões. Ainda segundo esse autor, modelos com 20 ou mais dimensões devem ser estudados para se verificar as dimensões supérfluas e/ou combiná-las.

Será abordado apenas o esquema “ESTRELA”, que é o mais utilizado para a construção de *Data Warehouse* e possui uma estrutura totalmente diferente do modelo E-R.

2.17.1.1 Esquema Estrela

A FIGURA 3, apresentada, em Kimball (1998a), é um exemplo de modelo dimensional tipicamente no formato estrela tradicional, no qual pode ser observada uma única tabela-fato, no centro, cercada por várias tabelas dimensão:

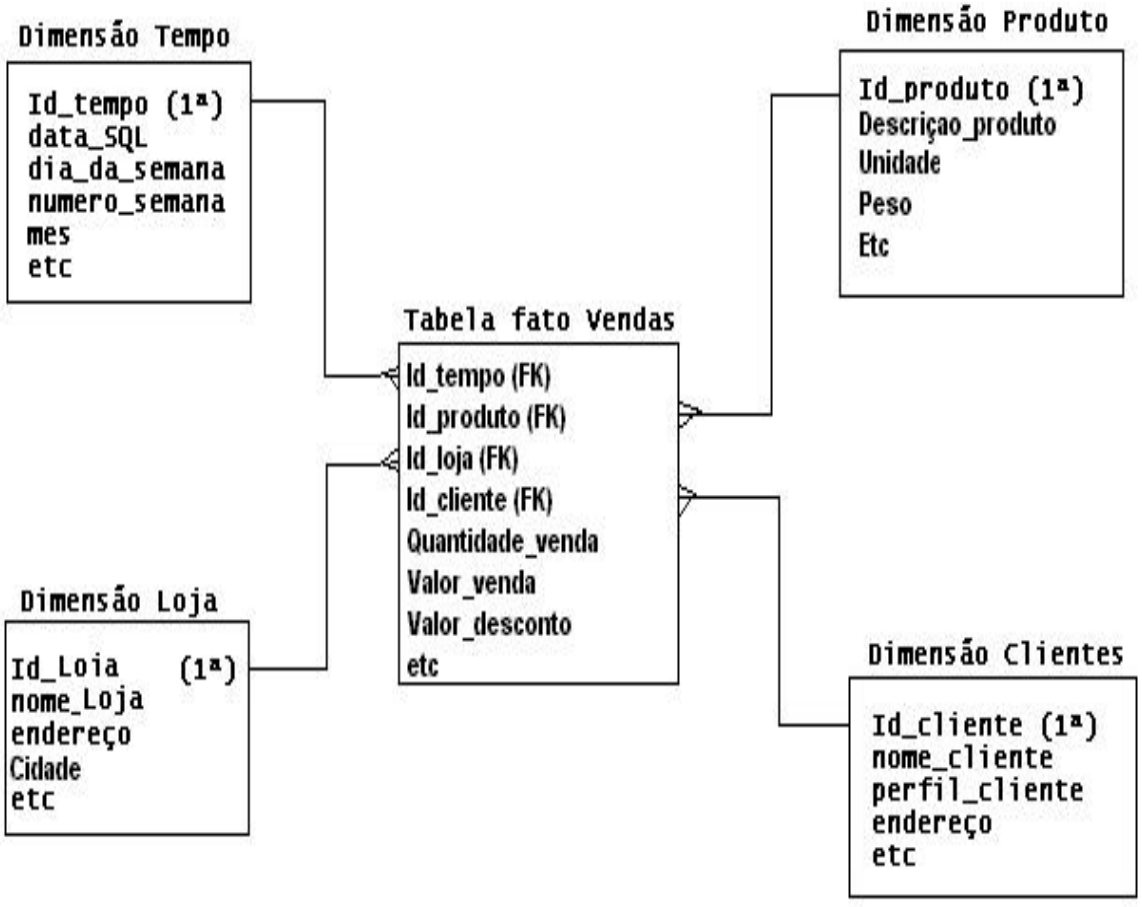

FIGURA 3 – EXEMPLO DE UM ESQUEMA ESTRELA TRADICIONAL

2.17.1.2 Vantagens do Esquema Estrela

Segundo Poe (1988), a utilização de um Esquema Estrela para processamento analítico possibilita alguns benefícios em relação a uma estrutura relacional, dos quais destacam-se os seguintes aspectos:

- o projeto de banco de dados multidimensional provê rápido tempo de resposta;
- permite otimizar o banco de dados para trabalhar com um projeto mais simples de banco de dados melhorando a execução do planejamento;
- permite projetar o banco de dados como o usuário final habitualmente pensa e usa os dados analiticamente;
- simplifica o entendimento e a navegação dos metadados por desenvolvedores e usuários finais;
- possibilita um maior número de ferramentas de acesso aos dados, sendo que alguns produtos disponíveis no mercado exigem o projeto de um Esquema Estrela.

2.17.1.3 Limitações do Esquema Estrela Tradicional

O Esquema Estrela Tradicional normalmente é usado em grande parte dos projetos de bancos de dados analíticos. Existem algumas situações, entretanto, que se deve abdicar de seu uso, como por exemplo, quando uma tabela dimensional possui uma quantidade muito grande de registros e atributos. É imperativo ter em mente que um projeto de banco de dados analítico deve ser conduzido no sentido de obter o máximo de vantagens do ambiente de processamento analítico como, por exemplo, a integração e a qualidade dos dados, usabilidade e desempenho, dentre outras.

Segundo Poe (1998), para atingir esses objetivos, muitas vezes, é necessário adotar soluções alternativas, tais como as variações do Esquema Estrela apresentadas, em lugar de aderir fielmente a um esquema tradicional, por querer adotá-lo sem analisar as implicações decorrentes. É necessário estabelecer um compromisso entre a técnica de

projeto de banco de dados tradicional e o projeto de banco de dados multidimensional de modo a se obter sucesso no desenvolvimento de um *Data Warehouse.*.

2.17.1.4 Variações do Esquema Estrela Tradicional

Poe (1998) apresenta variações no Esquema Estrela Tradicional, como, por exemplo, Esquemas Estrela com múltiplas tabelas-fato, tabelas associativas e tabelas externas.

☛ **Esquema Estrela com múltiplas tabelas-fato** ☚ **A** implementação de um banco de dados mais sofisticado pode requerer o uso de múltiplas tabelas-fato. Em alguns casos, essas tabelas existem porque contêm fatos que não estão relacionados ou por causa da diferente periodicidade de tempo de carga. A FIGURA 4, retirada de POE (1998), apresenta um exemplo de Esquema Estrela com múltiplas tabelas-fato.

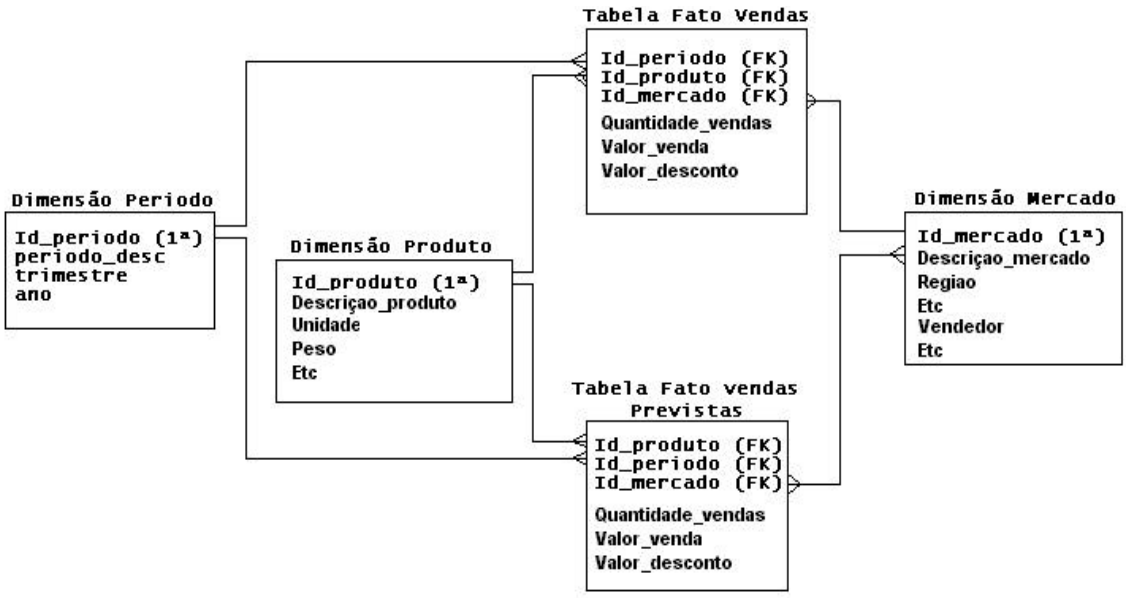

FIGURA 4 - EXEMPLO DE UM ESQUEMA ESTRELA COM MÚLTIPLAS TABELAS-FATO

☛ **Esquema Estrela com tabelas associativas** ☚ Uma tabela-fato também pode ser utilizada para definir relacionamento "muitos-para-muitos" entre certas dimensões de negócios. Este tipo de tabela é tipicamente conhecido como tabela associativa, cujo exemplo pode ser visualizado na FIGURA 5, retirada de Poe (1998). Nesse exemplo, cada produto pertence a um ou mais grupos e cada grupo contêm múltiplos produtos. Um relacionamento "muitos-para-muitos" é modelado pela criação de uma tabela-fato (na FIGURA 5 identificada pela tabela Grupo_Produto), que define todas as possíveis combinações das tabelas-dimensão,produtos e grupos.

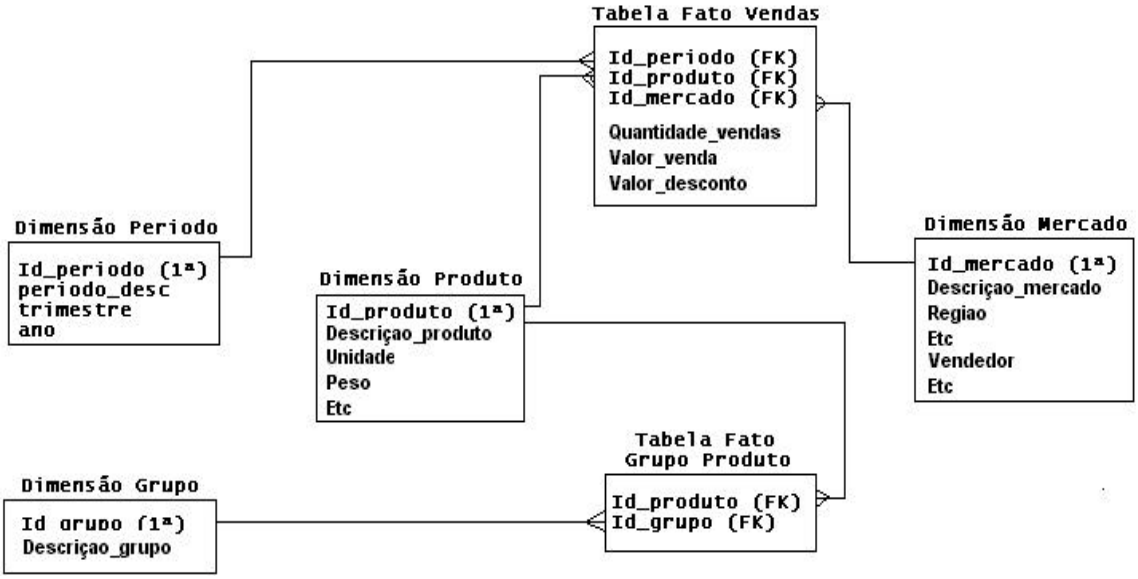

FIGURA 5 - EXEMPLO DE TABELA-FATO COMO TABELA ASSOCIATIVA EM UM ESQUEMA ESTRELA

☛ **Esquema Estrela com tabelas externas** ☚ Nesse esquema, uma ou mais tabelas-dimensão podem conter uma chave estrangeira que referencia a chave primária de outra tabela-dimensão, podendo também ser chamada de tabela-dimensão secundária. Tabelas externas podem formar uma cadeia, sendo possível representar uma hierarquia. A utilização desse esquema equivale à normalização de tabelas, apresentando, em conseqüência, algumas desvantagens tais como: menor compreensão do negócio por

parte do usuário, redução do desempenho na recuperação de informações, dentre outras Poe (1998). A FIGURA 6, retirada de Poe (1998), apresenta um exemplo de Esquema Estrela com tabelas externas. Nessa figura, podem ser observadas a tabela-dimensão Mercado e as suas chaves estrangeiras "Id_região" e "Id_vendedor", a partir da qual se ligam as tabelas externas Região e Vendedor, através de suas respectivas chaves primárias "Id_região" e "Id_vendedor".

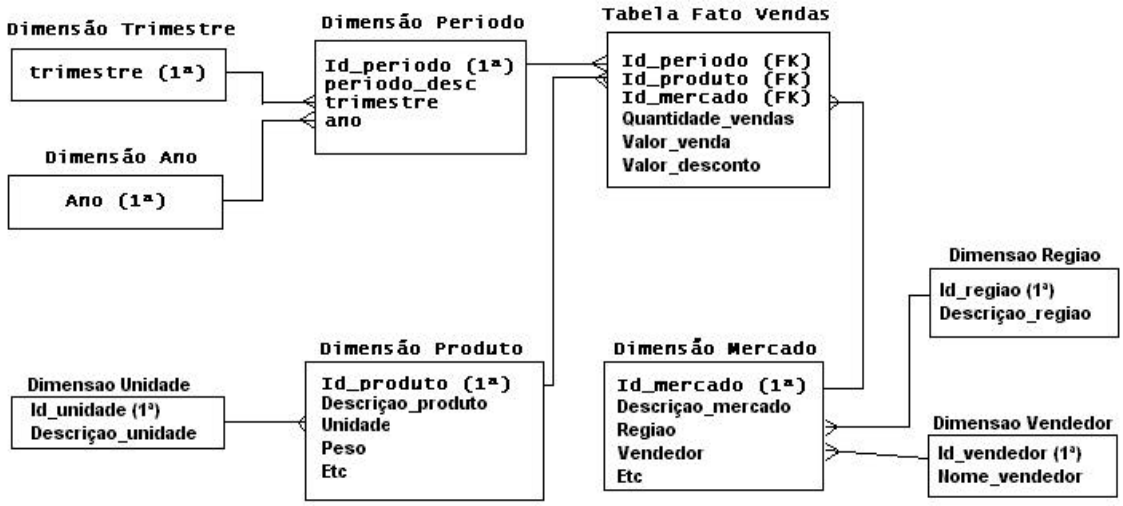

FIGURA 6 – EXEMPLO DE ESQUEMA COM TABELAS EXTERNAS

2.18 Ferramentas

As ferramentas foram avaliadas a partir do enfoque de arquitetura, de modo a verificar a sua aplicabilidade segundo os conceitos de funcionalidades da arquitetura das áreas internas e externas.

2.18.1 Principais Tipos de Ferramentas F*ront End*

Além de permitirem a consulta dos dados, as ferramentas *Front End* devem poder realizar operações com os dados consultados, transformando-os em informações verdadeiramente úteis ao usuário.

Estas ferramentas constituem as chamadas "*aplicações sobre o negócio*", as quais permitem que um usuário especializado possa realizar consulta, também especializada, em um banco de dados conciso e bem estruturado. Desta maneira, as questões em reuniões ou até mesmo no dia-a-dia do trabalho, podem ser respondidas rapidamente e com índice de acerto muito elevado.

A seguir, temos algumas das operações realizadas pelas "*aplicações sobre o negócio*":

- seleção do conjunto de dados necessário à consulta ao *Data Warehouse*;
- cálculo e transformação dos dados antes de sua apresentação ao usuário final;
- apresentação propriamente dita dos resultados das pesquisas e consultas, através de gráficos, tabelas, relatórios, indicação de pontos críticos etc;
- propagação dos resultados a outros usuários, através de impressos, correio eletrônico ou até mesmo, pela rede interna da empresa (intranet) e ainda, se desejável, pela internet.

De acordo com Mor (1998), os tipos de ferramentas *Front End* utilizados para realizar consultas aos bancos de dados analíticos são os seguintes:

TABELA 2 – PRINCIPAIS TIPOS DE FERRAMENTAS *FRONT END*

Tipo de Ferramenta	Questão básica	Exemplo de resposta	Usuário típico e suas necessidades
Consultas e relatórios	"O que ocorreu ?"	Relatórios mensais de vendas	Usuários "intermediários", visão estática de dados
OLAP	"O que aconteceu e por quê ?"	Vendas por produto em cada loja da cidade	Usuários "intermediários", visão multidimensional de dados históricos
EIS	"O que eu preciso saber agora ?"	Quadros de destaque: Narrativas de problemas – chave	Usuários ocupados, informações de alto nível ou resumidas
Data Mining	"O que interessa", "O que pode acontecer"	Modelos de previsão	Usuários altamente qualificados, tendências obscuras entre os dados

FONTE: MOR (1998)

2.18.1.1 Geradores de Relatórios

Os representantes das primeiras gerações de ferramentas desenvolvidas para permitir o acesso aos dados foram os chamados Geradores de Relatórios e de Consultas, os quais atuam fornecendo respostas a consultas do tipo *ad-hoc* (por demanda).

Nestas ferramentas, os usuários devem selecionar as opções dos menus e botões que lhes são apresentados indicando esta ou aquela informação-chave, e o programa verificará a existência da mesma e apresentará apenas os registros que contiverem características que atendam à solicitação desejada.

Desta forma, o usuário fica limitado a consultas simplificadas, utilizando cruzamentos rudimentares de dados. Por outro lado, pode atender inicialmente a usuários com pouca ou nenhuma qualificação, servindo, até mesmo, de motivador para operadores do sistema que desejem obter respostas rápidas aos seus questionamentos sem a necessidade de se aprofundarem no uso das ferramentas. É permitido, também, utilizar fórmulas prontas do tipo "média", "desvio-padrão", e outros cálculos estatísticos tradicionais.

A vantagem deste tipo de ferramenta está no fato de que o usuário não necessita despender horas de aprendizado em SQL, uma vez que o gerador cria estes *queries* de acordo com as seleções nos menus apresentados na tela.

2.18.1.2 Ferramentas OLAP

É o processamento voltado para a análise dos dados originados pelo uso de ferramentas transacionais empregadas no dia-a-dia da empresa para controle de suas atividades. É comum associarmos a tecnologia OLAP à manipulação multidimensional dos dados. Estas estruturas de dados permitem que estes sejam apresentados e analisados sob a ótica do gerente ou do tomador de decisão.

Estas visualizações são possíveis porque o modelo de dados é projetado para contemplar o formato de dimensões, as quais refletem a representação da realidade destes dados sob a ótica de quem irá analisá-los. Neste tipo de ferramenta, o usuário pode solicitar as mais variadas consultas, de acordo com as sumarizações desejadas, ficando a cargo deste usuário a definição de cada uma das visões que melhor se adequar

à resposta esperada, o que demanda maior qualificação em relação à ferramenta "geradores de relatório" citada no item anterior.

As ferramentas OLAP são as representantes da segunda geração das aplicações de acesso a dados. No item anterior, temos os representantes da primeira geração, os quais geram apenas consultas estáticas que não podem ser manipuladas.

Na segunda, os dados são apresentados em planilhas ou gráficos que podem ser facilmente alterados de acordo com a visão desejada. Usando estas ferramentas OLAP, o usuário atento e experiente consegue visualizar uma possibilidade de localizar uma informação mais palpável e interessante que aquela demonstrada na consulta atual, bastando que mude uma ou mais dimensões de posição, visualmente, com o auxílio do mouse. Intuitivamente, este operador pode mudar todo o panorama dos dados sem que para isso seja necessário produzir linhas de código ou executar operações muitas vezes inexistentes nas ferramentas tradicionais. Embora os dados estejam dispostos de uma forma que o modelo multidimensional exija para seu funcionamento, o usuário pode dispô-los da forma que mais lhe aprouver, permitindo que sejam feitas "abstrações" com o mínimo esforço, além de não obrigar o sistema a fazer o acesso às informações através de um único caminho. Apesar desta maleabilidade, podem ser armazenadas determinadas consultas que permitirão ao usuário buscá-las novamente sempre que necessário, agilizando e padronizando, quando desejável, o processo de pesquisa nos dados. Segundo Kimball (1996), as mais freqüentes funções oferecidas pelas ferramentas OLAP são:

- *Tabelas cruzadas*: as tradicionais planilhas eletrônicas. A diferença reside no fato de que os dados são apresentados em planilhas com mais de duas dimensões, normalmente quatro ou mais.
- *Drill-down*: serve para solicitar uma visão mais detalhada em um conjunto de dados, podemos dizer que o usuário "mergulha" nos dados.
- *Roll-up*: consiste na operação inversa ao *drill-down*, ou seja, apresenta os dados cada vez mais agrupados ou sumarizados.

- *Pivoting*: serve para adicionar ou rearranjar as dimensões das tabelas. Para que seja realizada esta operação pode-se utilizar as funções de "arrastar e soltar" com o *mouse*;
- *Slide-dice*: é a função que faz fixar uma informação de dimensão ou reduzir as dimensões de apresentação dos dados.

Existem, ainda, funções que servem de indicadores de problemas, como semáforos e sinalizadores, além dos indicadores de tendências, relatórios de exceção, previsões (projeções, simulações) etc.

À medida que se formulam determinadas consultas ao sistema, nota-se a necessidade de investigar outras características e peculiaridades numa infinidade de indagações que podem ser respondidas num tempo razoavelmente curto, se não instantâneo, a partir da base de dados carregada e devidamente certificada.

Existem diversos tipos de ferramentas OLAP de acordo com o tipo de banco de dados em que são implementadas:

- as ROLAP (*relational*) são assim denominadas porque utilizam as tecnologias dos SGBDR;
- as ferramentas MOLAP (*multidimensional*) baseiam-se em bancos de dados multidimensionais. Estes últimos armazenam as informações em estruturas de dados especializadas para o tipo de modelagem de um sistema de *Data Warehouse.*

Além destes dois tipos de ferramentas, temos as DOLAP (*dice*) que utilizam estruturas de cubos de dados, na maioria das vezes na própria estação de trabalho do usuário, o que permite que o mesmo não necessite estar conectado a um servidor para realizar as consultas ao seu sistema de *Data Warehouse.*

As ferramentas ROLAP estão sendo amplamente utilizadas pelos usuários na atualidade, aproveitando as estruturas de dados relacionais dos bancos de dados tradicionais, porém organizando estes dados de forma multidimensional.

A busca por informações realmente significativas pode se tornar bastante trabalhosa e cansativa à medida que os dados do *Data Warehouse* ficam cada vez mais volumosos. O usuário pode ter dificuldades em encontrar as informações importantes para ele no momento de formular uma nova consulta para análise. Pensando nisto, os projetistas criaram eventos chamados *alertas*, a fim de que, toda vez que um

determinado fato ocorresse, o próprio sistema enviasse uma mensagem ao responsável para que este tomasse as providências necessárias.

Estas características têm relação direta com a freqüência de atualização ou *refresh* do *Data Warehouse*, e as mudanças somente poderão ser verificadas se os dados forem carregados para o *Data Warehouse* a partir da base de dados transacional no intervalo de tempo desejado. Um exemplo disto é a necessidade de carregar os dados diariamente no sistema de *Data Warehouse*, caso os usuários devam ser avisados todos os dias a respeito da movimentação dos estoques da matriz e das filiais. Se os dados não forem atualizados diariamente não será possível verificar qual das lojas está com deficiência ou, por outro lado, não está conseguindo vender os produtos de acordo com o esperado. O uso de ferramentas OLAP deve ser feito por quem possui uma clara compreensão dos modelos de dados da empresa, além de saber quais são as funções analíticas necessárias aos executivos para a tomada de decisão

2.18.1.3 Ferramentas de *Data Mining*

As ferramentas de mineração de dados procuram por informações interessantes e úteis em banco de dados, atuando na área denominada de *Descoberta de Conhecimento em Banco de Dados*. Estas ferramentas tentam encontrar correlações entre os dados armazenados e indicam ao administrador do sistema estes relacionamentos, permitindo que este avalie a descoberta dedutiva apresentada.

Uma das características interessantes apresentadas pelas ferramentas de *Data Mining* é a de realizar previsões do tipo: "o que pode acontecer?" respondendo ao usuário com base na história anterior dos dados armazenados no sistema. Para que estas previsões sejam feitas, o sistema de *Data Mining* utiliza-se de técnicas de Mor (1998), Inteligência Artificial (IA), estatística, *Case-Based Reasoning* (CBR), redes neurais, descoberta por regra, detecção de desvio, programação genética, etc.

Ao contrário das ferramentas OLAP explicadas no item anterior, nas quais o usuário formula perguntas ao sistema, no *Data Mining*, entregamos ao sistema uma grande quantidade de informações para que este nos devolva resultados de pesquisas em busca de tendências ou agrupamentos de dados. Freqüentemente, o usuário que consulta a base procura obter confirmações para algumas de suas suspeitas sem se preocupar em

buscar outras ocorrências de fatos, até mesmo pelo grande número de variáveis envolvidas no processo. Estas ferramentas "descobrem" para o usuário os fatos inusitados e os apresenta para sua apreciação.

O termo *Data Mining* tem gerado alguma confusão quanto ao entendimento de sua denominação e função. Conforme Campos (1997), é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados, usando-as para efetuar decisões cruciais.

Podemos utilizar *Data Mining* com os seguintes objetivos:

- *explanatório*: demonstrar algum evento ou medida auferida;
- *confirmatório*: serve para confirmar uma determinada suspeita;
- *exploratório*: examinar os dados tentando descobrir relacionamentos que não foram previstos pelos usuários e apresentá-los para análise.

Existem diversas técnicas que dão suporte a operações realizadas pelos sistemas de *Data Mining*, as quais citamos a seguir:

- *associações*: são relacionamentos significativos entre itens e dados armazenados. O objetivo deste tipo de operação é encontrar tendências que são detectadas pelo grande número de transações que possam ser usadas para entender e explorar padrões de comportamento dos dados;
- *padrões seqüenciais*: de forma semelhante ao item anterior, podemos identificar eventos que ocorram de tempos em tempos;
- *séries temporais similares*: identificam séries similares que estão armazenadas na base de dados e que variam de forma semelhante ao longo de um período de tempo;
- *classificação e regressão*: utilizam dados armazenados para criar modelos de comportamento de variáveis. É criado um "conjunto de treinamento", que corresponde a um grupo inicial de registros tomados como padrão, classificando-se os demais registros a partir destes padrões. Uma vez definido o padrão de comportamento das variáveis, pode-se determinar quais registros estão fora deste padrão, determinar o distanciamento deste padrão, o que pode confirmar, e de certa forma

explicar, a verificação de algumas anomalias encontradas posteriormente;

- *clusterização*: a informação disponível é segmentada em conjuntos definidos e homogêneos baseados em atributos específicos. Este conceito já é conhecido em diversas áreas, porém em *data mining* foi especializado a fim de permitir sua aplicação em itens não-numéricos. Neste tipo de algoritmo, não é informado ao sistema os tipos de classes existentes, ficando a cargo do computador descobri-las a partir das alternativas encontradas na base de dados;
- *detecção de desvios*: pode-se encontrar anomalias na base de dados através da computação das informações, comparando-as com padrões definidos e regras que não podem ser quebradas.

A *Descoberta de Conhecimento em Bases de Dados* envolve diversos processos, dentre os quais o *Data Mining*, passando desde a seleção dos dados, seu pré-processamento, transformação e interpretação, como mostra a FIGURA 7.

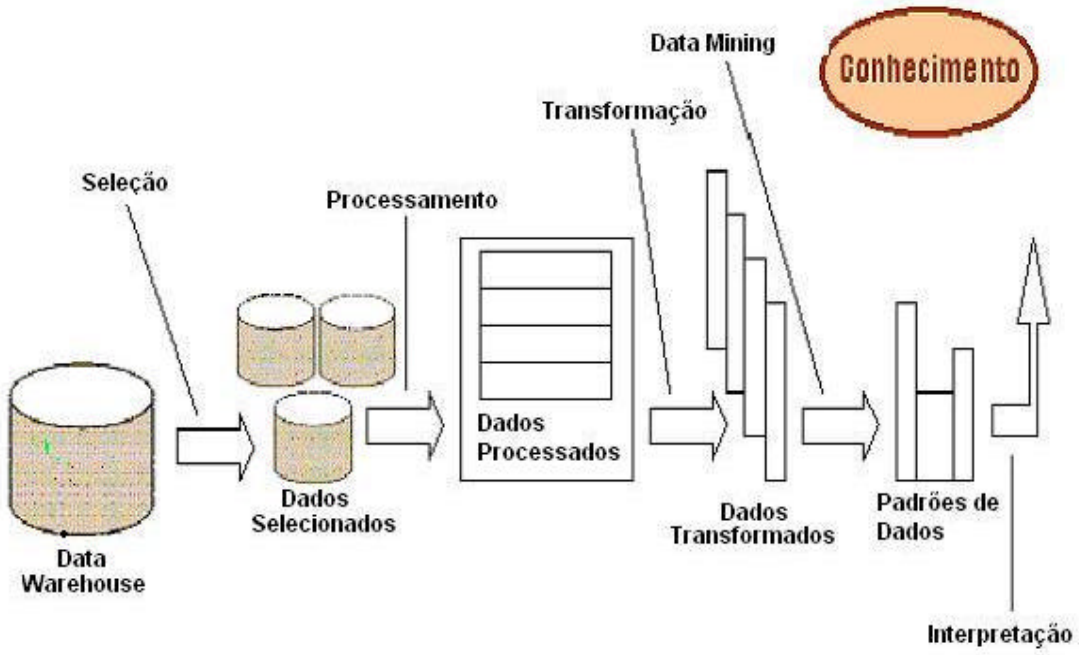

FIGURA 7 - PRINCIPAIS ESTAPAS DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

2.18.1.4 Sistema de Informações para Executivos

Estes sistemas visam satisfazer as necessidades de altos executivos de empresas que não querem e, possivelmente, não devem ser auxiliados por intermediários para chegar às informações desejadas que embasem as suas decisões. Conforme Mor (1998), as características destes sistemas são as seguintes:

- são utilizados por executivos de alto nível, que necessitam de informações corretas e de fácil acesso, a qualquer momento, contendo, inclusive, informações com textos explicativos de auxílio ao usuário;
- servem para acompanhamento e controle em nível estratégico da empresa;
- são ferramentas altamente flexíveis e personalizáveis quanto à interface e ao uso pelos executivos, que comumente não possuem conhecimento na área de modelagem e consulta de baixo nível a banco de dados;
- contêm recursos gráficos de alta qualidade para que as informações possam ser rapidamente analisadas;
- são fáceis de usar para reduzir o tempo de treinamento com os usuários;
- fazem uso intenso de dados de fontes externas à empresa, tais como concorrentes, clientes, fornecedores, governo etc, para comparações.

As informações apresentadas aos executivos podem ser de cinco tipos:

- *narrativas de problemas-chave*: enfatizam o desempenho de um modo geral, os problemas conhecidos e as suas origens na empresa, acompanhadas de textos explicativos e gráficos que facilitam o entendimento das informações demonstradas;
- *quadros de destaque*: são apresentações resumidas elaboradas pelo próprio usuário a partir de suas percepções a respeito dos problemas. Nestes quadros, pode-se grifar o desempenho positivo ou negativo da empresa em relação aos fatores críticos para o sucesso;

- *finanças de alto nível*: a saúde financeira da empresa está representada neste item formatada em números absolutos e taxas de desempenho comparativas;
- *fatores-chave*: são os indicadores de desempenho, que proporcionam medidas específicas dos fatores críticos para o sucesso no nível da empresa. Quando as metas não são atingidas, devemos enfatizar estes fatores na forma de problemas.
- *Relatórios detalhados de responsabilidades sobre os indicadores de desempenho*: monitoram as gerências e outros indivíduos que estão diretamente ligados às atividades críticas para o sucesso da empresa.

Concluindo, a diferença entre sistemas de EIS e *Data Mining* pode ser entendida da seguinte forma: se você tem perguntas específicas e sabe os dados de que necessita para as respostas, então use um EIS. Por outro lado, quando você não sabe como perguntar, mas precisa de respostas para problemas, use *Data Mining*.

2.18.1.5 Ferramentas de Visualização de Dados

As ferramentas que permitem a visualização dos dados de um sistema de *Data Warehouse* assumem um importante papel no sentido de cativar o usuário e tornar mais fácil e rápido o uso do sistema. Existem ferramentas que apresentam as informações na forma de linhas e colunas, em planilhas, onde os dados são dispostos de uma forma que torna a sua análise mais trabalhosa do que um gráfico, que apresenta visualmente os pontos extremos, máximos e mínimos.

Com o uso destas técnicas, pode-se encontrar informações que estariam dispersas na massa de dados e que, através de ferramentas estatísticas, poderiam não ser evidenciadas.

2.19 Metodologias de Desenvolvimento de um Data Warehouse

Seguindo Barbieri (2001), os primeiros projetos de Data Warehouse muito provavelmente caminharam pelas trilhas metodológicas originadas basicamente em duas fontes de inspiração: Bill lnmon ou Ralph Kimball.

2.19.1 Metodologia Segundo *Ralph Kimball*

O Ciclo de Vida para a implementação do *Data Warehouse,* segundo Ralph Kimball, está ilustrado na FIGURA 8. Este diagrama descreve a seqüência de tarefas requerida para o projeto, desenvolvimento e aplicação do *Data Warehouse.* O diagrama da FIGURA 8 mostra a metodologia de projeto na qual módulo serve como um indicador. As fases descritas a seguir foram retiradas de Kimball (1998a).

☛ **Planejamento do projeto** ☚ O ciclo de vida inicia-se com o planejamento de projeto. Envolve a definição e a abrangência do projeto do *Data Warehouse,* incluindo avaliação e justificação do negócio; tarefas criticas, devido à alta visibilidade e custos associados. O planejamento de projeto enfoca os recursos e o nível de habilidade requerida do pessoal, juntamente com a programação das tarefas associadas e a sua duração.

O planejamento de projeto identifica todas as tarefas associadas com o ciclo de vida empresarial e os comentários das partes envolvidas, serve como base para a administração contínua do projeto do *Data Warehouse.*

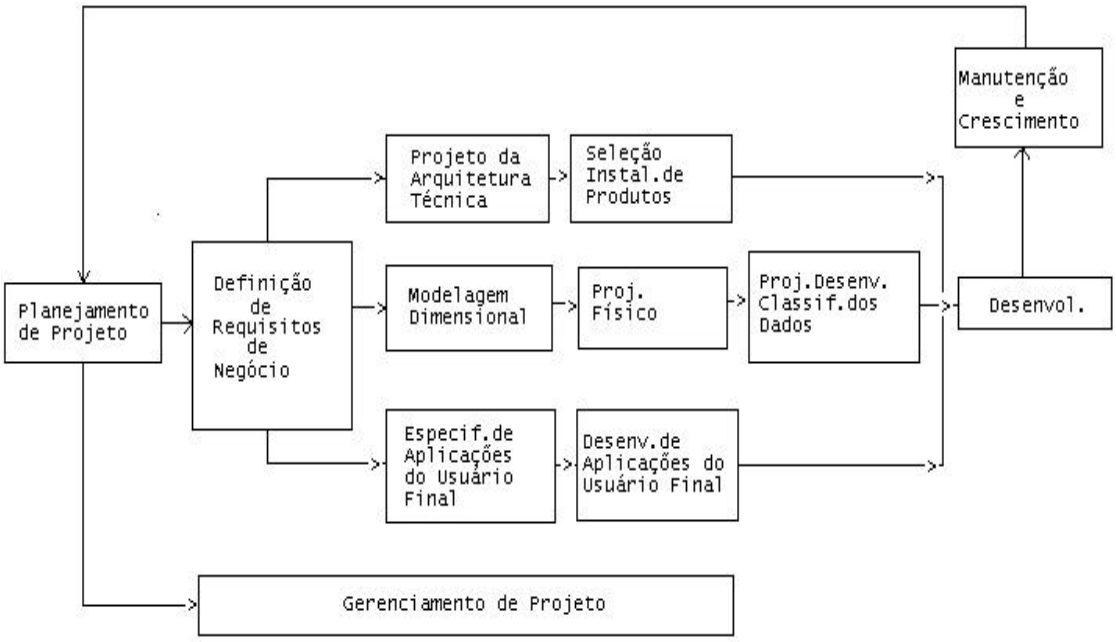

FIGURA 8 – DIAGRAMA DO CICLO DA VIDA
FONTE: KIMBALL (1998a)

☛ **Definição de requisitos de negócio** ☚ A probabilidade para o sucesso do *Data Warehouse é* aumentada quando se entende a exigência dos usuários. Projetistas do *Data Warehouse* têm que entender os fatores-chave que dirigem o negócio para determinar as exigências efetivamente empresariais e traduzi-las em considerações de implementação.

☛ **Modelagem dimensional** ☚ A definição das exigências empresariais determina os dados que precisam ser armazenadas no *Data Warehouse* para atender às exigências dos usuários. Inicia-se construindo uma matriz que representa os processos de negócio e a dimensionalidade. A matriz serve como uma "fotocópia" para assegurar que o *Data Warehouse* está extensível com o passar do tempo. A partir dessa matriz, uma análise de dados mais detalhada é realizada e juntada às exigências empresariais. Um Modelo Dimensional é desenvolvido, identificando as dimensões associadas, atributos e fatos. O projeto lógico do banco de dados é completado com estruturas apropriadas e relacionamento entre as chaves.

☛ **Projeto físico** ☚ O enfoque principal do projeto físico do banco de dados é definir as estruturas necessárias para apoiar o projeto lógico do banco de dados. Os elementos primários deste processo incluem a definição da nomenclatura padrão e a construção do ambiente de banco de dados.

☛ **Projeto e desenvolvimento da classificação de dados** ☚ O projeto e desenvolvimento da plataforma de classificação dos dados são tipicamente os mais subestimados no projeto do *Data Warehouse*. O processo de classificação de dados possui três passos principais: extração, transformação e povoamento. O processo de extração expõe questões de qualidade de dados que estiveram ocultos dentro do sistema operacional. Já que a qualidade dos dados tem um significante impacto na credibilidade do *Data Warehouse*, necessita-se resolver esses problemas de qualidade durante a classificação dos dados. Para aumentar sua complexidade, necessita-se projetar e construir dois processos de classificação, um para o povoamento inicial e outra para as cargas regulares.

☛ **Projeto da arquitetura técnica** ☚ Ambientes do *Data Warehouse* requerem a integração de numerosas tecnologias. O projeto de arquitetura estabelece a arquitetura global. Deve-se considerar três fatores: as exigências empresariais, ambientes técnicos atuais e o planejamento estratégico.

☛ **Seleção e instalação de produtos** ☚ Ao se usar o projeto de arquitetura, os componentes específicos como a plataforma de *hardware* e o sistema de administração de banco de dados devem ser avaliados e selecionados.

Um processo de avaliação técnica é definido junto com fatores de avaliação específicos para cada componente arquitetônico. Uma vez que os produtos foram avaliados e selecionados, eles são instalados e testados para assegurar uma integração de fim-a-fim apropriada dentro do ambiente do *Data Warehouse*.

☛ **Especificação de aplicações do usuário final** ☚ Especificações de aplicações descrevem o modelo de relatório, parâmetros e cálculos exigidos. Estas especificações asseguram que a equipe de desenvolvimento e os usuários tenham um entendimento comum das aplicações a serem entregues.

☛ **Desenvolvimento de aplicação do usuário final** ☚ O desenvolvimento das aplicações do usuário envolve a configuração das ferramentas e a construção dos relatórios. Opcionalmente, aplicações que são construídas tendo acesso a ferramentas automatizadas proporcionam maior produtividade para a equipe de desenvolvimento de aplicação; além disso, oferecem um poderoso mecanismo para os usuários facilmente modificarem os modelos de relatório existentes.

☛ **Desenvolvimento** ☚ Representa a convergência de tecnologia, dados e aplicações de usuário. Um planejamento é exigido para assegurar que tudo seja ajustado corretamente. Deve ser estabelecido apoio ao usuário e estratégias de realimentação antes que qualquer usuário tenha acesso ao *Data Warehouse*.

☛ **Manutenção e crescimento** ☚ Intenso trabalho segue o desenvolvimento inicial do *Data Warehouse*. É necessário continuar enfocando os usuários, proporcionando apoio contínuo e treinamento. Devem ser medidos e anotados a aceitação e desempenho do *Data Warehouse* no passar do tempo.

As mudanças devem ser vistas como um sinal de sucesso, não de fracasso. Devem ser estabelecidos processos de prioridades para tratar esta demanda de modificações, para se obter uma evolução e crescimento. Depois de identificar as prioridades de projeto, se regressa para o começo do ciclo de vida, construindo o que já foi estabelecido no ambiente do *Data Warehouse* com um enfoque nas novas exigências.

☛ **Gerenciamento de projeto** ☚ O gerenciamento de projeto assegura que as atividades do ciclo de vida permaneçam dentro do esperado e em sincronização. Elas acontecem ao longo do ciclo de vida e têm enfoque no monitoramento do estado do

projeto, localização do assunto e controle das mudanças para preservar os limites do *Data Warehouse.*

2.19.2 Metodologia Segundo *Inmon*

A metodologia desenvolvida por Willian H. Inmon é composta por vários passos. Inicialmente é necessário que um modelo de dados tenha sido definido. As principais áreas de interesse precisam ser identificadas, definidas claramente as fronteiras do modelo, separando dados primitivos de dados derivados, identificando chaves, atributos, agrupamentos de atributos, relacionamento entre agrupamentos de atributos, dados repetitivos e tipos de dados para cada área de interesse. Os passos são ilustrados na FIGURA 9 e descritos a seguir, retirados de Inmon (1997).

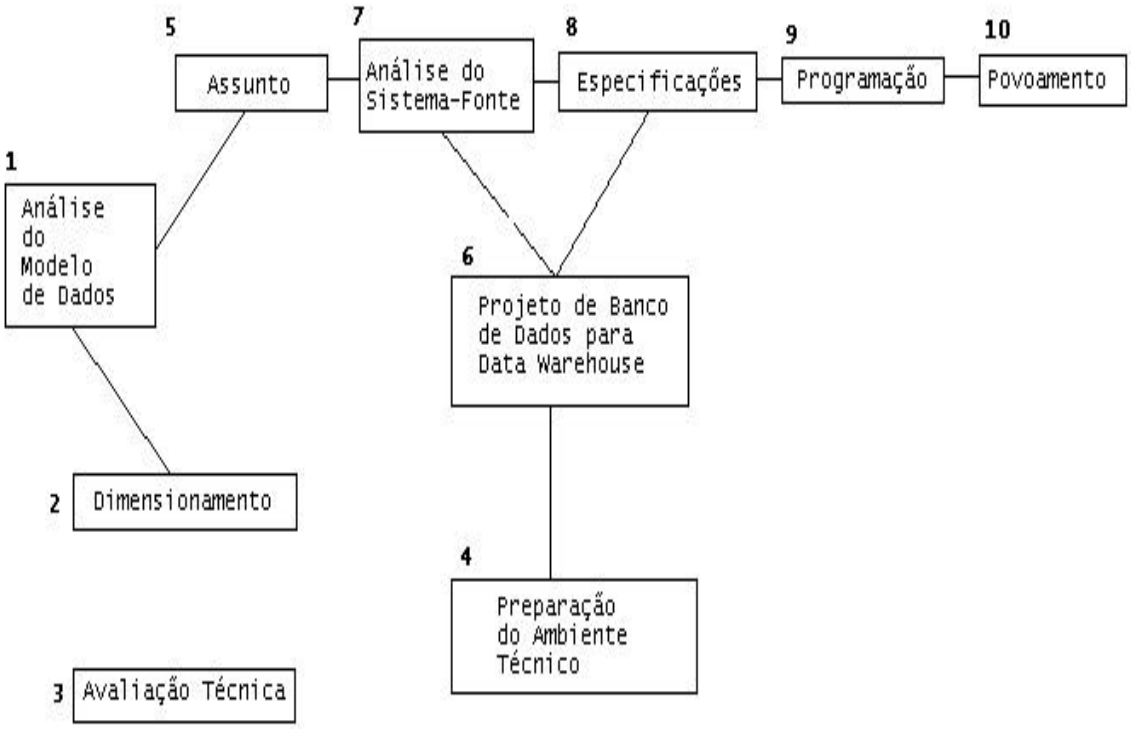

FIGURA 9 – DESENVOLVIMENTO DO *DATA WAREHOUSE*
FONTE: INMON (1997)

☛ **Análise do modelo de dados** ☚ Este passo resulta na confirmação de que a organização construiu um sólido modelo de dados. Se o modelo não atender aos critérios especificados, o andamento do projeto deve ser interrompido até que o modelo seja elevado a um padrão aceitável de qualidade.

☛ **Dimensionamento** ☚ Uma vez que o modelo de dados tenha sido analisado e transferido a um nível de qualidade aceitável, o próximo passo consiste em realizar o dimensionamento, que é uma estimativa do ambiente do Sistema de Apoio à Decisão (SAD). Se o volume de dados vai ser um problema, é importante saber disso no início. O dimensionamento simplesmente projeta, em termos brutos, que quantidade de dados o *Data Warehouse* vai conter.

O resultado do dimensionamento é simples, se o *Data Warehouse* irá conter grandes quantidades de dados, é necessário levar em consideração a possibilidade de existência de vários níveis de granularidade. Se o *Data Warehouse* não for destinado a conter uma enorme quantidade de dados, não existe a necessidade de planejar o projeto de vários níveis de granularidade.

☛ **Avaliação técnica** ☚ Os requisitos técnicos para o gerenciamento do *Data Warehouse* são muito diferentes dos requisitos e considerações técnicas para o gerenciamento de dados e processamento no ambiente operacional. Esse é o motivo pelo qual é tão comum a existência de um depósito central de dados - Sistema de Apoio a Decisão (SAD).

Quando apropriadamente conduzido, a definição técnica do *Data Warehouse* satisfaz aos seguintes critérios:

- capacidade de gerenciar grandes quantidades de dados;
- capacidade de permitir que os dados sejam acessados de modo flexível;
- capacidade de organizar dados de acordo com um modelo de dados;
- capacidade tanto de receber quanto de enviar dados para uma ampla variedade de tecnologias;
- capacidade de periodicamente carregar grandes quantidades de dados;

- capacidade de acessar um registro por vez.

☛ **Preparação do ambiente técnico** ☚ Uma vez que a configuração da arquitetura do *Data Warehouse* tenha sido definida, o próximo passo consiste em identificar tecnicamente como a configuração pode ser acomodada. Algumas das questões típicas que precisam ser tratadas aqui são:

- a quantidade de DASD (por exemplo, o *winchester)* necessária;
- que enlace, quer fora da rede ou em rede, será necessário;
- o volume de processamento previsto;
- como diminuir ou atenuar conflitos de processamento entre programas de acesso concorrente;
- o volume de tráfego que será gerado pela tecnologia que controla o *Data Warehouse;*
- a natureza do tráfego, interrupções curtas ou longas, geradas pela tecnologia que controla o *Data Warehouse.*

Quando esse passo é apropriadamente conduzido, não existe impedimento técnico para o sucesso. Entre os componentes técnicos que devem estar instalados, alocados, ligados e prontos para receber dados encontram-se as redes, o DASD, o sistema operacional que gerência o DASD, a interface do *Warehouse,* o software usado para gerenciar o *Data Warehouse.*

☛ **Análise das áreas de interesse** ☚ Neste ponto, é selecionada a área de interesse a ser povoada. A primeira área de interesse a ser selecionada deve ser suficientemente grande para ter sentido e suficientemente pequena para ser implementada. Se, por acaso, uma área de interesse for verdadeiramente grande e complexa, um subconjunto desta pode ser escolhido para sua implementação. O resultado deste passo é um escopo de empreendimento em relação a uma área de interesse. Quando essa fase é conduzida corretamente, o resultado é uma definição de que área de interesse (ou assunto) deve ser povoada a seguir. Nos primeiros povoamentos, geralmente são selecionadas pequenas áreas de interesse. Quando tudo é feito de forma apropriada, o assunto selecionado para

ser povoado atende às necessidades do nível corrente de desenvolvimento do *Data Warehouse*.

☛ **Projeto do *Data Warehouse*** ☚ O *Data Warehouse é* projetado com base no modelo de dados. Algumas das características do projeto mais recente incluem:

- acomodação dos diferentes níveis de granularidade, caso existam vários;
- orientação de dados para os principais assuntos da empresa;
- presença de somente dados primitivos e dados derivados públicos;
- ausência de dados do Sistema de Apoio a Decisão (SAD);
- A variação em relação ao tempo de cada registro de dados;
- desnormalização física de dados onde é aplicável, ou seja, onde a performance exigir;
- criação de artefatos de dados por meio dos quais as informações que estiverem no ambiente operacional são passadas para o *Data Warehouse*.

O resultado deste passo é um projeto físico de banco de dados para o *Data Warehouse*. Percebe-se que, no início, nem todo o *Data Warehouse* precisa ser projetado em detalhes. É recomendável inicialmente projetar as principais estruturas do *Data Warehouse e*, então, preencher os detalhes em um momento posterior.

☛**Análise do sistema-fonte** ☚ Uma vez que o assunto a ser povoado é definido, a próxima etapa consiste em identificar, no ambiente de sistemas existentes, a fonte de dados para o assunto. É normal que existam várias fontes para os dados do Sistema de Apoio a Decisão (SAD). É neste momento que são tratadas as questões de integração. A seguir, estão listadas as questões que devem ser tratadas aqui:

- estrutura de chave e escolha de chave à medida que os dados passam do ambiente operacional para o ambiente Sistema de Apoio a Decisão (SAD);
- o que fazer quando existem diversas fontes de dados para escolher, o que fazer quando não existe fonte de dados para escolher, quais

transformações, codificação, decodificação, conversões, precisa ser feitas à medida que os dados são selecionados para a passagem para o *Data Warehouse;*

- de que modo a variação em relação ao tempo será criada a partir dos dados de valor corrente;
- como a estrutura de Sistema de Apoio a Decisão (SAD) será criada a partir da estrutura operacional;
- de que modo os relacionamentos operacionais serão representados no ambiente Sistema de Apoio a Decisão (SAD).

O resultado desse passo é o mapeamento de dados do ambiente operacional para o ambiente Sistema de Apoio a Decisão (SAD).

☛ **Especificações** ☚ Uma vez que a interface entre os ambientes operacionais e Sistema de Apoio a Decisão (SAD) tenha sido delineada, o próximo passo consiste em formalizá-la em termos de especificações de programas. O resultado dos passos é a descrição dos programas que serão usados para efetuar a passagem dos dados do ambiente operacional para o ambiente do *Data Warehouse.*

☛ **Programação** ☚ Este passo inclui todas as atividades-padrão de programação como o desenvolvimento de pseudocódigo quando corretamente executado. O código que é gerado nesse passo é eficiente, documentado, com possibilidade de ser facilmente alterado, exato e completo.

☛ **Povoamento** ☚ Este passo é a execução dos programas Sistema de Apoio à Decisão (SAD) anteriormente desenvolvidos. As questões aqui tratadas são as seguintes:

- freqüência de povoamento;
- eliminação de dados povoados;
- obsolescência dos dados povoados, ou seja, a execução de programas de resumo de conferência;
- gerenciamento de vários níveis de granularidade;

- renovação dos dados de amostra viva, caso tenham sido construídas tabelas de amostra viva.

O resultado desse passo consiste em um *Data Warehouse* povoado e funcional.

3 ESTUDO DE CASO - A IMPLANTAÇÃO DO *DATA WAREHOUSE* EM UMA COOPERATIVA MÉDICA

Este capítulo tem o objetivo de apresentar um estudo de caso da implantação do *Data Warehouse* em uma Cooperativa Médica, aplicando a metodologia de Ralph Kimball.

3.1 Metodologia Segundo Ralph Kimball

Para Kimball (1998b), o *Data Warehouse* é um local em que as pessoas podem acessar seus dados, conforme já foi dito em parágrafo anterior deste trabalho. As metas fundamentais de um *Data Warehouse* podem ser desenvolvidas andando-se pelos corredores de uma organização de grande porte e ouvindo as conversas nas gerencias. Os temas que se repetem em todos os diálogos são: "Possuímos montanhas de dados nesta empresa, mas não conseguimos acessá-los." Nada deixa o chefe mais enfurecido do que duas pessoas apresentando o mesmo resultado do negócio, mas com números diferentes. "Queremos acessar os dados de todas as formas". "Todos sabem que alguns dos dados não estão bons". Essas preocupações são tão universais, que determinaram as necessidades fundamentais de um *Data Warehouse*..

Ainda segundo esse autor, *Data Warehouse* é uma fonte de dados consultáveis da organização, formado pela união de todos os *Data Marts* correspondentes.

A metodologia aplicada na Cooperativa foi a de Ralph Kimball, na qual a abordagem é *Data Marts Departamental*. De acordo com Kimball (1996), Data Mart é um sistema especializado que fornece informações departamentais ou para uma aplicação relacionada.

Os custos de uma implementação "por inteiro", em termos de recursos consumidos e o impacto no ambiente operacional da Cooperativa justificam esta estratégia. Abordagem incremental de implantação de um *Data Warehouse* prioriza os temas mais estratégicos, deixando para acrescentar os demais em etapas posteriores.

3.2 Planejamento do Projeto

"Qual a produção de cada cooperado e quantos exas este gera por consulta?"

"Qual a média de produção por especialidade e qual o desvio padrão da área?"

"Qual o percentual de exames auto-gerados?"

"Qual a utilização do cliente dentro de uma clínica?"

"Qual a taxa de utilização de cada plano, cada cliente e cada família?".

A Cooperativa já pode responder a estas perguntas e outras, depois de intensos estudos para definir e implementar a mais moderna tecnologia em informática na área de estatística e controle de produção médica. Pela primeira vez em toda a sua história, a Cooperativa pode obter as informações necessárias com precisão e seriedade.

Através do projeto *Data Warehouse*, as informações sobre a produção médica podem ser cruzadas de acordo com as prioridades da Cooperativa. Esse controle indispensável será o principal aliado dos cooperados na obtenção das melhorias no valor da renumeração médica por parte da Cooperativa, que poderá atuar com segurança e rigidez nos desvios comprovados e que prejudicam a Cooperativa e todos os cooperados do sistema.

As informações processadas mensalmente pela área de produção médica geram uma quantidade significativa de dados relativos ao desempenho de seus prestadores de serviços e utilização dos seus diversos planos de saúde. Com a implementação do *Data Warehouse*, é possível disponibilizar, com critérios, dados consolidados, gerando uma fonte única de informação de forma muito mais rápida aos usuários finais.

A Cooperativa não possui, no sistema transacional, um módulo para tratamento de estatística. Toda a aplicação sobre o negócio era solicitada ao Departamento de Informática para desenvolvimento de programas específicos, sendo que, para incorporar uma nova consulta, são necessárias várias horas de programação e de definições para atingir resultados satisfatórios.

Neste contexto, faz-se necessária a implantação de um *Data Warehouse* com *Data Marts Departamentais*, pela carência de flexibilidade e falta de autonomia dos usuários que o sistema possui e, também, para uma boa implementação de projetos envolvendo o conceito de *Customer Relationship Management (CRM), Data Mining* e a extração de relatórios de *Bussiness Intelligence (BI),* que dependem, em larga escala, da estrutura desses do *Data Warehouse*.

Os principais objetivos da implantação do *Data Warehouse* na Cooperativa foram:

- fonte única de informações com dados consolidados;
- sistemas transacionais em uso;
- conhecer a realidade da Cooperativa para poder traçar um efetivo planejamento de ações;
- localizar distorções e incorreções praticadas, na tentativa de orientar os cooperados e suprimir algumas práticas que prejudicam o desenvolvimento da Cooperativa;
- educar e conscientizar os cooperados para que eles se comprometam com a saúde financeira de sua Cooperativa;
- auxiliar aos cooperados na definição de técnicas e procedimentos, que, muitas vezes, não são do seu conhecimento, tendo em vista a velocidade das mudanças ocorridas na medicina em todo mundo.

De acordo com a justificativa, construiu-se um protótipo de um *Data Warehouse* para a Produção Médica, em que foram definidas diversas etapas do projeto, abaixo descritas:

- **Análise das informações necessárias para cada departamento**: Foram entrevistados os usuários com cargos gerenciais e de direção. O objetivo era

reconhecer quais as informações necessárias para cada um dos departamentos de forma a transformá-las em estatísticas.

- **Análise dos dados disponíveis**: O primeiro passo para a construção de um *Data Warehouse* é a identificação das diversas fontes de dados. Para cada departamento, avaliamos quais as informações estão disponíveis.
- **Validação e importação dos dados para o MS SQL Server 2000**: Após a identificação e seleção dos dados necessários para atender às necessidades levantadas, os mesmos são importados para o banco de dados *MS SQL Server 2000*, através da criação de programas ou mecanismos de importação.
- **Construção de mecanismos para a atualização dos dados**: Foi estudada a melhor forma para a importação dos dados disponíveis de acordo com a periodicidade das informações selecionadas. A atualização dos dados pode ser diária, semanal, mensal, etc. Definida a periodicidade, foram criados mecanismos que fizessem a importação automática dos dados.
- **Criação dos CUBOS estatísticos**: Após a importação e formatação dos dados foram criados os cubos estatísticos desejados.
- **Configuração das permissões de acesso*:*** Foram criados os usuários e grupos de usuários que poderão acessar os cubos. Esses grupos terão perfis de acesso às informações existentes de acordo com o critério definido pela Cooperativa, o que garantirá a segurança no acesso às informações.
- **Treinamento dos usuários**: Após a conclusão dos passos anteriores, os usuários foram treinados para aprender como manipular as informações existentes seja na criação das estatísticas, seja na criação de gráficos e relatórios. Para isso foi ministrado um curso sobre o *Data Warehouse*, qual dado pode-se extrair e de como usar o MS Excel para essa finalidade. O treinamento envolveu, ainda, o estudo de comandos estatísticos como média, desvio padrão, etc, permitindo a capacitação dos usuários.

3.3 Definição de Requisitos de Negócio

A fase de definição dos requisitos de negócio da Cooperativa decorreu de acordo com a metodologia de Kimball. O escopo foi o da Produção Médica, uma vez que já existe uma cultura de análise de informações, o que, em tese, "quebraria" as possíveis barreiras iniciais à implantação dos *Sistemas de Apoio à Decisão* (SAD).

É possível ter controle dos valores, investigar em que momentos e porquê estes valores estão acima ou abaixo dos totais. Esta área é bastante utilizada para definição de protótipos, pois é a que mede as entradas e saídas efetivas de capital da Cooperativa.

3.4 Modelagem Dimensional

O sistema transacional que está em uso na Cooperativa é o *Dataflex* para *Unix*. Neste sistema, temos diversos arquivos relacionados que servem basicamente para armazenar os dados relativos aos procedimentos de consulta, exames, internações, atendimentos ambulatoriais, cadastro de empresas, clientes, prestadores de serviços etc.

Quando o cliente é atendido por um prestador de serviços, gera um ou mais eventos para o sistema transacional, que verifica dados básicos do cliente como: código de identificação, nome, data de nascimento, plano, coberturas, carências, financeiro etc, para posterior liberação.

Na FIGURA 10, temos um modelo E-R reduzido apenas com as entidades relevantes para o projeto, embora existam muitas outras entidades que não foram consideradas no desenvolvimento do OLAP.

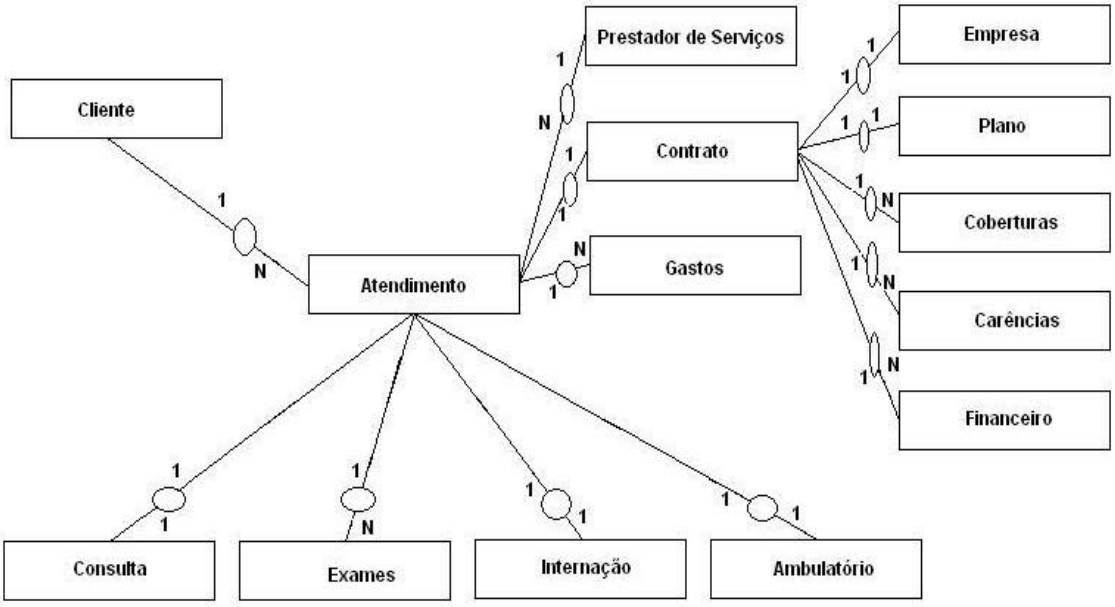

FIGURA 10 - MODELO E-R REDUZIDO DO SISTEMA TRANSACIONAL – COOPERATIVA MÉDICA

A Cooperativa utiliza *Data Marts Multidimensionais*, aonde a informação é organizada de acordo com o ESQUEMA ESTRELA, baseado em dois conceitos: dimensão e fato.

Conforme Kimball (1998b), iniciou-se construindo uma matriz que representa os processos de negócio e a dimensionalidade, apresentada na FIGURA 11, que ilustra uma tabela-fato, a Produção Médica, em que se encontram procedimentos de consultas, exames, internações e ambulatórios e que está relacionada com as tabelas de dimensões: Tempo, Cliente, Empresa e Prestador de Serviços.

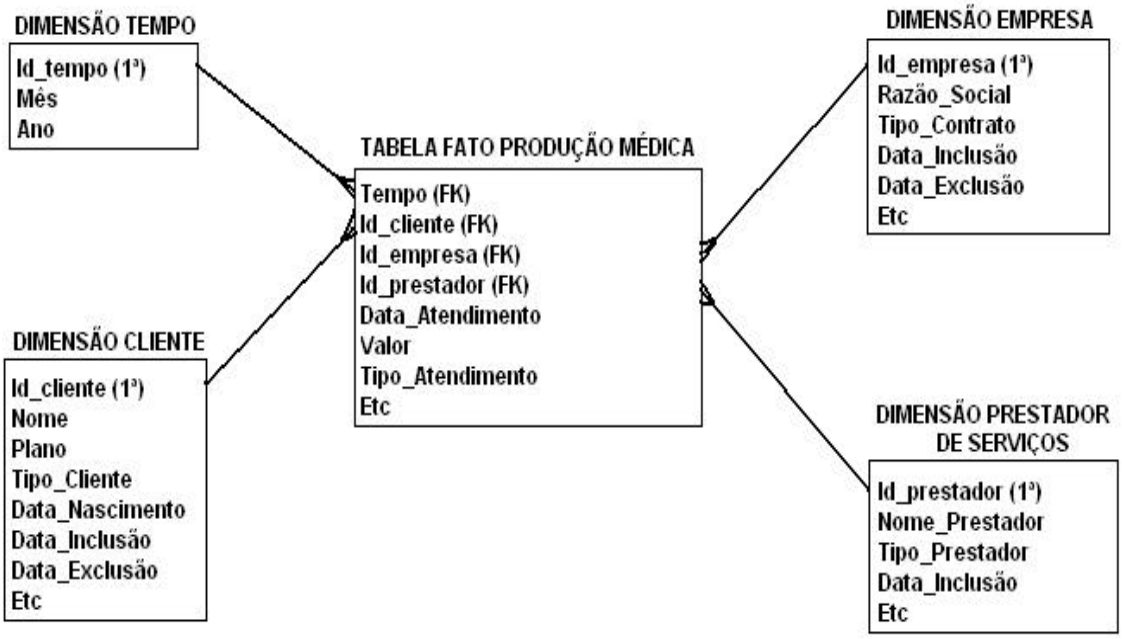

FIGURA 11 - MODELAGEM – ESQUEMA ESTRELA – COOPERATIVA

3.5. Projeto Físico

Lambert (1996), afirma que um *Data Warehouse* é um sistema complexo que permite integrar um conjunto de componentes, tais como diversos tipos de software e de hardware, redes de computadores, sistemas de comunicações de dados, servidores, mainframes e sistemas de administração de bancos de dados, com também muitas pessoas de diferentes unidades organizacionais, com objetivos diferentes.

Para organizar os dados, são necessários novos métodos de armazenamento, estruturação e novas tecnologias para a geração e recuperação dessas informações. Essas tecnologias já estão bem difundidas e oferecem diferentes opções de ferramentas para cumprir todas as etapas.

Essas tecnologias diferem dos padrões operacionais de sistemas de banco de dados em três maneiras:

- dispõem de habilidade para extrair, tratar e agregar dados de múltiplos sistemas operacionais em *Data Marts* ou *Data Warehouses* separados;
- amazenam dados freqüentemente em formato de cubo (OLAP) multidimensional, permitindo rápido agregamento de dados e detalhamento de análises (*drill down*);
- disponibilizam visualizações informativas, pesquisando, reportando e modelando capacidades que vão além dos padrões de sistemas operacionais freqüentemente oferecidos.

O projeto físico aplicado na Cooperativa, ilustrado na FIGURA 12, está definido sob uma arquitetura transacional com *IBM RISC 6000 B80*, com sistema operacional AIX versão 4.3.3.0-10. O ambiente é baseado em servidor Unix e o SGBD é o DataFlex versão 3.05C (produto da Data Access Corporation).

A arquitetura informacional é Servidor IBM Netfinity 5000. O ambiente é baseado em servidor Windows 2000 Server e o SGBD é Microsoft MS SQL Server 2000.

A existência de uma rede de microcomputadores ponto-a-ponto com Windows 98 e XP definiu a arquitetura do sistema baseada em Cliente/Servidor.

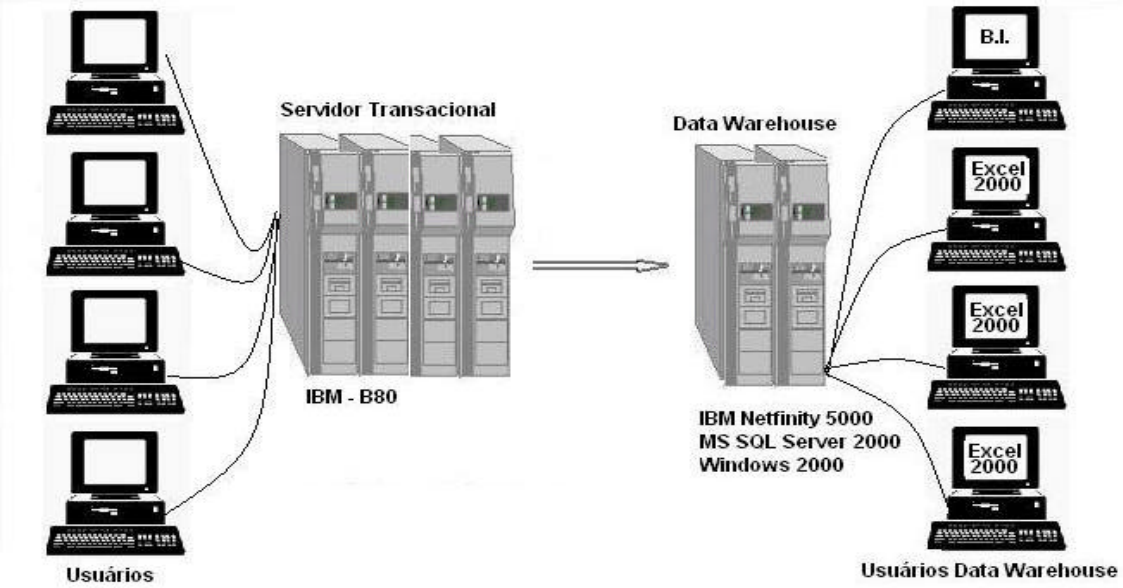

FIGURA 12 – PROJETO FÍSICO – COOPERATIVA MÉDICA

3.6 Projeto e Desenvolvimento da Classificação dos Dados

Segundo Kimball (1998), o processo de ETL compreende as atividades de extração, limpeza, transformação, carga e indexação e serve para preparar os dados provindos dos sistemas OLTP para o uso no *Data Warehouse* .

Uma das dificuldades do projeto foi relativa à importação e tratamento dos dados do DataFlex para o MS SQL Server 2000. Foram criados diversos programas extratores em DataFlex, que geraram arquivos-textos trazidos da base transacional para serem transformados, integrados e limpos antes de serem efetivamente carregados para o MS SQL Server 2000, conforme ilustrado na FIGURA 13.

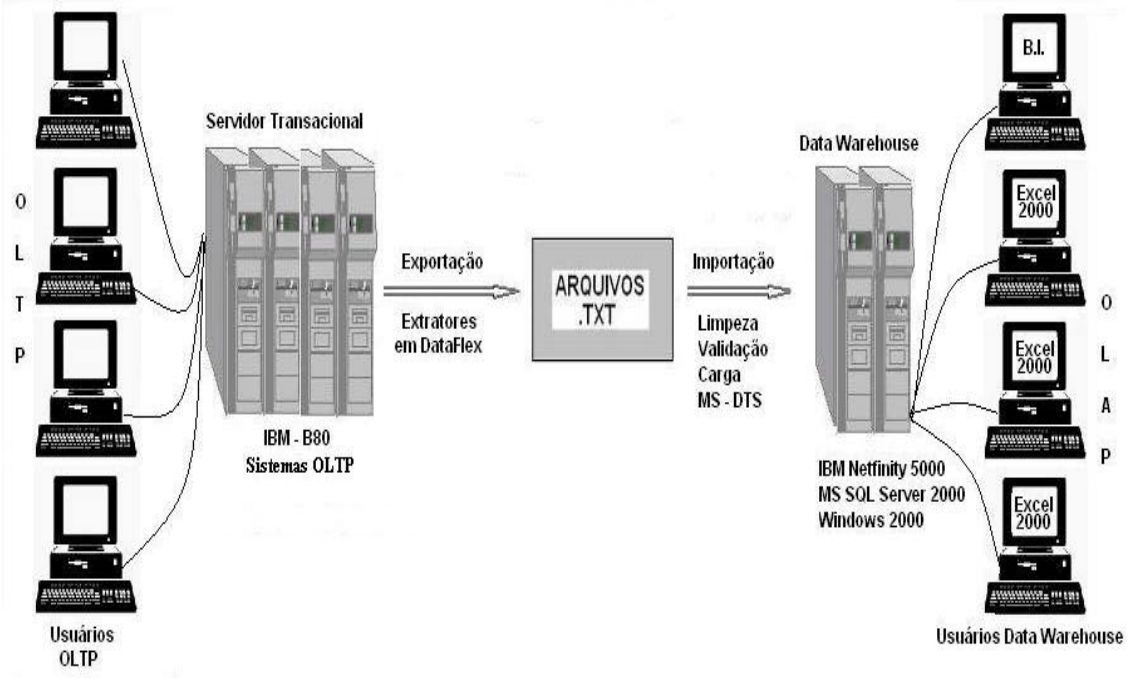

FIGURA 13 - TRANSFORMAÇÃO OLTP PARA OLAP – COOPERATIVA MÉDICA

A partir daí, foram criados os diversos depósitos de dados (*Data Marts*) para serem manipulados pelos usuários, conforme ilustrado na FIGURA 14 (pagina seguinte).

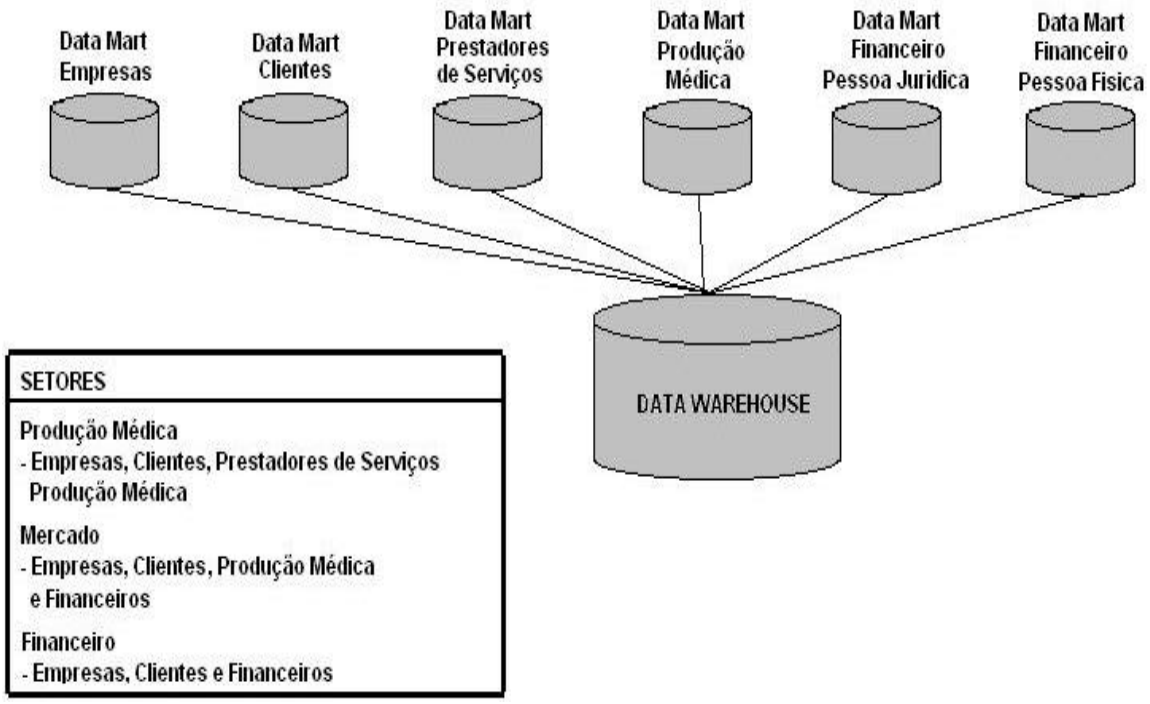

FIGURA 14 - MODELO CONCEITUAL DOS *DATA MARTS* – COOPERATIVA MÉDICA

3.6.1 Extração

Esta etapa foi a mais demorada e desgastante do projeto. A extração transacional dos dados atualmente é feita por programas específicos desenvolvidos na linguagem DataFlex que geram arquivos-textos posteriormente transferidos para MS SQL Server 2000. Todo esse processo é executado devido ao DataFlex não possuir conexão com o drive OBDC.

Os extratores geram informações de empresas, clientes, prestadores de serviços, produção médica e financeiros. Os arquivos-textos foram importados para tabelas intermediárias do banco de dados, através da ferramenta Microsoft Server 2000 DTS Package.

3.6.2 Limpeza – Redundância – Validação

A etapa consistiu em limpar das tabelas intermediárias estes dados a fim de que os mesmos obedecessem a um padrão que permitirá futuras comparações sem que haja a necessidade de executar operações de conversão durante a realização de consultas, o que possivelmente tornaria o processo de pesquisa extremamente lento e trabalhoso em alguns casos. Nesta etapa, identificou-se informações redundantes e dados sem valor para o negócio através de scripts específicos. Atualmente esse processo de limpeza/validação já tem definidas as regras, pelas quais é executado um script no MS Server Query Analyzer. Durante o processo, as dificuldades apareceram nas seguintes tabelas de fato:

Tabela Empresas:

- falta de padronização nos campos: razão social, endereço, bairro e cidade;
- classificação dos grupos das empresas;
- classificação dos tipos de contratos;
- campo data de exclusão – datas incorretas.

Tabela de Clientes:

- falta de padronização nos campos: nome, endereço, bairro e cidade;
- classificação dos tipos de planos – descrição;
- classificação dos tipos de clientes;
- campo sexo – em branco ou outros;
- campo data de nascimento – sem data;
- campo plano – valor em branco;
- campo data de exclusão – datas incorretas.

Tabela de Prestadores de Serviços:

- falta de padronização nos campos: nome, endereço, bairro e cidade;
- classificação das especialidades e sub-especialidades médicas;
- classificação dos tipos e grupos de Prestadores de Serviços;

- campo data de exclusão – datas incorretas.

Tabela da Produção Médica:

- identificação dos procedimentos autogerados;
- identificação dos pacotes de procedimentos;
- classificação dos serviços (atos) em diagnóstico ou não com base nos códigos tabela AMB;
- tabela de complemento da nota;
- classificação do tipo de pacientes;
- identificação do tipo de procedimentos: internação ou ambulatorial.

3.6.3 Carga

Após a limpeza/validação, foi executado o processo para transferência das tabelas intermediárias para as tabelas de produção, através do MS Server Query Analyser com script de carga (*insert*).

3.6.4 Transformação

Após todas as etapas acima citadas estarem prontas, foi executado o processamento dos cubos através do MS Server OLAP Manager e comunicado aos usuários do *Data Warehouse* a disponibilidade da informação através de E-mail.

3.6.5 Tempo

Para os *Data Marts* Produção Médica e Financeiros foi utilizada a competência, que é composta do mês e ano do faturamento; os demais não são dados históricos, devido a cargas semanais.

3.6.6 Metadados

O repositório de metadados deverá ser desenvolvido na forma de páginas *HTML* (*hypertext markup language*) pela facilidade de implementação, padronização e disponibilidade a qualquer usuário do *Data Warehouse* através da Intranet da Cooperativa. Atualmente, os metadados são disponibilizados para o usuário *Data Warehouse* através de E-mail.

☛ **Metadados Clientes** ☚ Pode-se ver a quantidade de clientes classificados por faixa etária, idade, plano, empresa, cidade, estado, sexo, situação (ativo/inativo), estado civil, opcionais, etc. Essas informações podem ser cruzadas para obter-se, por exemplo, "clientes x por faixa etária x plano", "plano x empresa". etc. A seguir, serão listados alguns campos encontrados nesse *Data Mart*:

CAMPO	CONTEÚDO
Cliente	Código do Cliente
Nome	Nome do cliente
Data Nascimento	Data de nascimento
Data Inclusão	Data de inclusão do cliente no contrato
Data Exclusão	Data de exclusão do cliente no contrato
Sexo	Sexo
Estado Civil	Estado civil
Opcionais	Opcionais
Plano	Descrição do plano do cliente
Empresa	Contrato da empresa do Cliente

☛ **Metadados Produção Médica** ☚ Pode-se cruzar diversos campos referentes à Produção Médica de uma competência, trimestre ou um ano específico. A seguir, serão listados alguns campos encontrados nesse *Data Mart*:

CAMPO	CONTEÚDO
Auto-gerado	Indica se o serviço foi auto-gerado, ou seja, se requisitante do serviço é o prestador do mesmo.
Cliente	Qual o cliente do serviço prestado
Código Serviço	Qual o código do serviço
Competência	Competência na qual o serviço foi pago (mês/ano)
Empresa	Qual é o tipo de contrato: particular ou empresarial
Especialidade Requisitante	Qual a especialidade do requisitante do serviço
Especialidade Prestador	Qual a especialidade do Prestador de Serviços
Faixa Etária	Qual a Faixa etária do Cliente dos serviços
Credenciado	Qual o Hospital, Laboratório ou Clínica.
Plano	Qual o plano do Cliente do serviço
Prestador	Qual o Prestador dos Serviços
Requisitante	Qual o Requisitante dos Serviços
Sexo	Qual o sexo do Cliente
Tipo do Item	Qual o tipo do item: Taxa ou Diária.
Cooperativa	Qual a Cooperativa do Cliente
Qtde Notas	Quantidade de notas
Qtde Clientes	Quantidade de Clientes
Qtde Serviços	Quantidade de serviços
Total Pago	O total pago ao Prestador de Serviços. Inclui todos os campos abaixo e outros como valor pago UTI, material, medicamentos etc
Valor Pago Honorário Médico	Valor pago como honorário médico
Valor Pago CO	Valor pago como custo operacional

Com isso, pode-se saber todos os serviços realizados por especialidade, por faixa etária, o total pago para cada prestador de serviços, quais os serviços por prestador de serviços, qual a quantidade de clientes atendidos por prestador de serviços, quantidade de clientes atendidos por empresa ou plano, ou qualquer uma das combinações dos campos acima.

☛ **Metadados Financeiro Pessoa Jurídica** ☚ Permite cruzar os campos referentes aos pagamentos realizados pelas empresas clientes da Cooperativa. A seguir, serão listados alguns campos encontrados nesse *Data Mart*.:

CAMPO	CONTEÚDO
Empresa	Número do contrato da empresa
Contrato	Tipo do contrato: Custo Operacional ou Pré-Pagamento
Grupo	Grupo da Empresa
Tipo	Tipo da Empresa
Data Pagamento	Data do pagamento
Competência	Qual a competência do pagamento
Data do Cancelamento	Se a fatura foi cancelada representa a data do cancelamento
Valor	Valor total da fatura
Etc	Outros campos

☛ **Metadados Financeiro Pessoa Física** ☚ Permite cruzar os campos referentes aos pagamentos realizados pelos Clientes de plano particular. A seguir listamos alguns campos encontrados nesse *Data Mart*.

CAMPO	CONTEÚDO
Cliente	Código do Cliente
Contrato	Tipo do contrato: Custo Operacional ou Pré-Pagamento
Plano	Plano do Cliente
Idade	Idade do Cliente – titular
Estado Civil	Estado civil do titular
Competência	Qual a competência do pagamento (mm/aaaa)
Data do Vencimento	Data do vencimento
Valor	Valor total de pagamento
Data do pagamento	Data do pagamento
Local do Pagamento	Local onde foi realizado o pagamento
Etc	Outros campos

☛ **Metatados Prestadores de Serviços** ☚ Pode-se avaliar os Prestadores de Serviços da Cooperativa por especialidade, tempo de Cooperativa, idade etc. É possível cruzar essas informações para obter, por exemplo, Prestadores de Serviços por faixa etária, Prestadores de Serviços por tempo de Cooperativa, Prestadores de Serviços por especialidade etc. A seguir, serão listados alguns campos encontrados nesse *Data Mart*.

CAMPO	CONTEÚDO
Prestador	Código do Prestador
Nome	Nome do Prestador
Data Nascimento	Data de nascimento
Data Inclusão	Data de inclusão
Data Exclusão	Data de exclusão
Sexo	Sexo
Estado Civil	Estado civil
Especialidade	Descrição da especialidade
Sub-especialidade	Descrição da sub-especialidade

3.7 Projeto da Arquitetura Técnica

A arquitetura proposta por Kimball é denominada como “arquitetura de três camadas + “ou “arquitetura Molap”.

A arquitetura aplicada, proposta por Kimball (1998), devidamente adaptada ao contexto da Cooperativa, é a de três camadas, ilustrada na FIGURA 15, que parte do princípio que para a transformação dos dados do sistema operacional para o sistema analítico são necessários dois passos:

- a reconciliação dos dados originários da camada de dados de tempo real que têm o propósito de ser uma fonte única e definitiva de dados para os *Sistemas de Suporte a Decisão (SSD)*, constituindo-se de uma visão lógica do modelo de dados de todo o empreendimento. Os dados originários da camada de dados de tempo real são limpos de modo a eliminar as inconsistências, irregularidades e padronizar o formato dos dados. Um papel importante da camada de dados reconciliados é suportar novas e imprevisíveis necessidades dos usuários, quando não puderem ser satisfeitas pela camada de dados derivados;
- a derivação de dados, a partir da camada de dados reconciliados sobre a qual são realizadas diversas atividades, tais como combinação, transformação, dentre outros processos. Essas atividades visam popular a camada de dados derivados com os dados necessários para atender às requisições dos usuários e *Sistemas de Suporte a Decisão (SSD)*, a fim de satisfazer suas necessidades de informações e possibilitar a tomada de decisões. Normalmente os usuários acessam a camada de dados derivados, que contêm um conjunto comum de consultas predefinidas;
- facilita atendimento de novas necessidades para o usuário;
- aumenta o espaço utilizado;
- é ideal para construção de *Data Warehouse* integrado com *Data Marts*.

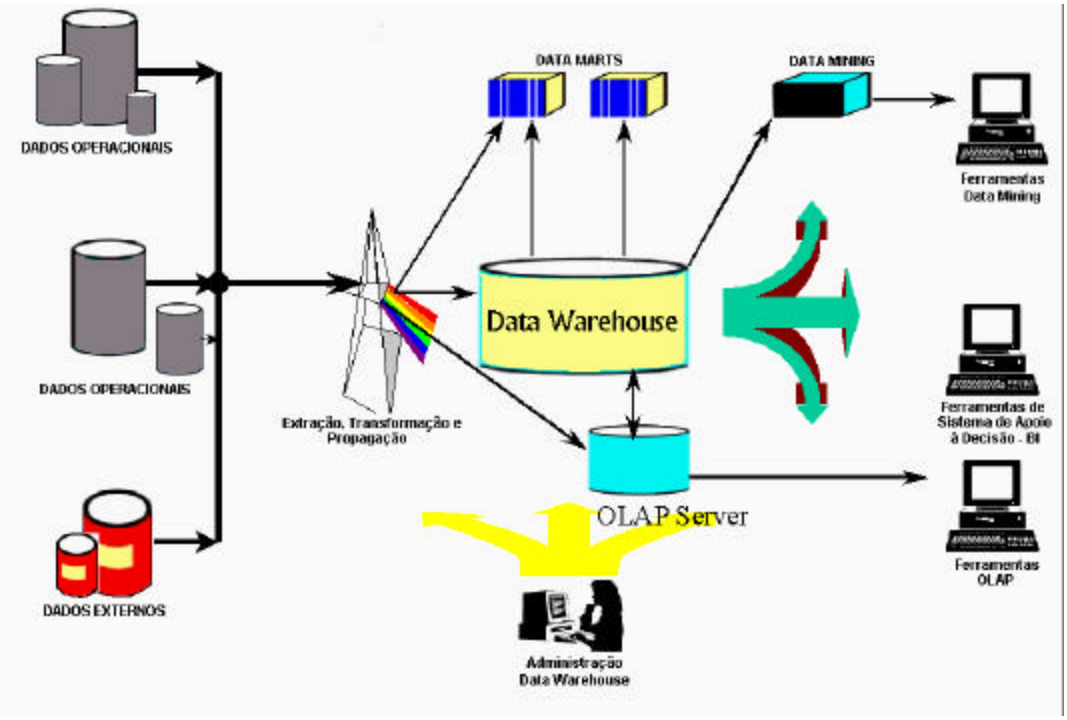

FIGURA 15 - ARQUITETURA UTILIZADA – TRÊS CAMADAS – COOPERATIVA MÉDICA

A vantagem da arquitetura utilizada é a separação de funções no ambiente de um *Data Warehouse*. Nessa arquitetura, o aumento do espaço de armazenamento é realizado de forma controlada e compreensível.

3.8 Seleção e Instalação de Produtos

Esta etapa foi executada após a criação do projeto de arquitetura técnica. As ferramentas utilizadas na implementação e validação do modelo dimensional obtido por intermédio do estudo de caso foram o *MS SQL Server 2000* para implantação do modelo relacional e o modelo dimensional, o *Olap Services 2000* para implementação das tabelas de dimensão e fato e o *Front Page* para a visualização dos dados. Todas foram de grande valia e sua funcionalidade nada deixou a desejar.

A *MS SQL Server 2000* é parte integrante dos servidores corporativos, em uma plataforma integrada, abrangente para construir e operar aplicações distribuídas em suas diversas funcionalidades e se mostrou como um ambiente adequado para o armazenamento, gerenciamento do *Data Warehouse* e análise das informações, através de técnicas que permitem o aprofundamento em níveis de detalhes das consultas. Esta tecnologia possibilita que análises complexas com grande volume de dados possam ser realizadas com resultados excepcionais. Além disso, todas as características do produto ajudam a diminuir o tráfego de rede e proporcionar melhores resultados de pesquisas.

3.9 Especificação de Aplicações do Usuário Final

Nesta etapa, procurou-se identificar as áreas prioritárias e, a partir destas, definiu-se um conjunto padronizado de aplicações destinadas aos usuários que necessitam ter acesso "ad hoc" aos dados do *Data Warehouse*. As atividades que compreenderam esta etapa foram as seguintes:

- priorização e identificação dos relatórios; desenvolvimento de uma estrutura geral que permita aos usuários acessar os diversos relatórios de forma estruturada (e.g. menu de relatórios) e de determinação de estruturas padronizadas de relatórios para os usuários finais;
- revisão das estruturas padronizadas e documentação;
- aceitação do projeto pelo usuário final;
- revisão do projeto.

3.10 Desenvolvimento de Aplicações do Usuário Final

Nesta etapa, foram desenvolvidas as aplicações necessárias de acordo com o levantamento realizado na etapa anterior.

3.11 Ferramentas *Front End*

Para Kimball (1998a), as ferramentas OLAP proporcionam consultas e apresentação dos dados com maior flexibilidade para o usuário.

As ferramentas aplicadas na Cooperativa foram OLAP e Sistema de Gestão, citados abaixo:

☛ **OLAP** ☚ O *MS SQL Server 2000* possui uma ferramenta de OLAP (estatística) embutida no produto e constitui-se num ambiente de consulta bastante completo e flexível, que permite a realização de diversas operações do universo das ferramentas analíticas, possibilitando ao usuário final "navegar" pelos dados do cubo. Assim, após a importação e tratamento dos dados foi criado o cubo estatístico no *MS OLAP Services*. As ferramentas utilizadas para *Front End* para *Data Warehouse* são os softwares *Microsoft Excel 2000* ou *Excel XP* pelos usuários *Data Warehouse*, conforme ilustra a FIGURA 16, que traz um modelo de planilha.

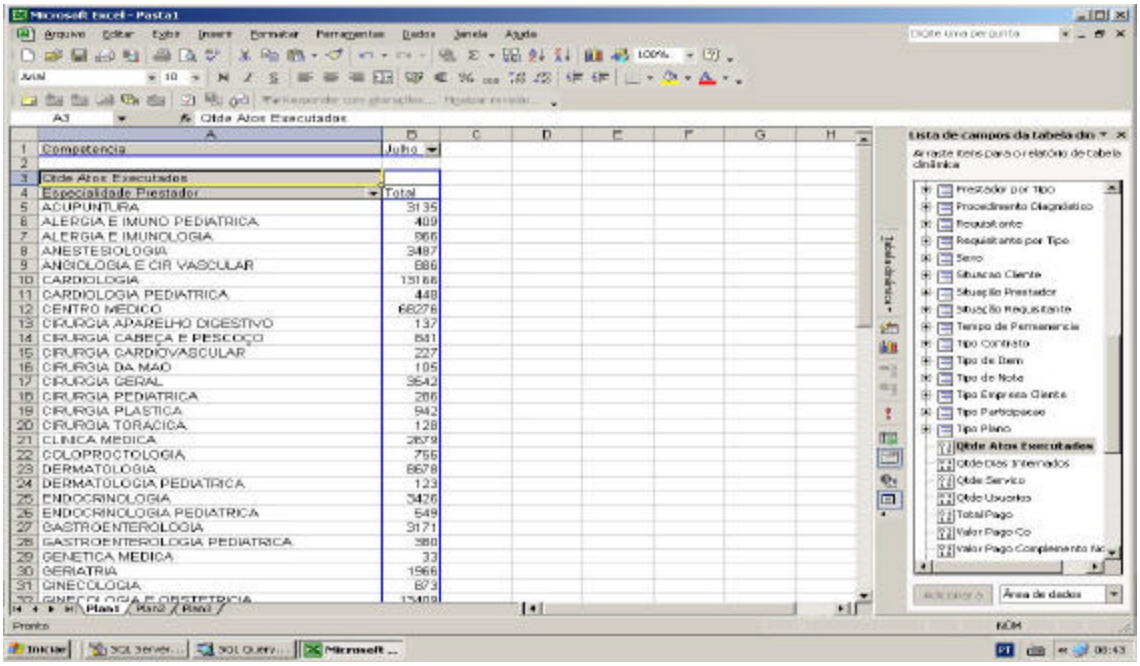

FIGURA 16 - Excel OLAP – MODELO DE PLANILHA – COOPERATIVA MÉDICA

A integração do *MS Excel* com *SQL Server 2000* proporcionou um modo fácil de criar e distribuir relatórios gráficos e interativos, permitindo que os usuários possam visualizar e analisar os dados que necessitam com ferramentas conhecidas.

3.12 Desenvolvimento

Nesta etapa, foram desenvolvidas as aplicações necessárias de acordo com levantamentos realizados na etapa de especificações de aplicações de usuário final que compreendem a seleção do ambiente de desenvolvimento dos relatórios, revisão e desenvolvimento de aplicações padronizadas; verificação da previsão dos dados; desenvolvimento de estruturas de navegação e documentação das aplicações do usuário final, desenvolvimento de procedimentos de manutenção e atualização das aplicações de usuário final; aceitação do projeto pelo usuário final e revisão do projeto.

3.12.1 Sistema de Gestão

O Sistema de Gestão tem como responsabilidade a manutenção dos indicadores obtidos pela Cooperativa; é desenvolvido em *Visual Basic* e é utilizado como ferramenta de *Front End*. A seguir, serão relacionados os tipos de relatórios definidos:

- **Extrato de produção de um Prestador de Serviços** – Discrimina a produção como executante (consultas, visitas hospitalares e procedimentos diagnósticos e terapêuticos), exames solicitados, custos com produção e exames, procedimentos autogerados e indicadores (média de exames, seus custos por consulta e percentual de reconsultas).
- **Analítica de Produção Médica de um Prestador de Serviços** – Oferece a data na qual o serviço (consulta, cirurgia, etc) foi realizado, discriminação do código do serviço, sua quantidade, nome do cliente e valor pago.

- **Analítica de requisições por Prestador de Serviços** – Lista, por datas, os códigos dos serviços, nomes dos clientes, prestadores de serviços e valores pagos.
- **Lista de Clientes por Prestador de Serviços** – Relaciona todos os clientes por nomes, quantidades de serviços e valores pagos.
- **Lista de serviços prestados por um Prestador de Serviços** – Sintetiza, por códigos de serviço da tabela AMB, todos os atendimentos prestados no período desejado.
- **Extrato ambulatorial de um Cliente:** discrimina a data do atendimento, procedimento realizado e sua quantidade (consulta, exame, fisioterapia, etc), identificando o Prestador de Serviços (Médico, Clínica, Laboratório) e custos. Acompanha pesquisa de satisfação.
- **Extrato hospitalar de um Cliente:** dados similares ao extrato ambulatorial, porém relativos à conta hospitalar, também com pesquisa de satisfação.

Nas figuras na pagina seguinte, é apresentado o Sistema de Gestão.

FIGURA 17 - SISTEMA DE GESTÃO – ABERTURA – COOPERATIVA MÉDICA

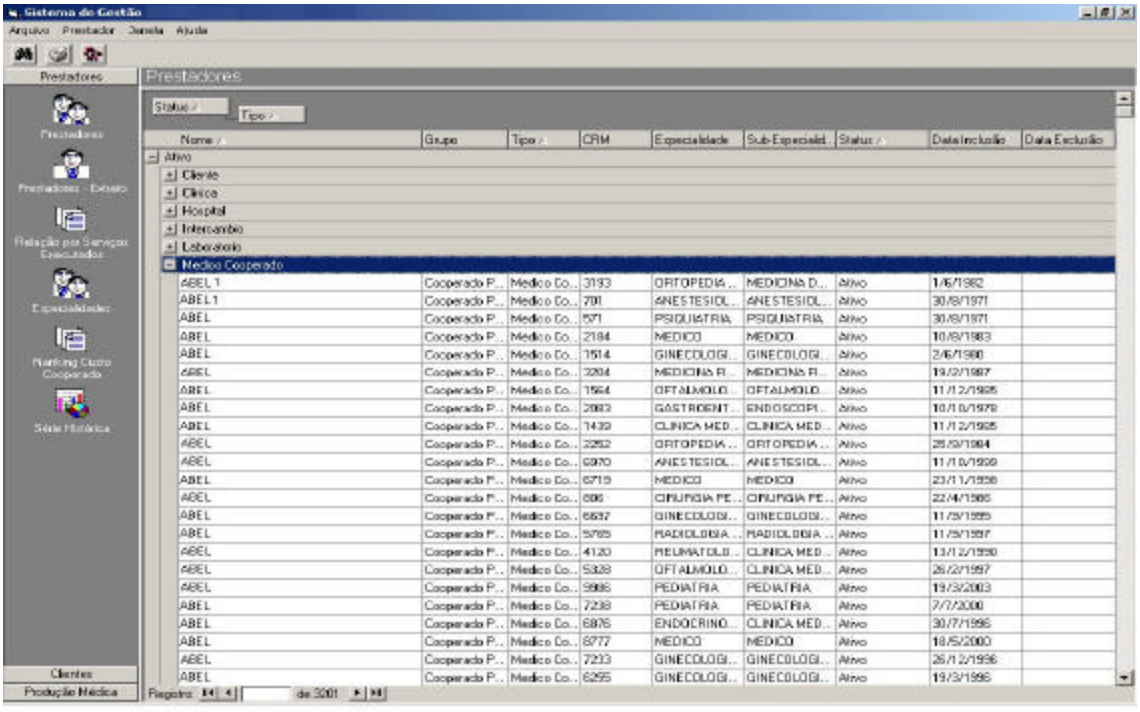

FIGURA 18 - SISTEMA DE GESTÃO – MENU PRINCIPAL – COOPERATIVA MÉDICA

3.13 Manutenção e Crescimento

Quanto a esta etapa, foi observado que seria possível, com os recursos disponíveis em infra-estrutura, assegurar contínua disponibilidade, desempenho e expansão do *Data Warehouse.*

O treinamento dos usuários do *Data Warehouse* é proposto para observar as dificuldades e as visões de cada usuários do *Data Marts.*

3.14 Gerenciamento de Projeto

A etapa "*gerenciamento do projeto Data Mart/Data Warehouse*" consistiu no acompanhamento rigoroso de todas as atividades previstas e a utilização das experiências adquiridas na fase de protótipo.

Esta etapa consiste no acompanhamento contínuo do desenvolvimento do projeto piloto, no monitoramento dos trabalhos realizados e no acompanhamento dos cronogramas.

3.14.1 Usuários *Data Warehouse*

A Diretoria Executiva da Cooperativa identificou, dentro de cada *Data Mart*, quem são os usuários que terão acesso permitido. Para esse grupo foi efetuado um treinamento inicial no acesso aos cubos, através da ferramenta *Front End Excel 2000 OLAP*, juntamente com apresentação dos metadados. Esses usuários também participaram na validação de cada cubo.

3.14.2 Segurança da Informação

Pela facilidade de acesso e uso da ferramenta, os usuários do *Data Warehouse*, durante o treinamento da ferramenta *Front End*, foram alertados sobre o sigilo das informações. Nesse período, foi apresentado pela Diretoria Executiva da Cooperativa o "*Manual de Código de Ética*", o qual trata das regras sobre a segurança da informação. O acesso aos dados foi liberado aos principais decisores de forma irrestrita. A segurança dos dados é realizada através de procedimentos de autorização e autenticação. A proteção dos dados é realizada através da execução da política de backup diário, mensal e semanal.

4 CONCLUSÃO

A tecnologia de *Data Warehouse* mostra-se muito interessante para empresas que possuem grandes volumes de dados gerados e acumulados durante sua existência e necessitam recuperá-los de uma forma que eles possam auxiliar os administradores destas empresas a tomarem decisões estratégicas de forma rápida e segura .

Os processos de extração, filtragem, carga e recuperação dos dados são bastante complexos, exigindo que pessoas altamente capacitadas façam parte do projeto para que os objetivos sejam atingidos no menor espaço de tempo possível e sem gastos de recursos desnecessários.

Um dos pontos fundamentais na elaboração e implantação do *Data Warehouse* é recursos humanos. Todo suporte computacional elaborado e implantado servirá para dar suporte aos usuários, ajudando-os a tomar decisões importantes para a organização. De nada vale realizar um grande investimento apenas no ambiente computacional e não investir nas pessoas que farão este ambiente ganhar utilidade.

Como o *Data Warehouse* não é um sistema ou programa, mas sim um ambiente que necessita ser adaptado à necessidade das empresas, é normal que cada ambiente de *Data Warehouse* possua características próprias, inviabilizando seu uso para outros objetivos que não os descritos no início do projeto.

A modelagem dimensional é um dos fatores críticos de sucesso em um projeto de *Data Warehouse.*

Para Barbieri (2001), em um projeto de *Data Warehouse* os dados fundamentalmente importantes são aqueles consolidados nas dimensões específicas.

Um modelo de dados bem elaborado e de acordo com os objetivos da empresa torna o *Data Warehouse* muito mais consistente e robusto.

Segundo Barbieri (2001), a abordagem de Bill Inmon se concentrou inicialmente no estilo tradicional de construção de Bancos de Dados que busca uma forte integração entre todos os dados da empresa. Isto seria representado num modelo único, integrado e coeso que se mostrou tígido e de difícil consecução. A abordagem de Kimball possui um estilo mais simples, centrada e incremental. Diferente da abordagem anterior, a metodologia esquema estrela, que transforma dados em tabelas de fato e em

tabelas de dimensão. Kimball aponta para projetos de *Data Marts* separados que deverão ser integrados à medida da evolução.

A metodologia proposta por Kimball apóia-se sobremaneira sobre a etapa "experimentação", que se constitui em um grande "laboratório". Nela, ganha-se experiências e elimina-se incertezas do projeto através da realização de uma série de testes intensivos e repetitivos em todos os sentidos (arquitetura, infra-estrutura, produtos, modelagem dimensional etc). Porém, o uso de um estudo prático ajudou a uma maior familiarização e solidificação dos conceitos, que começaram a ser apreendidos no estudo teórico.

Segundo o que foi visto durante esse trabalho, em especial no estudo de caso, a criação de um *Data Warehouse* é uma tarefa bastante trabalhosa, pois o *Data Warehouse* não apenas armazena dados, mas é também um conjunto de ferramentas para consultas, análise e visualização de informações. Desta forma, espera-se que o trabalho que ora está concluído, possa ser disponibilizado e utilizado por empresas que queiram implantar um *Data Warehouse*, e que este venha contribuir efetivamente para o desenvolvimento e crescimento das corporações.

Trabalhos futuros

Como trabalhos futuros, sugere-se várias questões como:

- potencializar o uso do *Data Warehouse* através da socialização. É preciso que o processo de gestão do conhecimento esteja alinhado com os processos de negócio da empresa. Se o conhecimento não for relevante, dificilmente haverá interesse, as pessoas tenderão a não usá-lo, não contribuindo com a empresa e, conseqüentemente, não havendo as atualizações adequadas.
- Desenvolvimento de uma ferramenta para mineração de dados (*Data Mining*, *BI* e *CRM*), visto que a base de dados multidimensional já esta modelada.
- A transferência de dados do ambiente operacional para o ambiente multidimensional é, um dos maiores problemas, uma vez que se pode encontrar diversas fontes para um mesmo dado. É possível desenvolver um mecanismo para transformar e carregar os dados na base multidimensional.

- Aprofundar mais os conhecimentos teóricos de tópicos que não foram abordados ou foram vistos de formas superficial neste trabalho.

5 REFERÊNCIAS BIBLIOGRÁFICAS

BARBIERI, C. **BI-Bussiness Intelligence – Modelagem & Tecnologia**. Rio de Janeiro: Axcel Books do Brasil, 2001.

CAMPOS, Maria Luiza & Filho, ROCHA , Arnaldo V. ***Data Warehouse*** **-** XVII Congresso da Sociedade Brasileira de Computação. 1997.

_____ . ***Data Warehouse***. NCE – UFRJ. Nov., 1998.

CORBA. **Architecture. Common Object Architecture**. Disponível em http://www.omg.org/gettingstarted/corbafaq.htm Acesso em: : 01 out. 2002.

DAL'ALBA, Adriano. **Um estudo sobre *Data warehouse*.** Disponível em www.geocities.com/siliconvalley/port/5072 Acesso em: 15 out. 2002.

DATE, C. **Introdução a sistemas de bancos de dados**. 3.ed. Rio de Janeiro: Campus, 1986.

GARDNER, Stephen R. **Building the *Data Warehouse*. Communications of the ACM**, 41(9): 52-60. Sept., 1998.

GRAY, Paul & Watson, HUGH J. **Decision Support in the *Data Warehouse***. New Jersey, Prentice Hall PTR, 1998.

INMON, W.H. **Como Construir o *Data Warehouse*** .Rio de Janeiro: Campus, 1997.

KIMBALL, Ralph. **The *Data Warehouse* Toolkit**. New York: John Wiley & Sons, 1996.

_____ . ***Data Warehouse* Toolkit** . Ed. Makron Books,1998.

KIMBALL, R. et al. **The Data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses**. New York: John Wiley & Sons, 1998a.

KIMBALL, Ralph. **Data warehouse Toolkit**. São Paulo: Makron Books, 1998b.

POE, Vidette; KLAUER, Patricia & Brobst, Stephen. **Building a *Data Warehouse* for decision support**. New Jersey: Prentice Hall PTR, 1998.

SEN, Aru & Jacob, VARGHESE S. Industrial**. Strenght Data Warehousing** . Communications of the ACM, 41 (9): 29-31, Sept; 1998.

SIL, Abraham S. **Database System Concepts**. 3rd ed. McGraw Hill, 1996.

WELDON, J.L. **Warehouse Cornerstones**. Revista Byte, V22, n.1, janeiro 1997. Disponível em: http://www.byte.com/art/9701/sec7/art2.htm. Acesso em: 02 nov. 2002.

www.microsoft.com/brasil/solucoes. Acesso em: 10 out. 2002.

www.microsoft.com/brasil/sql/parceiros. Acesso em: 10 out. 2002.

www.microsoft.com/brasil/sql/bi. Acesso em: : 10 out. 2002.

www.datawarehouse.inf.br/datawarehousing.asp. Acesso em: 10 out. 2002.

www.relativa.com.br/sol/EmpresaProjetoDWUnimed.asp. Acesso em: 10 out. 2002.

6 GLOSSÁRIO

- *Ad-Hoc Query (consulta eventual):* qualquer consulta que não possa ser determinada antes do momento da consulta ser emitida. Uma consulta que consiste em SQL construído dinamicamente, em geral por ferramentas de consulta residentes na estação de trabalho do usuário final.
- *Banco de dados Relacional*: sistema de banco de dados que suporta todos os comandos SQL padrão.
- *Cubo:* banco de dados multidimensional, geralmente referindo-se a um caso simples de produto, mercado e tempo.
- *Data Mining:* técnica que utiliza ferramentas de software geralmente orientadas para o usuário que não sabe exatamente o que está pesquisando, mas procura identificar determinados padrões ou tendências. O Data Mining (garimpagem de Dados) é um processo que separa grandes quantidades de dados de forma a identificar relacionamentos entre eles.
- *Data Warehouse:* conjunto de tabelas que armazenam os dados dos sistemas de operação (ERPs, tarifadores etc…) em um modelo multidimensional, possibilitando a exploração direcionada dos mesmos, melhorando as possibilidades de análises operacionais e gerenciais.
- *Heurística:* método analítico usado na resolução de problema.
- *Metadados (Metadata):* são dados a respeito de dados. Exemplos de metadados incluem as descrições de elementos de dados, descrições de tipos de dados, atributos/propriedades, faixas/domínios, métodos e processos. O ambiente do repositório abrange todos os recursos de metadados.
- *ODBC (Open Database Connectivity):* conectividade de base de dados aberta. Padrão para acesso a banco de dados do SQL Access Group consortium adotado pela Microsoft.
- *OLAP (On-Line Analytical Processing):* Processamento analítico on-line.
- *Query:* termo que designa uma consulta a um banco de dados.
- *Refresh:* processo de extrair dados de um ambiente e de movê-los para um outro ambiente substituindo, a cada vez, os dados antigos pelos novos.

- *Slice and Dice:* termo usado para descrever a função de analise de dados complexos proporcionada por algumas ferramentas de consulta e analise.
- *SQL (Structure Query Language):* linguagem de consulta estruturada. Linguagem de consulta para acessar sistemas de base de dados ODBC, DRDA ou não relacional.