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

Alessandra Costa Smolenaars Dutra

DESENVOLVENDO DATA WAREHOUSES BASEADOS EM INTRANET

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Vitório Bruno Mazzola

Florianópolis, Fevereiro e 2001

DESENVOLVENDO DATA WAREHOUSES BASEADOS EM INTRANET

Alessandra Costa Smolenaars Dutra

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistema de Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Prof. Vitório Bruno Mazzola, Dr.

Prof. Fernando A. O. Gauthier, Dr.

Banca Examinadora

Prof. Vitório Bruno Mazzola, Dr.

Prof. Murilo Silva de Camargo, Dr.

Prof. Roberto Willrich, Dr.

Prof. Marta Maria Leite

Este trabalho é para
meu marido Florentino, minha
filha Marcela e meu orientador Vitório
que colaboraram com a realização deste,
incentivando-me.

SUMÁRIO

Resumo

O objetivo deste trabalho é estudar e conhecer a tecnologia de datawarehouse para que ele possa ajudar os empresários a descobrir novas formas de competir em uma economia globalizada, trazendo melhores produtos ou serviços para o mercado, mais rápido do que os concorrentes, sem aumentar o custo do produto ou do serviço.

O Data warehose é um banco de dados especializado, o qual integra e gerencia o fluxo de informações a partir dos bancos de dados corporativos e fontes de dados externas à empresa. Não existem ainda metodologias formais para implementação de um datawarehouse, ela deve ser adaptada às características e às expectativas de cada empresa.

Um datawarehouse oferece os fundamentos e os recursos necessários para um Sistema de Apoio a Decisão eficiente, fornecendo dados integrados e históricos que servem desde a alta direção, até as gerências de baixo nível.

Um dos desafios da implantação de um datawarehouse é justamente a integração destes dados, eliminando as redundâncias e identificando informações iguais que possam estar representadas sob formatos diferentes em sistemas distintos.

Neste trabalho foi traçado um histórico dos sistemas de informação, apresentando as principais eras da tecnologia da informação, os seus impactos e suas tendências tecnológicas sobre elas, os conceitos, tipos, objetivos, características e arquiteturas de um Data Warehouse.

São abordados os vários modelos de Data Warehouse que podem ser utilizados no seu desenvolvimento, finalizando com um estudo de caso para o desenvolvimento de um Data Warehouse para a área educativa, tendo como modelo as escolas Yázigi de Florianópolis.

Abstract

The aim of this paper is studying and getting to know the Data Warehouse's technology in order that it can help businessmen to find out new ways of competing with others in a globalized economy, bringing better products or services to the market faster than the competitors without increasing the cost of them.

The Data Warehouse is a specialized data base which integrates and manages the flow of information from corporative data bases and outward data sources. There are not any formal methodologies for the implementation of a Data Warehouse; it still must be adapted to the characteristics and expectations of each company.

A Data Warehouse offers the theory and necessary resources to an efficient System of Decision and Support, giving integrated data and reports that will be helpful not only to the direction's body but all the sections as well.

One of the Data Warehouse's use's challenge is exactly the integration of these data, eliminating redundancies and identifying similar information that can be represented under different aspects distinct systems.

This paper draws a report on information systems, presenting the main eras of the information technology, its impacts, tendencies, concepts, types, objectives, characteristics and architectures of a Data Warehouse.

Also, many Data Warehouse's models that can be used on its development are shown here, ending with a case study for the educational area, having the Yázigi schools from Florianópolis as a model.

1.0 Introdução

Com a evolução da tecnologia de informação e o crescimento do uso de computadores interconectados, praticamente todas as empresas de médio e grande porte estão utilizando sistemas informatizados para realizar seus processos mais importantes, o que com o passar do tempo acaba gerando uma enorme quantidade de dados relacionados aos negócios, mas não relacionados entre si. Estes dados armazenados em um ou mais sistemas operacionais[1] de uma empresa são um recurso, mas de modo geral, raramente servem como recurso estratégico no seu estado original. 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 o apoio ao processo de tomada de decisões das organizações. Estas decisões 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.

Em termos simples, um data warehouse, ou em português, Armazém de Dados, pode ser definido como um banco de dados especializado, o qual integra e gerencia o fluxo de informações a partir dos bancos de dados corporativos e fontes de dados externas à empresa. Um data warehouse é construído para que tais dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais. A função do data warehouse é tornar as informações corporativas acessíveis para o seu entendimento, gerenciamento e uso. Como o data warehouse está separado dos bancos de dados operacionais, as consultas dos usuários não impactam nestes sistemas, que ficam resguardados de alterações indevidas ou perdas de dados. O data warehouse não é como 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.

---

[1] Sistemas Operacionais : Sistemas que são utilizados diariamente dentro de uma empresa.

Um data warehouse oferece os fundamentos e os recursos necessários para um Sistema de Apoio a Decisão (SAD) eficiente, fornecendo dados integrados e históricos que servem desde a alta direção, que necessita de informações mais resumidas, até as gerências de baixo nível, onde os dados detalhados ajudam a observar aspectos mais táticos da empresa. Nele, os executivos podem obter de modo imediato, respostas para perguntas que normalmente não possuem respostas em seus sistemas operacionais e, com isso, tomar decisões com base em fatos, não com intuições ou especulações.

Com o surgimento do data warehouse são necessários novos métodos de estruturação de dados e novas tecnologias, tanto para armazenamento, como para recuperação de informações. A necessidade destes novos métodos e tecnologias surgiu da constatação, primeiro de que existe uma necessidade de informação não atendida pelos aplicativos comerciais convencionais que atuam a nível operacional do negócio, e segundo, pelo fato de que a tecnologia de armazenamento de dados utilizada nestes aplicativos não atende às necessidades detectadas. Graças aos avanços nos bancos de dados relacionais, no processamento paralelo e na tecnologia distribuída, finalmente a tecnologia da informação pode permitir que qualquer organização elabore um data warehouse.

Como as empresas demoram vários anos para gerar e armazenar um volume considerável de informações, é normal que estes dados estejam espalhados por diversos locais e que tenham sido gerados por sistemas desenvolvidos em diferentes ambientes e linguagens. Um dos desafios da implantação de um data warehouse é justamente a integração destes dados, eliminando as redundâncias e identificando informações iguais que possam estar representadas sob formatos diferentes em sistemas distintos.

O objetivo deste trabalho é estudar e conhecer a tecnologia de data warehouse para que ela possa ajudar os empresários a descobrir novas formas de competir em uma economia globalizada, trazendo melhores produtos ou serviços para o mercado, mais rápido do que os concorrentes, sem aumentar o custo do produto ou do serviço. Não existem ainda metodologias formais para implementação de um data warehouse, ela deve ser adaptada às características e às expectativas de cada empresa, mas o principal objetivo em todas elas é o de descobrir maneiras diferentes de atuar no mercado e quais as mudanças internas que devem ocorrer para atender as novas realidades.

No capítulo 2 foi traçado um histórico dos sistemas de informação, apresentando as principais eras da tecnologia da informação, os seus impactos e suas tendências tecnológicas sobre elas.

No capítulo 3 são descritos os componentes fundamentais de Data Warehouse, os principais tipos de Data Warehouse, uma introdução a tecnologias Intranet, os vários conceitos existentes sobre Data Warehouse e Data Warehouse para Intranet.

Ainda neste capítulo são apresentados os objetivos de um Data Warehouse, as suas diversas características e as suas diversas arquiteturas que podem ser utilizadas nos projetos de DW.

No capítulo 4 são abordados os vários modelos de Data Warehouse que podem ser utilizados no seu desenvolvimento.

No capítulo 5 são descritos diversos aspectos sobre o projeto e o desenvolvimento de sistemas de Data Warehouse. Foram citadas as justificativas do projeto, a seleção do modelo de Dados, a identificação das necessidades de Informações para análise e ainda a definição da duração dos dados. Foi feito também um estudo de caso para o desenvolvimento de um Data Warehouse para as escolas Yázigi, especificando suas tabelas e suas funções.

Finalmente, no capítulo 6, são tecidas as conclusões sobre os resultados obtidos e as sugestões para futuros trabalhos relacionados.

2.0 Histórico da Informação

A história da tecnologia de informação nos computadores pode ser classificada em três fases ou eras principais:

- **A era do Hardware**, que começou em 1945 com o desenvolvimento do ENIAC (o primeiro e gigantesco computador eletrônico) e teve como foco a produção de computadores mais rápidos e mais poderosos para o processamento de dados comercialmente.

- **A era do Software**, que inicio em 1975 com a introdução do Altair 8800, o primeiro computador pessoal para o público em geral, e ganhou destaque em 1981 com a introdução do vitorioso e popular computador pessoal da IBM. A nova geração dos computadores pessoais acionou uma demanda por uma nova geração de softwares práticos e de fácil manuseio e distribuição.

A partir desta nova era, os computadores não eram mais exclusivamente de programadores e digitadores. A indústria de software tinha usuários leigos, como empresários ansiosos por informação, que precisavam de aplicativos simples para explorar os recursos dos seus novos computadores.

- **A era do conteúdo**, que representa uma das mudanças mais significativas efetuadas pela Internet, enfatizando a criação e gerenciamento de conteúdo em vez da lógica do aplicativo, além de desenvolver a colaboração e troca de informações entre usuários. A era do conteúdo, no entanto, está dando seus primeiros passos, crescendo e se modificando para responder á demanda por informação da comunidade global. Esta era enfatiza a criação e gerenciamento do conteúdo (dados brutos otimizados por meio da comunicação e compartilhamento) e o aprimoramento da colaboração entre os usuários.

Durante a era do hardware, a tecnologia da informação forneceu um eficiente meio para o processamento de informações de transações comerciais como pedidos, faturas e

reservas. A era do software começou como resultado de mudanças na tecnologia de hardware – modificações que possibilitaram aos fabricantes de hardware a construção e comercialização de computadores 'pessoais' relativamente baratos para o público em geral.

Atendendo à demanda, os fabricantes forneceram aos usuários recursos de processamento de informações 'pessoais' na forma de aplicativos de software, principalmente planilhas e processadores de texto e distribuíram esses aplicativos em disquetes. Com o progresso da era do software, os usuários continuaram a desejar mais e mais aplicativos comerciais e os computadores 'pessoais' ficaram repletos. Ao passo que as corporações 'diminuíram' os sistemas de computadores mainframe, aumentavam continuamente os seus computadores de mesa e suas redes.

O impacto principal da era do software não foi simplesmente a qualificação de uma grande classe de usuários leigos a procura de informações e buscando aplicativos, mas sim o crescimento do 'bloatware'(programas repletos de recursos, dos quais o usuário só usa uma pequena fração) que ocasionou inevitavelmente a era do conteúdo.

A transição para a era do conteúdo começou com a popularização da Internet e da Web. Na era do conteúdo, a atenção se desloca para o conteúdo em lugar da tecnologia de hardware e software utilizada para transmiti-los. Ao contrário da era do software, na qual os usuários ficavam atarefados em carregar aplicativos em disquete (ou CD-ROM) usando então a lógica do aplicativo para acessar ou criar conteúdos úteis, a Internet (e a Web) apresenta primeiramente o conteúdo ao usuário, ocultando o aplicativo. O usuário recebe apenas a lógica do aplicativo necessária para apresentar o conteúdo e retém apenas a lógica do aplicativo requerida por sua estação de trabalho local.

A era do conteúdo enfatiza a criação e o gerenciamento de conteúdo (dados brutos otimizados por meio de comunicação e compartilhamento) e o aprimoramento da colaboração entre os usuários. Embora o correio eletrônico esteja se mostrando o primeiro grande representante da era do conteúdo, a era ainda está em seus primeiros passos e os usuários começam a entender o potencial deste último estágio na evolução da tecnologia de informação. Em muitos aspectos, a tecnologia da Internet não é uma extensão da computação cliente/servidor; representa uma passagem ampla e fundamental

para a próxima era na tecnologia de informações – redes de computadores distribuídas em larga escala.

Duas tendências tecnológicas significativas estão em andamento na era do conteúdo:

1. A ênfase se expandiu de aplicativos OLTP (On-Line Transaction Processing – processamento de transações on-line) para Aplicativos OLAP (On-Line Analytic Processing - Processamento Analítico on-line) e Data warehouse (Armazenamento de Dados).

2. As arquiteturas de sistemas do tipo cliente/servidor estão passando por uma transformação e freqüentemente emergem na forma de Intranets [TAN98].

Essas duas tendências tecnológicas estão interligadas pelo fato de lidarem com os assuntos relacionados ao gerenciamento e a transmissão de informações. O data warehouse reconhece que os dados são importantes ativos que devem ser estruturados para OLAP. As intranets estão criando um novo modelo para comunicações e colaboração no âmbito da empresa. Em conjunto, constituem uma infra-estrutura de informações fundamentalmente diversa para organizações interessadas em alcançar um retorno ótimo para seus investimentos em capital intelectual.

3.0 Data Warehouse

3.1 Conceitos

3.1.1 Data Warehouse

O ambiente de dados para suporte aos processos de gerência e tomada de decisão é fundamentalmente diferente do ambiente convencional de processamento de transações. No coração deste ambiente está a idéia do data warehouse, integrando e consolidando dados disponíveis em diferentes acervos para fins de exploração e análise, ampliando o conteúdo informacional destes acervos para atender as necessidades de nível estratégico da empresa.

O data warehouse pode ajudar as organizações na proteção de seus conjuntos de informações e a tornar os dados mais acessíveis durante a tomada de decisões.

Um data warehouse é uma modalidade de implementação de uma base de dados informal, voltada ao armazenamento de dados compartilhados, obtidos a partir dos ambientes de base de dados operacionais. Trata-se tipicamente de uma base de dados de assuntos que permite ao usuários penetrar o vasto repositório de dados operacionais da empresa para acompanhar e fazer frente às tendências de negócios e facilitar os esforços de previsão de planejamento.

Segundo W.H.Inmon, considerado um pioneiro no tema, "um data warehouse é uma coleção de dados orientada por assuntos, integrada, variante no tempo, e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão" [IHMO92]

Data warehouse é um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas e ad-hoc, relatórios analíticos e de suporte a decisão, conforme Harjinder [HAR96].

Segundo Barquini [BAR96], data warehouse é uma coleção de técnicas e tecnologias que juntas disponibilizam um enfoque pragmático e sistemático para tratar com o problema do usuário final de acessar informações que estão distribuídas em vários sistemas da organização.

Para entender o que é um data warehouse, é importante fazer uma comparação com o conceito tradicional de banco de dados. Conforme [BAT86], "um banco de dados é uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de uma empresa específica". Os dados mantidos por uma empresa são chamados de "operacionais" ou "primitivos". Batini[BAT86] refere-se aos dados no banco de dados como "dados operacionais", distinguindo-se de dados de entrada, dados de saída e outros tipos de dados.

Levando em consideração esta definição sobre dados operacionais, pode-se dizer que um data warehouse é, na verdade, uma coleção de dados derivados dos dados operacionais para sistemas de suporte à decisão. Estes dados derivados são, muitas vezes, referidos como dados "gerenciais", "informacionais" ou "analíticos" [INM96].

Os bancos de dados operacionais armazenam as informações necessárias para as operações diárias da empresa, são utilizados por todos os funcionários para registrar e executar operações pré-definidas, por isso seus dados podem sofrer constantes mudanças conforme as necessidades atuais da empresa. Por não ocorrer redundância nos dados e as informações históricas não ficarem armazenadas por muito tempo, este tipo de BD não exige grande capacidade de armazenamento.

Já um data warehouse armazena dados analíticos, destinados às necessidades da gerência no processo de tomada de decisões. Isto pode envolver consultas complexas que necessitam acessar um grande número de registros, por isso é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. Um data warehouse armazena informações históricas de muitos anos e por isso deve ter uma grande capacidade de processamento e armazenamento dos dados que se encontram de duas maneiras, detalhados e resumidos.

Na Tabela 1 estão relacionadas algumas diferenças entre bancos de dados operacionais e data warehouse bem como as diferenças dos dados que eles manipulam segundo os seguinte autores: [INM96] [BAR96] [KIM96] [ONE97].

Tabela 1 – Diferenças entre banco de dados operacionais e data warehouse

Características	*Bancos de dados Operacionais*	*data warehouse*
Objetivo	Operações diárias do negócio	Analisar o negócio
Uso	Operacional	Informativo
Tipo de processamento	OLTP	OLAP
Unidade de trabalho	Inclusão, alteração, exclusão	Carga e consulta
Número de usuários	Milhares	Centenas
Tipo de usuário	Operadores	Comunidade gerencial
Interação do usuário	Somente pré-definida	Pré-definida e ad-hoc
Condições dos dados	Dados operacionais	Dados Analíticos
Volume	Megabytes – gigabytes	Gigabytes- terabytes
Histórico	60 a 90 dias	5 a 10 anos
Granularidade	Detalhados	Detalhados e resumidos
Redundância	Não ocorre	Ocorre
Estrutura	Estática	Variável
Manutenção desejada	Mínima	Constante
Acesso a registros	Dezenas	Milhares
Atualização	Contínua (tempo real)	Periódica -em batch
Integridade	Transação	A cada atualização
Número de índices	Poucos/simples	Muitos/complexos
Intenção dos índices	Localizar um registro	Aperfeiçoar consultas

Com base nestes conceitos podemos concluir que o data warehouse não é um fim, mas sim um meio que as empresas dispõem para analisar informações históricas podendo utilizá-las para a melhoria dos processos atuais e futuros.

Data warehouses são resumos de dados retirados de múltiplos sistemas de computação normalmente utilizados há vários anos e que continuam em operação.

Data warehouse são construídos para que tais dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais. Os dados de um Data warehouse podem ser compostos por um ou mais sistemas distintos e sempre estarão separados de qualquer outro sistema transacional, ou seja, deve existir um local físico onde os dados desses sistemas serão armazenados.

Os Data warehouses contem informações como avaliações de desempenho operacional e inteligência competitiva que facilitam a tomada de decisões.

3.1.2 Data warehouse Para Intranet

Um Data warehouse para intranet é uma combinação de tecnologias que permite aos usuários gerar dinamicamente uma consulta a um banco de dados, fazer análise de dados e formatar o resultado como arquivos de texto ou de imagem para ser exibido em qualquer navegador (browser). Esta é a chave para a utilidade definitiva de um data warehouse para intranet; não requer software algum além do navegador. O data warehouse e o software de análise são acessados por meio de uma URL(Universal Resource Locator – localizador universal de recursos). As consultas dos usuários ao data warehouse criam relatórios dinamicamente e imagens para apresentação no navegador – proporcionando acesso a informações precisas e objetivas quando e onde o usuário precisar.

O método usado no projeto e gerenciamento de um data warehouse para intranet é o mesmo do data warehouse tradicional. O que distingue o primeiro é a capacidade de atender às necessidades de um grande número de usuários leigos. Neste aspecto, a facilidade de operação, o desempenho, a escalabilidade e a segurança são ampliadas. Os data warehouse projetados para a Intranet têm a possibilidade de atingir um grande número de usuários, cada um com seu grau de conhecimento e sua própria definição de facilidade de uso. Um data warehouse para intranet evolui a partir de uma combinação de bancos de dados distribuídos. O desempenho e a segurança podem ser aprimorados pela criação de data marts (depósito de dados) separados, potencialmente menores.

Além de proporcionar uma solução para o gerenciamento de dados distribuídos na rede, um data warehouse para intranet requer que funções de OLAP sejam também distribuídas na rede, um data warehouse para intranet requer que funções de OLAP sejam também distribuídas. Apesar de todas as vantagens, um data warehouse para intranet engloba desafios técnicos e de gerenciamento significativos que podem impedir uma percepção completa, por parte das organizações, de suas vantagens potenciais. Proporcionar aos executivos informações oportunas, precisas e completas com os quais tomar decisões mais rápidas e bem informadas não é tarefa fácil.

3.2 Componentes Fundamentais

Três tecnologias fundamentais estão convergindo para formar o fundamento de uma nova infra-estrutura de informação com a finalidade de acelerar o processo de tomada de decisão. O investimento nestas tecnologias ajuda a montar o capital intelectual necessário para que a sua empresa se sobressaia em relação à sua concorrência:

1. Armazenamento de Dados: a criação de um repositório completo e preciso de dados. Contêm o conteúdo estruturado basicamente em linhas e colunas de um banco de dados.

2. Processamento Analítico Online (OLAP): Proporciona as ferramentas necessárias para o acesso e análise de dados. As funções OLAP incluem consulta e relatório, análise multidimensional, análise estatística e garimpagem de dados (um tipo de exploração de dados).

3. Tecnologias Internet, especificamente Intranets: Desenvolvem a comunicação e a colaboração no âmbito da empresa. As intranets são precursoras de uma mudança no paradigma das redes de computadores, redefinindo o desenvolvimento de aplicativos cliente/servidor e estratégias de disposição de recursos.

Ampliando cada componente básico temos:

3.2.1 Componente Data Warehouse

Há três tipos de data warehouse que resultam das necessidades de apoio às decisões do usuário e dos aspectos comerciais típicos:

- Financeiro
- De Marketing
- Comportamentais

3.2.1.1 Os data warehouses financeiros

Os data warehouse financeiros monitoram o desempenho comercial em termos financeiros. Contêm históricos financeiros, que são acessados rapidamente, em geralmente dados sobre receitas e despesas. Estes dados são atualizados mensalmente ou por outro período de relatório que coincida com o calendário financeiro da empresa.

3.2.1.2 Os data warehouses de marketing

Os data warehouse de marketing são projetados para permitir que os usuários[2] avaliem o desempenho comercial de um produto ou serviço de múltiplos ângulos. Este tipo de informação é atualizado com freqüência, em geral semanalmente, e em muitos casos diariamente.

O data warehouse de marketing permite ao usuário a análise de dados em vários níveis hierárquicos em cada dimensão do banco de dados. As exigências da análise são muito complexas e altamente variáveis devido à impossibilidade quase total de previsão das perguntas do usuário.

3.2.1.3 Os data warehouse comportamentais

Os data warehouse comportamentais são usados em aplicativos classificados genericamente como "banco de dados de relacionamento", contendo informações sobre os clientes e seus hábitos.

Os data warehouse comportamentais são facilmente encontrados em seguradoras, instituições financeiras, companhias aéreas e empresas de seguro-saúde, entre outras - todas com uma necessidade principal de compreender os hábitos de seus clientes e de determinar como oferecer produtos e serviços que satisfaçam às necessidades individuais destes.

---

[2] Usuários : Pessoas que utilizam o sistema de informática em uma empresa/organização.

3.2.2 Componente On Line Analytic Processing (OLAP)

O OLAP gira em torno de quatro habilidades diferentes:

- Consulta e Geração de Relatórios
- Análise Multidimensional
- Análise Estatística
- Data Mining (garimpagem de dados)

A classe de ferramenta de Consulta e Relatório do OLAP permite ao usuário formular consultas a banco de dados sem precisar interagir com a linguagem de programação do banco de dados SQL. Funções OLAP mais complexas englobam uma análise de dados multidimensional, que inclui um conjunto robusto de capacidades computacionais e de navegação nos dados.

A análise estatística representa o próximo nível de complexidade OLAP. Ela tenta reduzir uma grande quantidade de dados a uma relação simples freqüentemente exposta como uma fórmula matemática.

O Data Mining é o tipo mais complexo de função analítica OLAP, usa sofisticados modelos de reconhecimento de padrões e algoritmos de aprendizado para identificar relações entre elementos de dados. O data mining projeta problemas não-lineares com grande números de variáveis, análise multiautomática, usando técnicas como algoritmos de árvores de decisões, rede neurais, lógica difusa e algoritmos genéricos.

As funções OLAP se situam entre o data warehouse e os componentes de apresentação/interface com o usuário de um aplicativo. As tecnologias da intranet e de navegação proporcionam os subsídios para a criação da interface com o usuário e das funções de apresentação que devem interagir com as funções OLAP. As próprias funções OLAP ligam-se ao data warehouse para recuperar os dados brutos necessários à análise e execução da análise de dados.

3.2.3 Tecnologias Intranet

As Intranets são variantes empresariais da tecnologia Internet que operam sobre redes internas TCP/IP e são separadas de redes públicas por firewalls. Embora as intranets possam incorporar o acesso à Internet, as intranets são privadas e seguras. Oferecem às empresas o potencial para adquirir enormes ganhos de velocidade com a qual a informação é distribuída dentro de uma empresa. Todos os benefícios da tecnologia Internet se aplicam as intranets, que também tiram vantagens do maior desempenho de arquiteturas de redes internas privadas. As intranets oferecem vantagens significativas em termos de custo e um modelo de distribuição de software que pode simplificar enormemente o desafio inerente ao suporte de um grande número de usuários.

3.3 Objetivos de um Data Warehouse

1. O data warehouse fornece acesso a dados corporativos ou organizacionais;

Acesso significa várias coisas. Os gerentes e analistas de uma organização devem poder conectar o data warehouse a partir de seu computador pessoal. Essa conexão deve ser imediata, quando solicitada e com alto desempenho. Acesso por meio de outra pessoa ou acesso não confiável e lento são inaceitáveis. Um acesso de alto desempenho significa que as menores consultas são executadas em menos de um segundo. Acesso significa também que as ferramentas disponíveis aos gerentes e analistas devem ser fáceis de usar.

2. Os dados do data warehouse são consistentes;

Consistência significa que quando duas pessoas solicitam o resultado das vendas na região Sudoeste em Janeiro, devem obter o mesmo número. Consistência também significa que quando essas pessoas solicitarem ao data warehouse a definição do elemento de dado "vendas", recebam uma resposta útil especificando o que elas estão recuperando do banco de dados. Além disso, consistência significa que se os dados de ontem não forem totalmente carregados, o analista deve ser avisado de que a carga de dados não está completa e que provavelmente será concluída amanhã.

3. Os dados no data warehouse podem ser separados e combinados usando-se qualquer medição possível no negócio.

4. O data warehouse não consiste apenas em dados, mas também em um conjunto de ferramentas para consultar, analisar e apresentar informações.

Os componentes que especificam o hardware central do data warehouse, o software do banco de dados relacional e os dados propriamente ditos, representam

apenas 60% do que é necessário para um data warehouse bem sucedido. Os 40% remanescentes consistem no conjunto de ferramentas de front-end que consultam, analisam e apresentam os dados.

5. O data warehouse é o local em que publicamos dados confiáveis.

A responsabilidade de publicar é o âmago do data warehouse. Os dados não são simplesmente acumulados em um ponto central e depois liberados. Ao contrário, os dados são cuidadosamente coletados em várias fontes de informação, limpos, têm sua qualidade assegurada e então são liberados somente se forem adequados ao uso. Se os dados não forem confiáveis ou estiverem incompletos, o gerente de qualidade dos dados não permitira que eles fossem publicados para a comunidade de usuários. O gerente de qualidade de dados desempenha o mesmo papel que um editor de uma revista ou de um livro. É responsável pelo conteúdo e qualidade da publicação e por sua liberação.

6. A Qualidade dos dados no data warehouse impulsiona a reengenharia de negócios.

O data warehouse não pode aprimorar dados de baixa qualidade. Para aprimorar dados de baixa qualidade, os responsáveis pela entrada de dados e a gerência devem retornar à fonte dos dados com sistemas melhores, com gerenciamento melhor e com uma melhor visibilidade dos dados com valores adequados. Curiosamente, uma boa forma para justificar um projeto de reengenharia como esse é publicar os dados incompletos e deixar que surja uma pressão natural da organização, quando as pessoas perceberem o quanto esses dados poderiam ser valiosos se fossem de melhor qualidade. Desse modo, o data warehouse pode desempenhar um papel-chave nos esforços de reengenharia em uma organização.

3.4 Características do Data warehouse

Os dados de um DW devem ser classificados por assunto, além disso é importante que se faça à integração (normalização) de representação para facilitar as consultas, também se deve definir a granularidade temporal da informação e a forma de armazenar os dados, ter consciência de que dados em um DW não são modificados pois representam as informações em um determinado instante de tempo e podem estar fisicamente armazenados de diferentes formas. Essas são as principais características do DW, as quais são apresentadas no conceito do W.H.Inmon [INM97] e serão detalhadas a seguir.

3.4.1 Data warehouse orientado por assuntos

Um data warehouse orientado por assuntos refere-se ao fato do data warehouse armazenar informações sobre assuntos específicos importantes para o negócio da empresa, como por exemplo: bancos, clientes, contas corrente, contas de poupança, produtos, etc. Em contrapartida, o ambiente da empresa é organizado por aplicações funcionais. Por exemplo, em uma organização bancária, estas aplicações incluem empréstimos, investimentos e seguros.

Um DW sempre armazena dados importantes sobre temas específicos da empresa e conforme o interesse das pessoas que irão utilizá-lo. Uma empresa pode trabalhar com vendas de produtos alimentícios no varejo e o seu maior interesse ser o perfil de seus compradores, então o DW será voltado para as pessoas que compram seus produtos e não para os produtos que ela vende.

Portanto, ao se construir um DW deve-se discutir com o usuário quais os seus objetivos, definir as informações relevantes para o processo de análise, além de se preocupar com os tipos de análise que serão realizadas sobre os dados do DW.

3.4.2 Integrado

Esta é a característica mais importante do DW, pois é ela quem irá definir a representação única para os dados provenientes dos diversos sistemas que formarão a base de dados do DW. A maior parte do trabalho na construção de um DW está na análise dos sistemas operacionais e dos dados que eles contêm. Como não existem padrões de codificação, cada analista pode definir a mesma estrutura de dados de várias formas, fazendo com que dados que representam a mesma informação sejam representados de diversas maneiras dentro dos sistemas utilizados pela empresa ao longo dos anos.

Os dados devem ser organizados para fornecer uma fonte única de informação. Por exemplo, uma companhia de seguros pode ter informação sobre diferentes apólices do mesmo agente armazenada em diversos bancos de dados utilizando tecnologias radicalmente diferentes. Para tomar decisões efetivas sobre o relacionamento total com os clientes, os dados precisam ser apresentados em um formato comum. Além disso, se a companhia está para tomar decisões referentes a lucros, ela deve concordar em regras comuns de negócios tais como a mensuração da lucratividade.

3.4.3 Variante no tempo

Segundo W.H.Inmon [INM97] todos os dados no DW são precisos em algum instante no tempo, como eles podem estar corretos somente em um determinado momento, é dito que esses dados "variam com o tempo".

A variação no tempo pode apresentar-se de três maneiras:

1. Em um DW é normal que as informações sejam representadas em horizontes de tempo maiores de cinco anos chegando até o limite da idade dos dados ou em um período considerado satisfatório conforme a sua aplicação. Nas aplicações operacionais o período de tempo é muito mais curto, variando entre sessenta e noventa dias, pois é necessária um resposta rápida às exigências das tarefas diárias o que só pode ser conseguido com o processamento de poucos dados;

2. Assim como os dados, os metadados, que incluem definições dos itens de dados, rotinas de validação, algoritmos de derivação, etc. também possuem elementos temporais para que com eventuais mudanças nas regras do negócio a empresa não perca dados históricos;

3. Os dados armazenados corretamente no DW não serão mais atualizados tendo-se assim uma imagem fiel da época em que foram gerados.

Os dados ainda podem ser separados em duas categorias, a de dados atuais e de dados antigos.

Os dados detalhados atuais são os dados de maior interesses por refletir os acontecimentos mais recentes, são em grande volume porque são armazenados no nível mais baixo de granularidade e devem ficar armazenados em disco, um meio de acesso rápido mas de difícil gerenciamento. Os dados detalhados atuais fornecem uma visão do comportamento recente e podem permitir a utilização de técnicas como mineração de dados e descoberta de conhecimento. O horizonte de tempo, para esses dados, normalmente é de dois anos.

Os dados detalhados antigos são aqueles que não são acessados freqüentemente e por isso normalmente ficam armazenados em meios de armazenamento de baixo custo pois possuem um grande volume por ficarem em um nível de detalhe consistente com os dados detalhados atuais. Mesmo não ficando armazenados em outros meios, como fitas por exemplo, eles continuam fazendo parte do DW e podem ser carregados sempre que surgir necessidade de extrair informações.

Uma definição que deve ser feita é sobre o período de atualização dos dados que se refere ao tempo necessário para que uma alteração sobre dados do ambiente operacional reflita no DW. Um vez que os dados tenham sido colocados no ambiente operacional, as alterações precisam ser passadas para o DW, o problema é definir de quanto em quanto tempo isto deve ocorrer. Inmon [INM97] sugere que pelo menos 24 horas devem se passar entre o momento em que a alteração é observada pelo ambiente operacional e sua repercussão no DW.

Existem algumas razões para a existência deste intervalo, a primeira delas consiste no fato de que quanto mais rigidamente o ambiente operacional for emparelhado com o DW, mais dispendiosa e complexa será a tecnologia necessária. A segunda é que este

intervalo de tempo possibilita a estabilização dos dados, diminuindo a chance de o DW receber informações incorretas.

3.4.4 Não Volatilidade

Significa que o data warehouse permite apenas a carga inicial dos dados e consultas a estes dados, o chamado ambiente "load-and-access". Após serem integrados e transformados, os dados são carregados em bloco para o data warehouse, para que estejam disponíveis aos usuários para acesso. No ambiente operacional, ao contrário, os dados são, em geral, atualizados registro a registro, em múltiplas transações. Um data warehouse não requer este grau de controle típico dos sistemas orientados a transações.

Como descrito acima, em um DW não existem alterações de dados, somente a carga inicial e as consultas posteriores. Ele é definido assim pois as operações a nível de registro em modo on-line como são os sistemas transacionais, exigem um controle e um processamento muito grande, fugindo do objetivo principal do DW. Segundo W.H.Inmon[INM97] dizer que existe redundância de dados entre os sistemas transacionais e o DW demonstra a falta de conhecimento de como as coisas acontecem no DW.

Deve-se considerar que os dados passam por filtros antes de entrar no DW, com isso muitos dados nunca passam do ambiente transacional e outros são resumidos de certa forma que não são encontrados fora do DW. "Em outras palavras, a maior parte dos dados é física e radicalmente alterada quando passam a fazer parte do DW. Do ponto de vista de integração, não são mais os mesmos dados do ambiente operacional. À luz destes fatores, a redundância de dados entre os dois ambientes raramente ocorre, resultando em menos de um por cento de duplicações."[INM97].

3.4.5 Localização

Os dados podem estar fisicamente armazenados de três formas[CAM97]:

1. Armazenados em um único local centralizando o banco de dados em um DW integrado, procurando maximizar o poder de processamento e agilizando a busca dos dados;

2. Distribuídos por áreas de interesse, o que pode ser chamado de arquitetura federativa, com dados financeiros em um servidor, dados de marketing em outro e dados de manufatura em um terceiro lugar;

3. Armazenados por níveis de detalhes em que as unidades de dados são mantidas no DW. Pode-se armazenar dados altamente resumidos em um servidor, dados resumidos em um nível de detalhe intermediário em um segundo servidor e os dados mais detalhados (atômicos) em um terceiro servidor. Os servidores da primeira camada podem ser otimizados para suportar um grande número de acessos e um baixo volume de dados enquanto servidores nas outras camadas podem ser adequados para processar grandes volumes de dados mais baixo número de acessos.

Um DW pode possuir diferentes níveis de dados, que podem estar agrupados por idade, sintetização ou detalhe. Os componentes da estrutura são divididos em:

- Dados detalhados atuais
- Dados detalhados antigos
- Dados levemente resumidos
- Dados altamente resumidos

Para mudar de nível é necessário que ocorra um dos seguinte eventos: os dados são sintetizados, arquivados ou eliminados.

O processo de sintetização interage no nível mais alto de detalhamento (dados detalhados atuais) para os níveis seguintes (levemente e altamente resumidos). Quando termina determinado período de tempo (semana, mês, trimestre, ano), os dados são indexados por estes períodos e armazenados nos seus respectivos níveis de detalhamento. Para facilitar o acesso aos dados estes devem estar sintetizados e indexados de várias maneiras, portanto ao mesmo tempo em que ocorre o agrupamento por datas também pode ocorrer a sintetização por grupos e subgrupos.

Cada nível possui um horizonte de tempo definido para a permanência dos dados, então o fato dos dados serem transportados para níveis mais elevados não implica na sua exclusão do nível anterior. Um processo denominado processo de envelhecimento ocorre quando este limite é ultrapassado e então os dados podem ser transferidos para meios de armazenamentos alternativos ou passar de dados detalhados atuais para dados detalhados antigos.

3.4.6 Credibilidade dos Dados

A credibilidade dos dados é o mais importante para o sucesso de qualquer projeto. Discrepâncias simples de todo tipo podem causar sérios problemas quando se quer extrair dados para suportar decisões estratégicas para o negócio das empresas. Dados não dignos de confiança podem resultar em relatório inúteis, que não têm importância alguma, como uma lista de pacientes do sexo masculino e grávidos, por exemplo. "Se você tem dados de má qualidade e os disponibiliza em um DW, o seu resultado final será um suporte a decisão de baixo nível com altos riscos para o seu negócio", afirma Robert Craig [IDG98], analista do Hurwitz Group.

Coisas aparentemente simples, como um CEP errado, podem não ter nenhum impacto em uma transação de compra e venda, mas podem influir nas informações referentes à cobertura geográfica, por exemplo. "Não é apenas a escolha da ferramenta certa que influi na qualidade dos dados", afirma Richard Rist [IDG98], vice-presidente

Data Warehousing Institute. Segundo ele, conjuntos de coleções de dados, processos de entrada, metadados e informações sobre a origem dos dados, são de muita importância.

Outras questões como a manutenção e atualização dos dados e as diferenças entre dados para bancos transacionais e para uso em data warehouse também são crucias para o sucesso dos projetos. Além das camadas do DW propriamente dito, tem-se a camada dos dados operacionais, de onde os dados mais detalhados são coletados. Antes de fazer parte do DW estes dados passam por diversos processos de transformação para fins de integração, consistência e acurácia.

A Tabela 2 descreve um conjunto das características normalmente utilizadas para verificar a qualidade dos dados e indica algumas das maneiras de medir o nível da qualidade dos dados do DW. Nem todas as características da Tabela 2 precisam necessariamente ser averiguadas, deve-se escolher as que representam maior fator de risco para o ambiente proposto e trabalhar em cima destas características.

Características da Qualidade de Dados	*Descrição*	*Exemplo de Medida*
Precisão	Grau de Informações que estão corretas	Percentual de Correção
Abrangência	Grau de dados requisitados e atendidos	Percentual de Atendimentos
Consistências	Consistência de dados/liberdade de contradição	Percentual de Condições satisfeitas
Coerências	Coerência lógica que permite criar relações entre os dados.	Percentual de regras de integridade referencial suportadas
Tempo de Resposta	Tempo entre o pedido de consulta e a resposta	Relação entre a complexidade e o tempo de reposta
Singularidade	Singularidade dos dados de mesma natureza	Percentual dos dados que têm valores dentro dos domínios de valores permitidos

Tabela 2 – Conjunto de característica da qualidade de dados.

3.4.7 Granularidade

Granularidade diz respeito ao nível de detalhe ou de resumo contido nas unidades de dados existentes no DW. Quanto maior o nível de detalhes, menor o nível de granularidade. O nível de granularidade afeta diretamente o volume de dados armazenado no DW e ao mesmo tempo o tipo de consulta que pode ser respondida.

Quando se tem um nível de granularidade muito alto o espaço em disco e o número de índices necessários se tornam bem menores, porém há uma correspondente diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas.

A Tabela 3 exemplifica o conceito acima utilizando os dados históricos das vendas de um produto, um nível de granularidade muito baixo pode ser caracterizado pelo armazenamento de cada uma das vendas ocorridas para este produto e um nível muito alto de granularidade seria o armazenamento do somatórios das vendas ocorridas por mês.

Primeira Camada – Dados Resumidos Por Produto
Produto A1 – Maio de 1998 Valor Total de Vendas R$1.270,*00*
Produto A2 – Junho de 1998 Valor Total de Vendas R$ 4.543,00

Tabela 3 – Níveis de granularidade.

Com um nível de granularidade muito baixo, é possível responder a praticamente qualquer consulta, mas uma grande quantidade de recursos computacionais é necessária para responder perguntas muito específicas. No entanto, no ambiente de DW, dificilmente um evento isolado é examinado, é mais comum ocorrer à utilização de uma visão de conjunto dos dados.

Os dados levemente resumidos compreendem um nível intermediário na estrutura do DW, são derivados do detalhe de baixo nível encontrado nos dados detalhados atuais. Este nível do DW é quase sempre armazenado em disco. Na passagem para este nível os dados sofrem modificações, por exemplo, se as informações nos dados detalhados atuais são armazenadas por dia, nos dados levemente resumidos estas informações podem estar

armazenadas por semanas. Neste nível o horizonte de tempo de armazenamento normalmente fica em cinco anos e após este tempo os dados sofrem um processo de envelhecimento e podem passar para um meio de armazenamento alternativo.

Os dados altamente resumidos são compactos e devem ser de fácil acesso, pois fornecem informações estatísticas valiosas para os Sistemas de Informações Executivas, enquanto que nos níveis anteriores ficam as informações destinadas aos Sistemas de Apoio a Decisão (SAD) que trabalham com dados mais analíticos procurando analisar as informações de forma mais ampla.

O balanceamento do nível de granularidade é um dos aspectos mais críticos no planejamento de uma DW, pois na maior parte do tempo, há uma grande demanda por eficiência no armazenamento e no acesso aos dados, bem como pela possibilidade de analisar dados em maior nível de detalhes. Quando uma organização possui grandes quantidades de dados no DW, fazem sentido pensar em dois ou mais níveis de granularidade na parte detalhada dos dados. Na realidade, a necessidade de existência de mais de um nível de granularidade é tão grande que a opção de projeto que consiste em duplos níveis de granularidade deveria ser o padrão para quase todas as empresas.

Primeira Camada – Dados Resumidos Por Produto	*Segunda Camada – Dados Detalhados Por Produto*
Produto A1 – Maio de /1998	02/5/1998- Valor R$ 100,00 – Quantidade 20
Valor Total R$1.270,*00*	09/5/1998- Valor R$ 50,00 – Quantidade 10
	12/5/1998- Valor R$ 125,00 – Quantidade 25
	20/5/1998- Valor R$ 350,00 – Quantidade 70
	22/5/1998- Valor R$ 110,00 – Quantidade 22
	29/5/1998- Valor R$ 320,00 – Quantidade 64

Tabela 4 – Níveis duplos de granularidade para dados resumidos.

O chamado nível duplo de granularidade, ilustrado na Tabela 4, se enquadra nos requisitos da maioria das empresas. Na primeira camada de dados ficam os dados que fluem do armazenamento operacional e são resumidos na forma de campos apropriados

para a utilização de analistas e gerentes. Na segunda camada, ou nível de dados históricos, ficam todos os detalhes vindos do ambiente operacional, como há uma verdadeira montanha de dados neste nível, faz sentido armazenar os dados em um meio alternativo como fitas magnéticas.

Com a criação de dois níveis de granularidade no nível detalhado do DW, é possível atender a todos os tipos de consultas, pois a maior parte do processamento analítico dirige-se aos dados levemente resumidos que são compactos e de fácil acesso e para ocasiões em que um maior nível de detalhe deve ser investigado existe o nível de dados históricos. O acesso aos dados do nível histórico de granularidade é caro, incômodo e complexo, mas caso haja necessidade de alcançar esse nível de detalhe, lá estará ele.

3.4.8 Os Metadados

Metadados são normalmente definidos como "dados sobre os dados". Podem ser definidos também como uma abstração dos dados, ou dados de mais alto nível que descrevem dados de um nível inferior. Os metadados têm um papel muito importante na administração de dados, mas no DW podem ser considerados de suma importância pois, é a partir deles que as informações serão processadas, atualizadas e consultadas.

Como os usuários de DW procuram por fatos não usuais e relações não conhecidas previamente eles precisam examinar os dados e para isso necessitam conhecer a estrutura e o significado dos dados do DW, o que não ocorre em um ambiente operacional onde os usuários trabalham com aplicações que contém as definições de dados embutidas e simplesmente interagem com as telas do sistema sem precisar conhecer como os dados são mantidos pelo banco de dados.

Geralmente os metadados em um DW podem ser apresentados em três camadas diferentes:

1. Metadados operacionais: definem a estrutura dos dados mantidos pelos bancos operacionais, usados pelas aplicações de produção da empresa;

2. Metadados centrais do DW: são orientados por assunto e definem como os dados transformados devem ser interpretados, incluem definições de agregação e campos calculados, assim como visões sobre cruzamentos de assuntos;

3. Metadados do nível do usuário: organizam os metadados do DW para conceitos que sejam familiares e adequados aos usuários finais;

Os metadados podem ser classificados conforme a classe de seus componentes:

1. Mapeamento: descrevem como os dados de sistemas operacionais são transformados antes de entrarem no DW. Exemplos desta classe de metadados podem ser os que identificam campos fontes, mapeamentos entre atributos, conversões, codificações, padrões, etc.;

2. Histórico: com a evolução dos sistemas operacionais as regras de negócio da empresa podem mudar, cabe a estes metadados manter o histórico de mudanças destas regras, pois as regras certas devem ser aplicadas aos dados certos;

3. Miscelânea: esta classe define diversos tipos de metadados, informações da situação de desenvolvimento de partes do DW, informações sobre volume dos dados para estimativas de tempo e recursos, etc.;

4. Algoritmos de sumarização: mostram a relação entre os diferentes níveis de detalhes dos dados, indicando inclusive que nível de sumarização é mais adequado para um dado objetivo;

5. Padrões de acesso: mantém informações sobre freqüência e tipo de acesso aos dados.

Conforme visto anteriormente os dados sobre desempenho e monitoramento também são qualificados como metadados. Eles podem ser criados por processos que

monitoram atividades como extração, carga e uso dos dados. Dados que identificam questões relativas à qualidade dos dados também devem estar disponíveis para os usuários, afim de que estes identifiquem a acurácia de suas análises.

Segundo Inmon[INM97] os metadados englobam o DW e mantém informações sobre "o que está aonde" no DW. Tipicamente os aspectos sobre os quais os metadados mantêm informações são:

- A estrutura dos dados segundo a visão do programador;
- A estrutura dos dados segundo a visão dos analistas de SAD;
- A fonte de dados que alimenta o DW;
- A transformação sofrida pelos dados no momento de sua migração para o DW;
- O modelo de dados;
- O relacionamento entre o modelo de dados e o DW;
- O histórico das extrações de dados.

3.4.9 Fontes de metadados

Os metadados podem ser "encontrados" em vários locais durante o desenvolvimento de um DW. Em [ADE97] alguns tipos de metadados são classificados conforme suas fontes, essas fontes e o tipo de metadados que pode ser obtido através delas são:

1. Repositório de ferramentas CASE: Os dados contidos em ferramentas CASE,geralmente são estruturados, o que facilita a integração automática entre a origem dos metadados e o repositório do ambiente de Data warehouse. Pode-se extrair

informações sobre a origem dos dados, o fluxo dos dados (os processos que utilizam e transportam os dados), o formato dos dados e as definições de negócios.

2. Documentação do desenvolvimento dos sistemas operacionais: o tipo de metadados potencialmente disponível é idêntico ao item acima. A diferença é que normalmente a documentação de desenvolvimento dos sistemas não está estruturada o que pode dificultar o entendimento das origens e fluxos dos dados.

3. Código fonte dos sistemas operacionais: quando não existe uma documentação eficiente dos sistemas operacionais, é possível extrair as informações sobre eles através dos programas fontes. Como vasculhar todos os programas de um ou vários sistemas operacionais a procura de regras é um trabalho demorado e oneroso é possível simplesmente utilizá-los como forma de esclarecer dúvidas que a documentação não contempla, também cobre os mesmos tipos de informações das fontes anteriores.

4. Entrevistas: apesar de não ser uma fonte estruturada de informações, entrevistar profissionais da empresa que entendam do negócio, como gerentes e analistas, é de vital importância. Destas entrevistas pode se obter regras e informações que não estão explícitas na documentação dos sistemas como, requisitos para teste dos dados e indicadores de qualidade dos dados.

5. O próprio ambiente do DW: informações tais como freqüência de acesso às informações, em que nível de agregação, tempo de resposta de cada consulta, auditoria de acesso de informações por usuários, são informações interessantes de se manter, que podem ser geradas pelo próprio sistema ao longo de sua utilização, podendo ser usadas, dentre outros propósitos, para a criação de estruturas de metadados.

3.5 Arquitetura do Data Warehouse

CONSULTAR

Para ser útil o DW deve ser capaz de responder a consultas avançadas de maneira rápida, sem deixar de mostrar detalhes relevantes à resposta. Para isso ele deve possuir uma arquitetura que lhe permita coletar, manipular e apresentar os dados de forma eficiente e rápida. Mas construir um DW eficiente, que servirá de suporte a decisões para a empresa, exige mais do que simplesmente descarregar ou copiar os dados dos sistemas atuais para um banco de dados maior. Deve-se considerar que os dados provenientes de vários sistemas podem conter redundâncias e diferenças, então antes de passá-los para o DW é necessário aplicar filtros sobre eles.

O estudo de uma arquitetura permite compreender como o DW faz para armazenar, integrar, comunicar, processar e apresentar os dados que os usuários utilizarão em suas decisões. Um DW pode variar sua arquitetura conforme o tipo de assunto abordado, pois as necessidades também variam de empresa para empresa. É possível definir uma arquitetura genérica onde praticamente todas as camadas necessárias são apresentadas, conforme a arquitetura genérica vista a seguir, ou arquiteturas que utilizam somente algumas das camadas definidas como as arquiteturas em duas e três camadas e a arquitetura segundo valente, por fim, pode-se definir uma arquitetura baseada na origem dos dados e no fluxo que eles seguem pelo DW, como a arquitetura definida por Chaudhuri.

3.5.1 Arquitetura Genérica de Data Warehouse

Na figura 1 é descrita uma arquitetura genérica proposta por Orr[ORR96]. Esta descrição genérica procura apenas sistematizar papéis no ambiente de DW, 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 de um DW 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 DW formam o centro da arquitetura e são elas as responsáveis por manter e distribuir os dados. A camada de acesso à informação é formada por ferramentas que possibilitam os usuários extrair informações do DW. Todas as camadas desta arquitetura interagem com o dicionário de dados (metadados) e com o gerenciador de processos.

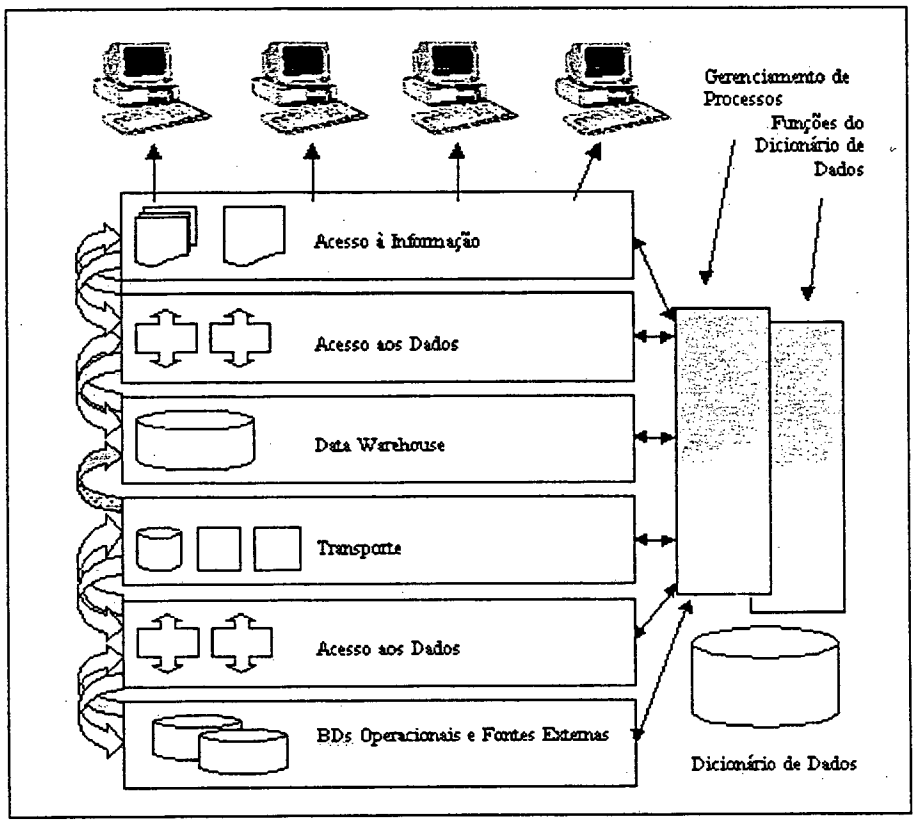

Figura 1 : Arquitetura Genérica de um Data warehouse

1.Camadas de bancos de dados operacionais e fontes externas: É composto pelos dados dos sistemas operacionais das empresas e informações provenientes de fontes externas que serão integradas para compor o DW;

2.Camada de acesso aos dados: Esta camada faz a ligação entre as ferramentas de acesso à informação e os bancos de dados operacionais. Esta camada se comunica com

diferentes sistemas de bancos de dados, sistemas de arquivos e fontes sob diferentes protocolos de comunicação, o que se chama acesso universal de dados;

3.Camada de transporte: Esta camada 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;

4.Camada do Data Warehouse: É o DW propriamente dito, corresponde aos dados utilizados para obter informações.

5.Camada de acesso à informação: Envolve o hardware e o software utilizado para obtenção de relatórios, planilhas, gráficos e consultas. É nesta camada que os usuários finais interagem com o DW, utilizando ferramentas de manipulação, análise e apresentação dos dados, incluindo-se as ferramentas de data-mining e visualização;

6.Camada de metadados (Dicionário de dados): Metadados são as informações que descrevem os dados utilizados pela empresa, isto envolve 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 DW para que ele mantenha sua funcionalidade e os usuários não precisem se preocupar onde residem os dados ou a forma com que estão armazenados;

7.Camada de gerenciamento de processos: É a camada responsável pelo gerenciamento dos processos que contribuem para manter o DW 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 DW;

Às vezes o DW pode ser simplesmente uma visão lógica ou virtual dos dados, podendo não envolver o armazenamento dos mesmos ou armazenar dados operacionais e externos para facilitar seu acesso e manuseio.

3.5.2 Arquitetura segundo Chaudhuri

Além de conhecer os componentes envolvidos na construção do DW é necessário compreender os fluxos de dados que ocorrem no sistema. Conforme [HAC95], "O verdadeiro valor de um sistema de DW não está em apenas colecionar dados, mas sim, gerenciar fluxos de dados". Chaudhuri [CHA97] propõe uma arquitetura, conforme o fluxo e a origem dos dados no sistema de DW, esta arquitetura pode ser dividida em:

1. Fontes de dados de onde o DW irá retirar os seus dados de origem;
2. Um conjunto de estruturas de dados analíticos armazenados: o DW do sistema;
3. Um Sistema Gerenciador de Banco de Dados (SGBD) otimizado para atender os requisitos analíticos dos sistemas de DW;
4. Um componente back end: conjunto de aplicações responsáveis por extrair, filtrar, transformar, integrar e carregar os dados de diferentes origens no DW;
5. Um componente front end: conjunto de aplicações responsáveis por disponibilizar aos usuários finais acesso ao DW;
6. Um repositório para armazenar e gerenciar os metadados do sistema.

Cinco principais fluxos fazem parte do sistema: fluxo de entrada (inflow), fluxo de saída (outflow), fluxo de subida (upflow), fluxo de descida (downflow) e o metafluxo (metaflow).

O primeiro fluxo é o de entrada dos dados no sistema (inflow), que envolve extrair, filtrar, transformar, integrar e carregar os dados de várias fontes no DW. Deve-se considerar as fontes de dados que pertencem à empresa e as fontes externas. O fluxo de entrada é geralmente implementado com ajuda de ferramentas especialmente desenvolvidas para este fim.

O segundo fluxo é o de descida dos dados (downflow), ou seja, em tempos pré-determinados, de dois a cinco anos dependendo da empresa, os dados armazenados no

DW passam para o estado de dados antigos [INM96]. Este é o fluxo que remove do DW aqueles dados considerados velhos, que já não são mais utilizados com freqüência.

O terceiro fluxo é o de subida dos dados (upflow), onde é enfocada a necessidade de colocar os dados em formatos mais acessíveis aos usuários finais. Este processo sumariza e agrupa os dados dentro de "visões" mais adequadas aos usuários finais e as aplicações front end que eles utilizam, tais como tabelas sumarizadas, planilhas, gráficos, páginas no formato Hyper Text Markup Language (HTML), banco de dados pessoais, entre outros formatos. Também é função do fluxo de subida a distribuição dos dados para os diferentes níveis do sistema como, por exemplo, Data Marts e bancos de dados pessoais localizados nas estações de trabalho dos usuários finais.

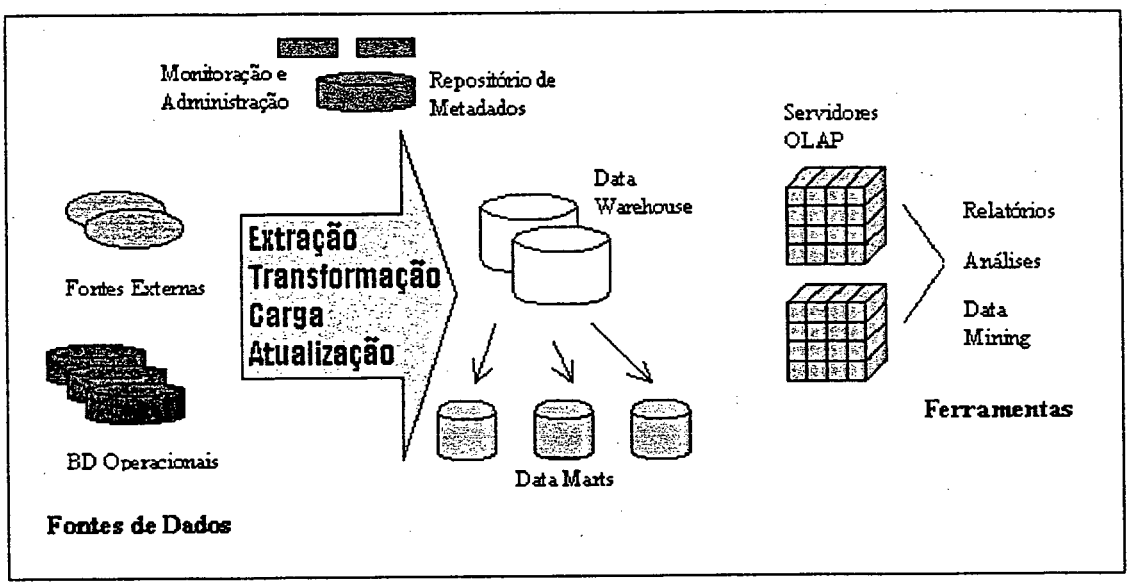

Figura 2: Arquitetura segundo Chaudhuri

O quarto fluxo é o de saída dos dados (outflow), cuja função é disponibilizar acesso aos usuários finais do sistema. Este processo é implementado através de uma variedade de ferramentas front end como, por exemplo, geradores de consulta e relatório, ferramentas com características On-line Analytical Processing (OLAP), pacotes estatísticos, ferramentas de Data Mining, ferramentas de visualização, Executive Information System (EIS), Decision Suport Systems (DSS), entre outras. As ferramentas

front end podem acessar tanto dados previamente preparados pelo fluxo de subida, quanto dados "brutos" e detalhados armazenados no DW.

O quinto e último fluxo pode ser chamado de metafluxo (metaflow), ao contrário dos quatro fluxos de dados citados acima, que descrevem como os dados se movem no DW o metafluxo move metadados, ou seja, dados sobre os outros fluxos. O repositório de metadados é responsável pela gerência do sistema como um todo, indicando de onde os dados vêm, como são transformados, quando são atualizados, o que significam, como são acessados e quem os vê.

3.5.2.1 Fontes de dados internas

Como regra básica, as fontes primárias dos dados de um DW são os sistemas de processamento de transações, os quais dão suporte ao dia a dia de uma empresa. Estes sistemas, também chamados de "sistemas On-line Transaction Processing (OLTP)", "sistemas operacionais" e "sistemas de produção", têm como principal objetivo garantir as operações básicas das empresas nas áreas de produção, administração e comércio, entre outras. Grandes projetos de DW chegam a tratar com mais de quarenta diferentes sistemas de produção [GEL96]. Todos estes sistemas acabam gerando grandes volumes de dados, os quais podem estar armazenados e organizados na forma de sistemas de arquivos, bancos de dados de arquitetura fechada ou aberta e bancos de dados distribuídos.

Além dos dados referentes às transações operacionais da empresa, podem ser necessários outros dados de fontes internas, geralmente não computadorizados como, por exemplo, as metas a serem atingidas em determinado ano ou uma pesquisa mensal que revela o grau de satisfação dos clientes em relação a determinados produtos ou serviços da empresa. Este tipo de informação raramente está disponível através de atividades normais de processamento de dados, necessitando que seja coletada, inserida no DW e mantida.

3.5.2.2 Fontes de dados externas

Outras fontes de dados para um sistema de DW são as fontes externas a empresa, principalmente como apoio para decisões nos níveis gerenciais mais altos. Dentre alguns exemplos estão informações econômicas regionais, sobre o setor de atuação da empresa, sobre concorrentes, preferências do consumidor, entre outras. Informações de fontes externas são geralmente compradas de empresas que mantém bancos de dados comerciais.

Muitas das informações podem estar em um formato não tradicional como, por exemplo, imagens, áudio e, principalmente, dados baseados em documentos. O conteúdo desses documentos, na medida do possível, deve ser armazenado eletronicamente e recuperado de acordo com suas características.

3.5.3.3 Arquitetura segundo Valente

Sabendo que nem todas as camadas descritas no modelo de Orr[ORR96] são necessárias para a implementação de um DW, é possível se definir uma arquitetura mais simples, conforme um estudo realizado por Valente [VAL96]. Nesta arquitetura pode-se ver as bases de dados que compõe o DW, tem-se o extrator que é responsável pela detecção automática de mudanças nas bases de dados. Sempre que existe uma mudança no conteúdo da base, a informação nova ou atualizada, que é relevante para o DW é propagada para o integrador.

O integrador é responsável em integrar o conteúdo fornecido pelos extratores e fornecê-lo ao DW. Para integrar os dados advindos de diversos tipos de bases de dados é necessário que estes passem por alguns processos. Primeiro, os dados devem ser ajustados ao modelo de dados conceitual utilizado pelo DW, então o dado deve ser unido com os dados já existentes, resolvendo possíveis inconsistências que possam existir entre as bases de dados e os dados do DW. Uma lista das tarefas do integrador pode conter itens como: especificar as diferenças das bases de dados, definir

relacionamentos entre dados de múltiplas bases, resolver duplicações e inconsistências e determinar como as informações serão integradas dentro do DW.

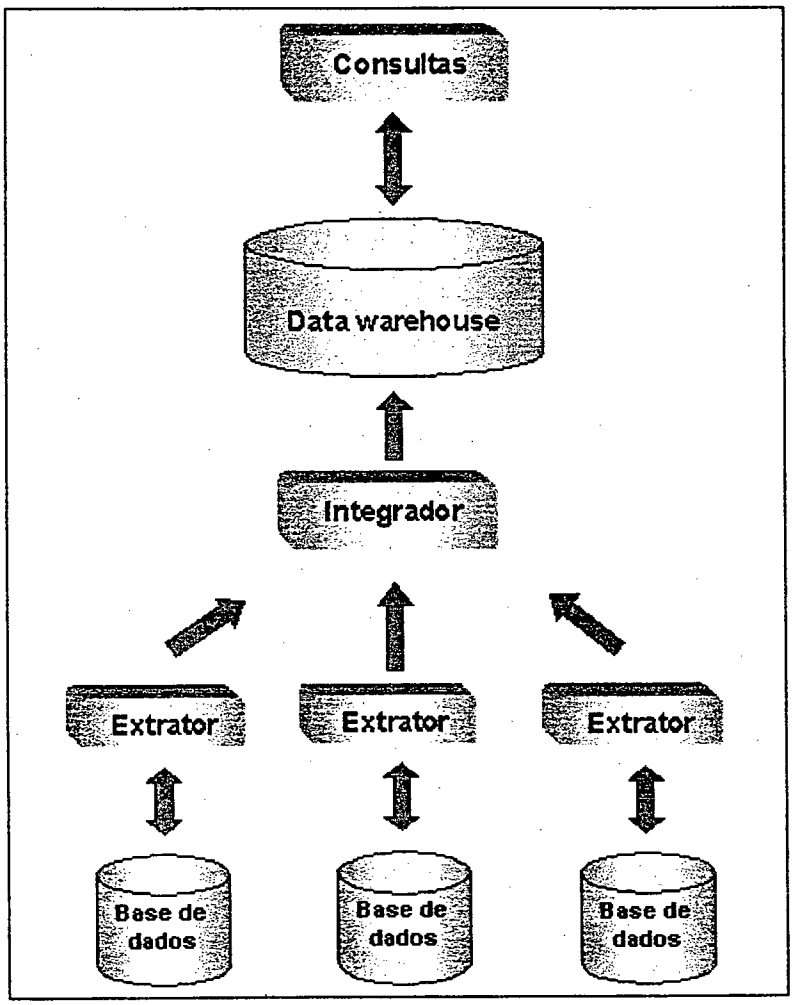

Figura 3: Arquitetura Segundo Valente.

3.5.3.1 Extração dos dados

Para que as modificações nas bases de dados sejam detectadas pelos extratores é necessário que antes seja feita uma tradução das informações das bases de dados para o modelo do DW. Quando alguma modificação de interesse é detectada pelos extratores, a informação é propagada em um formato genérico para o integrador. Quando a base dos dados operacionais é um banco de dados relacional é possível definir um conjunto de

gatilhos ou triggers e guardar as notificações de mudanças. Uma outra maneira é que o extrator examine o arquivo de atualizações (log) e retire as modificações que são de interesse do DW.

Para sistemas que não possuem a facilidade de gatilhos e arquivo de atualizações é necessário implementar funções que notifiquem os extratores a cada mudança na base ou criar um programa para descarregar a base de dados em um arquivo para que o extrator compare este arquivo com o que foi gerado anteriormente, detectando assim as modificações existentes.

Sempre que é realizada a carga de novos dados no DW é necessário que estes passem por um processo de limpeza, descartando dados errados, inserindo dados em formato padrão, eliminando duplicidades e inconsistências e realizando agregações e sumarizações.

3.5.3.2 Integrador

O integrador pode ser implementado como um mecanismo de regra base, recebendo as notificações dos extratores e integrando-as no DW [VAL96]. Cada regra é responsável pela manipulação de um determinado tipo de notificação e é implementada como um método em um sistema orientado a objetos. Quando o extrator gera um determinado tipo de notificação o método correspondente é chamado e então executa os processamentos necessários para integrar os dados no DW, durante este processo o integrador pode obter dados extras no DW ou em outras bases de dados.

3.5.3.3 Processamento de informações

Para receber respostas corretas de um DW é necessário elaborar perguntas certas, para isso o usuário deve ter conhecimento do assunto, experiência e capacidade de análise. É necessário primeiro definir o problema a ser resolvido, depois elaborar as perguntas que necessitam de respostas, verificar se os dados apropriados estão

disponíveis e se são suficientes para responder as questões. O processo de consulta consiste de cinco etapas:

1. Definição das consultas: O usuário deve definir a consulta de forma a traduzir as necessidades da empresas em termos que o DW possa responder, isso pode ser feito através de ferramentas comerciais ou aplicações;
2. Acesso e recuperação de dados: As ferramentas de acesso submetem a consulta e recuperam os dados apropriados, este processo pode envolver cálculos;
3. Cálculo, manipulação e análise: Análises adicionais, como cálculos e manipulações, podem ser feitas para transformar os resultados da consulta em informações;
4. Apresentação das informações: Os resultados podem ser apresentados em forma de relatórios, gráficos, texto em tela ou como dados pré-processados para análise posterior;
5. Disseminação da informação: As informações podem ser distribuídas aos interessados das diversas formas existentes, como relatórios, arquivos, correio eletrônico, etc.

As ferramentas para consulta dos dados normalmente possuem componentes gráficos e suportam um certo grau de multidimensionalidade, como sondagem inteligente, relatórios de intertabelas e análise de séries de tempo.

3.5.4 Outras arquiteturas

Segundo Pieter R. Mimno [BAR96], existem várias opções de arquiteturas diferentes que podem ser consideradas na implementação de um DW. Algumas delas são:

3.5.4.1 Arquitetura de duas camadas

Uma opção de arquitetura para o DW é utilizar um computador de alta capacidade como servidor. Isto é uma incorporação das aplicações utilizadas pelos usuários (front end) com os componentes do servidor (back end).

Aplicações front end construídas com ferramentas cliente/servidor fornecem uma interface gráfica amigável, suportam funções específicas da empresa, possibilitam o acesso transparente aos dados dos sistemas já existentes e escondem a complexidade e a falta de consistência dos bancos de dados atuais além de facilitar a utilização e a visualização dos resultados. Os sistemas operacionais de uma empresa podem estar em uso por 15 ou 20 anos e podem ter altas taxas de redundância.

Organizações como as companhias de seguros, que podem crescer com a compra de outras seguradoras, gradualmente acumulam múltiplos sistemas de computação, cada um deles com redundâncias e incompatibilidade de definições dos dados. A redundância e a falta de consistência dos dados pode dificultar a administração da empresa e o acesso aos dados e impede o desenvolvimento de novas aplicações front end. Uma das maneiras de tratar com esta situação é partir de um só sistema e construir uma espécie de "sistema guarda-chuva" que tenha facilidade de acesso aos dados do servidor principal.

Esta Arquitetura pode ser usada para construir um DW em duas camadas que consiste de componentes dos clientes (front end) e componentes do servidor (back end). Esta arquitetura é atrativa porque ela utiliza os sistemas existentes bem como os servidores de bancos de dados existentes e requer um investimento mínimo em hardware e software. Entretanto, a arquitetura em duas camadas não é escalonável e não suporta um grande número de usuários simultaneamente. Isto estimula o desenvolvimento de estações clientes muito pesadas, pois muito processamento é alocado para processar nestas estações.

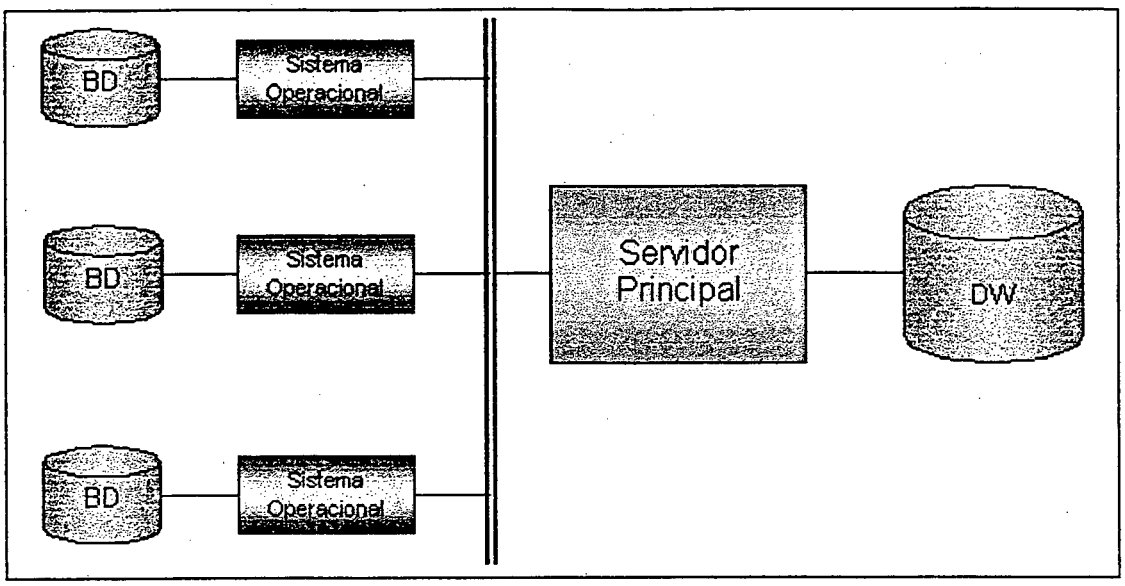

Figura 4: Arquitetura de Duas Camadas

3.5.4.2 Arquitetura de três camadas

Uma alternativa é utilizar a arquitetura de informação em múltiplas camadas. Esta arquitetura flexível suporta um grande número de serviços integrados, na qual a interface do usuário, as funções de processamento do negócio e as funções de gerenciamento do banco de dados são separadas em processos que podem ser distribuídos através da arquitetura de informação.

A arquitetura em três camadas é amplamente utilizada para DW. Na terceira camada ficam as fontes de dados. Dados e regras de negócio podem ser compartilhados pela organização, assim como o banco de dados para o DW, ficam armazenados em servidores de alta velocidade na segunda camada. Na primeira camada ficam as aplicações de interface com os usuários que devem ser gráficas e baseadas em rede.

No ambiente do DW, os servidores de banco de dados e os servidores de aplicações da segunda camada fornecem um acesso eficiente e veloz aos dados compartilhados. Os dados de um DW são tipicamente estáticos, por exemplo, não variam com o tempo e devem ser integrados, de natureza histórica e sumarizados ou agregados para que sejam significantes para os analistas de negócios. Dados operacionais e bancos de dados para o DW são freqüentemente armazenados em servidores fisicamente separados. Bancos de dados operacionais são otimizados para ter alto desempenho no

processamento de transações on-line, em inglês conhecido como On-line Transaction Processing (OLTP). Bancos de dados para DW são otimizados para ter alto desempenho em consultas e análises, em inglês conhecido como On-line Analytical Processing (OLAP).

É importante reconhecer que não existe uma arquitetura "correta" para DW. Para algumas organizações pode ser atrativo utilizar a arquitetura em duas camadas, por que ela minimiza o custo e a complexidade de construção do DW. Para outras que requerem grande performance e escalabilidade, a arquitetura em três camadas pode ser mais apropriada. Nesta arquitetura dados extraídos dos sistemas operacionais são filtrados, transformados e armazenados em servidores de bancos de dados de alta velocidade, os quais são utilizados para o acesso dos usuários finais. No planejamento do DW, as organizações devem examinar as alternativas disponíveis de arquiteturas e selecionar aquela que satisfaça os seus requisitos estratégicos e organizacionais.

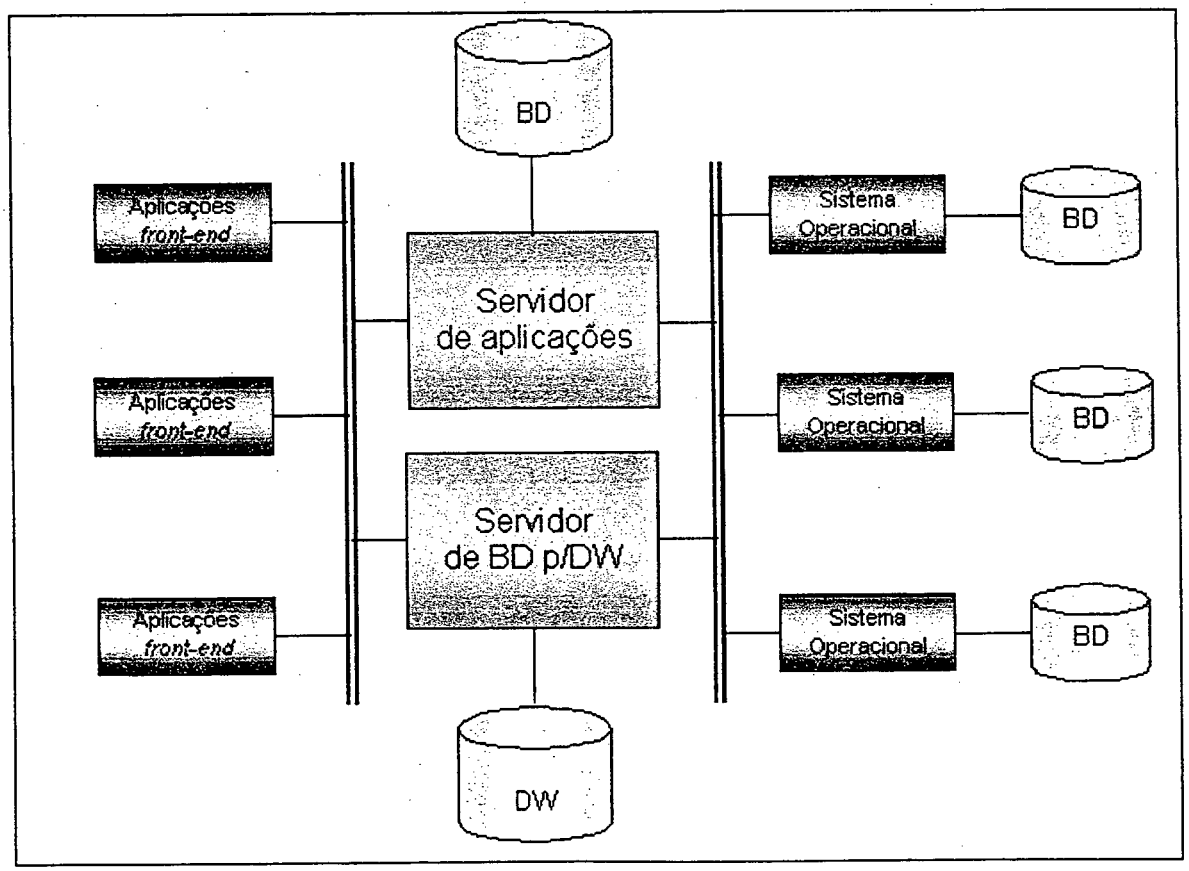

Figura 5: Arquitetura de Três Camadas

Foi

4.0 Modelo de dados

O modelo de dados tem um papel fundamental para o desenvolvimento interativo do DW. Quando os esforços de desenvolvimentos são baseados em um único modelo de dados sempre que for necessário unir estes esforços os níveis de sobreposição de trabalho e desenvolvimento desconexo serão muito baixos, pois todos os componentes do sistema estarão utilizando a mesma estrutura de dados.

Existe um grande número de enfoques sobre modelagem de dados já desenvolvidos por vários autores, a maioria deles pode ser usada para construir um DW. O primeiro modelo foi escrito por R.Kimball em [KIM96] e divide a modelagem dos dados em três partes: modelo empresarial, modelo dimensional e modelo físico. O segundo modelo apresentado foi escrito por W.H.Inmon em [INM93] e também divide a modelagem dos dados em três partes: a modelagem de alto nível, a modelagem de nível intermediário e a modelagem de baixo nível

4.1 Modelo de dados segundo R.Kimball

A construção de um modelo de dados ajuda a compreender as regras de negócio que o DW irá apoiar. Muitas vezes a equipe que está desenvolvendo o ambiente do DW ignora a fase de construção do modelo de dados por não ter tempo ou achar que não é uma parte fundamental para o desenvolvimento. Segundo [KIM96], um dos maiores problemas no desenvolvimento do DW é a compreensão dos dados, em seu artigo são descritos três modelos de dados, um modelo entidade-relacionamento normalizado das regras de negócio ou modelo empresarial, um modelo dimensional e um modelo físico.

4.1.1 Modelo empresarial

A análise do modelo de dados é o primeiro passo para desenvolver um modelo entidade-relacionamento normalizado das regras de negócio. Este é o modelo empresarial. Nesta fase não se deve pensar em como as informações serão recuperadas

nem em como serão utilizadas. Estas tarefas serão realizadas posteriormente. Nesta fase o importante é ter o foco na estrutura da informação, como os atributos e as relações entre elas. Por exemplo, se o DW que estiver sendo construído for sobre vendas em marketing, o tipo de perguntas que se deve procurar responder são:

1. Quem compra o produto? (os clientes e sua estrutura);
2. Quem vende o produto? (organização das vendas, os canais de distribuição e assim por diante);
3. O que é vendido? (estrutura de produto);
4. Quando é vendido? (estrutura de tempo);
5. Como e vendido? (contrato, telefone e assim por diante);
6. Quais são as características da venda (de promoção, venda normal, etc.)

Estas questões irão ajudar a descobrir quais são os dados relevantes para o DW e quais os dados que irão resultar em estruturas dimensionais. É muito comum iniciar um processo perguntando aos usuários o que eles querem ver. Existe um grande problema nisso. O problema é que eles irão pedir por coisas não respondidas no passado, e se for trabalhado no sentido de oferecer respostas para perguntas do passado os novos problemas que surgirem podem requerer informações novas que não estarão disponíveis. E o novo DW estará instantaneamente obsoleto. Pelo menos durante a fase de construção do modelo empresarial é importante desenvolver um modelo que contenha todos os dados que estão disponíveis nos sistemas operacionais da empresa e em fontes externas.

Isto demonstra uma das diferença entre os sistemas de suporte a decisão e os sistemas operacionais. Nos sistemas operacionais, o processo empresarial é quem define os requisitos dos dados. A utilidade de cada parte dos dados pode ser determinada avaliando seu valor dentro do processo empresarial. Considerando que o processo não muda, o valor dos dados não muda. Embora o processo empresarial determine a informação em uma aplicação de DSS, o processo empresarial pode e vai mudar.

É muito comum confundir as perguntas que deveriam ser feitas com as perguntas que se deseja perguntar com as perguntas que podem ser feitas. Um exemplo disto é a questão: "Que produtos compram as pessoas casadas?". Esta parece ser uma pergunta interessante e relativamente fácil responder. Para executar alguma análise de tendência, poderia ser perguntado: "O que compraram as pessoas casadas ano passado?". Esta pergunta parece com a primeira, entretanto sem a informação de estado civil do último ano seria impossível responder a questão. Este é só em exemplo de como uma informação, ou a falta dela, pode fazer uma diferença crítica para o ambiente de DW.

O DW pode ser considerado como uma estrutura orgânica, pois ele cresce e cresce em direções que não podem ser previstas. Se forem enfocadas exigências funcionais atuais, será desenvolvida uma solução altamente aperfeiçoada para os problemas de hoje. Porém, como os problemas mudam a solução pode ficar cada vez menos ótima. Eventualmente, estará regular e conseqüentemente ficará inaceitável. Um caminho melhor é trocar otimização por flexibilidade. A solução resultante poderá não ser a ideal mas será mais adaptável e conseqüentemente terá um tempo de vida maior.

O segundo passo para o desenvolvimento do modelo empresarial é a normalização do modelo, uma vez definidas todas as entidades e relações é possível normalizar o modelo. Pode-se utilizar a 3ª Forma Normal que é alcançada quando os atributos dependem exclusivamente da sua chave, há muitas vantagens em utilizar a 3ª FN.

1. A estrutura é notavelmente insensível a mudanças. Se for preciso mudar, acrescentar ou excluir, atributos que são relacionados a uma chave em particular, não deveria haver nenhuma razão para mudar outras entidades ou relações. De uma perspectiva do DW, isto significa que a estrutura do DW pode ser modificada facilmente para refletir mudanças organizacionais ou problemas empresariais novos.

2. Os caminhos estruturais para ter acesso a informação ficam mais claros. Considerando que este modelo é o resultado de regras de negócio é possível que surjam caminhos cíclicos. Isto significa que é necessário educar os usuários sobre os caminhos diferentes e que interpretação cada um deles têm. Os caminhos cíclicos devem ser eliminados pois a maioria das ferramentas de consultas não consegue trabalhar com eles.

3. Muito da discussão sobre as formas normais gira em torno da eliminação de problemas de atualização. Isto é , quais os tipos de problemas de atualização que são resolvidos movendo-se para uma forma normal mais alta. Considerando que um DW raramente é atualizado, a possibilidade destes problemas acontecerem é muito baixa. Este é o motivo pelo qual é aceitável falar sobre desnormalização do DW.

Também existem algumas desvantagens na utilização da 3 ª FN:

1. O desempenho pode ser comprometido, podendo ficar muito abaixo do desejado. Muito do trabalho que é feito para desnormalizar o modelo de dados é uma tentativa para alcançar objetivos de desempenho.

2. É possível que muitas relações pequenas sejam criadas e o usuário pode pensar que são uma única relação ou grupo de dados. O modelo dimensional reduz este tipo de problema até um certo ponto.

Considerando que o modelo empresarial não será implementado, estas desvantagens raramente são críticas, porém se levadas para o modelo dimensional estas desvantagens poderão reaparecer e será necessário um tratamento especial para cada situação.

O terceiro passo, para desenvolver um modelo entidade-relacionamento normalizado das regras de negócio, é a definição das restrições ou regras de integridade. Como norma geral, deveria ser obrigatória a definição das restrições de integridade dentro de um DW. As restrições de integridade ajudam a garantir a consistência dos resultados das consultas. Por exemplo, quando forem executadas operações de refinamento das informações, conhecidas por operações de drill-down, por uma estrutura dimensional, as restrições de integridade garantem que a resposta obtida seja sempre a mesma, como se tivesse sido feita uma pesquisa por toda a tabela. Nota-se que a não obrigatoriedade das restrições de integridade pode resultar em respostas diferentes para perguntas iguais.

Por outro lado a obrigatoriedade da definição das restrições de integridade pode fazer com que alguns registros não sejam selecionados durante o processo de carga. A

decisão sobre a definição de regras de integridade deve levar em conta o que é melhor para o DW que está sendo construído, respostas consistentes, mas possivelmente incompletas ou respostas completas mas possivelmente inconsistentes.

É possível que as fontes de dados criem situações onde não seja possível utilizar as restrições de integridade, neste caso é importante saber as implicações que isto pode ter sobre o DW e manter os usuários informados sobre o que pode acontecer em determinadas situações.

4.1.2 Modelo Dimensional

A resposta a perguntas complexas e que envolvam questões de análise dos negócios de uma empresa, normalmente requerem uma visão dos dados de perspectivas diferentes. As respostas a esse tipo de pergunta é que podem levar a tomadas de decisões acertadas ou não. As ferramentas baseadas em SQL podem ajudar na pesquisa de dados relacionados a este tipo de consulta, o que ocorre é que normalmente as respostas não são conseguidas em um tempo curto, principalmente pela falta de flexibilidade destas ferramentas.

Para exemplificar, imagine-se uma rede de supermercados que esteja querendo melhorar o desempenho de suas vendas ou saber se suas promoções estão trazendo bons resultados. Para isso, é necessário examinar os dados sobre as vendas disponíveis na empresa. Uma avaliação deste tipo requer uma visão histórica do volume de vendas sob múltiplas perspectivas, como por exemplo: volume de vendas por produto, volume de vendas por marca, volume de vendas por filial, volume de vendas por período de tempo.

Chama-se de dimensões as diferentes perspectivas envolvidas, no caso, produto, marca, filial e mês. Estas dimensões usualmente correspondem a campos não numéricos em um banco de dados. Considera-se também um conjunto de medidas, tal como vendas ou despesas com promoção. Estas medidas correspondem geralmente a campos numéricos em um banco de dados. A seguir, avaliam-se agregações destas medidas segundo às diversas dimensões, essas agregações ficam armazenadas para acesso futuro.

Por exemplo, calcula-se a média de todas as vendas por todos os meses por filial. A forma como estas agregações são armazenadas pode ser vista em termos de dimensões e coordenadas, dando origem ao termo multidimensional.

A modelagem multidimensional é o nome de uma técnica de projeto lógico freqüentemente usada para DW, cujo principal objetivo é apresentar o dado em uma arquitetura padrão e intuitiva, que permita acessos de alta performance [KIM96]. 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 a mês igual a maio e filial 4 é um ponto com coordenada [maio, filial 4]. Neste caso, mês e filial são duas dimensões e vendas é uma medida. Teoricamente, quaisquer dados podem ser considerados multidimensionais. Entretanto, o termo normalmente se refere a dados representando 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, para isso é necessário utilizar formas específicas de modelagem como o modelo "Estrela" e o modelo Floco de Neve" descritos a seguir.

4.1.2.1 Modelo Estrela

Conforme Daphnis Valente em [VAL96], tradicionalmente, modelos de bases de dados relacionais apresentam tabelas com relacionamentos complexos e com múltiplas uniões circulares entre dois pontos do modelo. Para a maioria dos usuários que utilizam ferramentas para compor suas consultas é necessário que o acesso a base de dados seja simples o suficiente para facilitar o acesso direto a base de dados. Para acomodar as necessidades de todos os usuários e facilitar a atualização do DW o projetista deve criar um modelo que o usuário final possa facilmente entender em termos do negócio.

O principal tipo de modelo dimensional, é o chamado modelo Star (Estrela), onde existe uma tabela dominante no centro, chamada tabela de fatos, com múltiplas junções conectando-a a outras tabelas, sendo estas chamadas de tabelas de dimensão. Cada uma das tabelas secundárias possui apenas uma junção com a tabela central. O modelo Estrela, tem a vantagem de ser simples e intuitivo, mas também faz uso de novos enfoques de indexação e união de tabelas.

A tabela de fatos contém milhares ou milhões de valores e medidas do negócio da empresa, como transações de vendas ou compras. Cada uma destas medidas é tomada segundo a interseção de todas as dimensões. Os fatos melhores e mais úteis são numéricos, continuamente valorados (diferentes a cada medida) e aditivos, já que estes facilitam a geração do conjunto de respostas. Uma outra característica da tabela de fatos é a esparsidade, ou seja, se não existe um cruzamento para alguns valores das dimensões, a tabela de fatos não armazena zeros.

As tabelas de dimensão armazenam as descrições textuais das dimensões do negócio. Cada uma dessas descrições textuais ajuda a definir um componente da respectiva dimensão. Uma das principais funções dos atributos de tabelas de dimensão é servir como fonte para restrições em uma consulta ou como cabeçalhos de linha no conjunto de resposta do usuário. Tabelas dimensões tendem a utilizar tipos caracteres ao invés de numéricos, de forma que suas linhas são muito mais longas mas em pouca quantidade ocupando uma pequena percentagem de espaço em disco. As tabelas de fatos podem utilizar até 95% da área destinada ao DW [BAR96].

Na maioria das vezes as dimensões representam hierarquias, como por exemplo, um produto, que é de uma marca ou categoria, que por sua vez pertence a uma sub-categoria, etc. Só que, na maioria das vezes, quando esta é representada na dimensão, não temos várias tabelas normalizadas com ligações um-para-muitos, e sim uma única tabela de dimensão. Isso faz com que a performance das consultas aumente muito, já que não são necessários joins para se obter os dados relacionados com algum assunto.

Outro fato importante é que como a tabela de fatos na verdade representa os relacionamentos muitos-para-muitos entre as tabelas de dimensões, esta tem como chave

primária uma chave composta de todas as chaves estrangeiras das tabelas de dimensão [KIM96].

Para um bom desempenho do modelo Estrela é necessário que os projetistas saibam antecipar, na modelagem do DW, as consultas mais freqüentes a serem realizadas pelos usuários. Com a redundância seletiva e relacionamentos pré-estabelecidos o projetista pode simplificar os dados facilitando seu acesso.

4.1.2.2 Variação do modelo Estrela

Outro tipo de estrutura bastante comum, conforme [CAM97], é o modelo de dados Snow Flake (Floco de Neve), que consiste em uma extensão do modelo Estrela onde cada uma das "pontas da estrela" passa a ser o centro de outras estrelas. Isto porque cada tabela de dimensão seria normalizada, "quebrando-se" a tabela original ao longo de hierarquias existentes em seus atributos. Este modelo pode ser ilustrado imaginando-se a classificação de um automóvel., onde a dimensão do produto possui uma hierarquia definida: categoria se divide em marca e marca se divide em produtos. Da mesma forma, a dimensão tempo inclui ano que contém mês e mês que contem dia-do-mês. Cada um destes relacionamentos muitos-para-um geraria uma nova tabela em um modelo Floco de Neve.

Kimball [KIM96] aconselha os projetistas a resistirem à tentação de transformar modelos Estrela em modelos Floco de Neve, devido ao impacto da complexidade deste tipo de estrutura sobre o usuário final, enquanto que o ganho em termos de espaço de armazenamento seria pouco relevante.

4.1.2.3 Vantagens do Modelo Estrela

Segundo Kimball, o modelo dimensional apresenta várias vantagens no que diz respeito a sua utilização para DW, dentre estas estão [KIM96]:

O modelo Estrela tem uma arquitetura padrão e previsível. As ferramentas de consulta e interfaces do usuário podem se valer disso para fazer suas interfaces mais amigáveis e fazer um processamento mais eficiente;

Todas as dimensões do modelo são equivalentes, ou seja, podem ser vistas como pontos de entrada simétricos para a tabela de fatos. As interfaces do usuário são simétricas, as estratégias de consulta são simétricas, e o SQL gerado, baseado no modelo, é simétrico;

O modelo dimensional é totalmente flexível para suportar a inclusão de novos elementos de dados, bem como mudanças que ocorram no projeto. Essa flexibilidade se expressa de várias formas, dentre as quais temos:

Todas as tabelas de fato e dimensão podem ser alteradas simplesmente acrescentando novas colunas a tabelas;

Nenhuma ferramenta de consulta ou relatório precisa ser alterada de forma a acomodar as mudanças;

Todas as aplicações que existiam antes das mudanças continuam rodando sem problemas;

Existe um conjunto de abordagens padrões para tratamento de situações comuns no mundo dos negócios. Cada uma destas tem um conjunto bem definido de alternativas que podem então ser especificamente programadas em geradores de relatórios, ferramentas de consulta e outras interfaces do usuário. Dentre estas situações temos:

Mudanças lentas das dimensões: ocorre quando uma determinada dimensão evolui de forma lenta e assíncrona;

Produtos heterogêneos: quando um negócio, tal como um banco, precisa controlar diferentes linhas de negócio juntas, dentro de um conjunto comum de atributos e fatos, mas ao mesmo tempo esta precisa descrever e medir as linhas individuais de negócio usando medidas incompatíveis;

Outra vantagem é o fato de um número cada vez maior de utilitários administrativos e processo de software serem capazes de gerenciar e usar agregados, que são de suma importância para a boa performance de respostas em um DW.

4.1.2.4 Comparação entre o modelo E-R e o modelo Estrela

Uma das principais diferenças entre o modelo E-R e o modelo Estrela é a complexidade. Por ser normalizado, o modelo E-R gera dezenas ou até centenas de tabelas conectadas entre si, o que faz com que ele se torne confuso e de difícil compreensão por parte dos usuários. Já o modelo multidimensional do tipo estrela apresenta uma estrutura muito mais simples, onde uma tabela central, a tabela de fatos, é ligada a várias tabelas de dimensões uma única vez, tornando o modelo de fácil compreensão [KIM95].

A construção do modelo E-R se baseia no micro-relacionamento entre os dados, ou seja, a escolha das entidades e seus relacionamentos é feita baseando-se em como as coisas acontecem. No modelo dimensional, a escolha das dimensões e atributos de fatos são baseados na opinião do usuário, o que faz com que o modelo reflita as suas necessidades [KIM95].

O modelo dimensional enxerga as informações de uma perspectiva histórica ao invés de transações atômicas, tal como nos diagramas E-R [RAD96].

O modelo multidimensional é globalmente consistente para toda a empresa, ao passo que o modelo relacional é consistente somente dentro de seu escopo. A principal razão disso é que o modelo dimensional tem como objetivo fornecer uma visão integrada dos dados da empresa [RAD96].

Os relacionamentos no modelo E-R são modelados explicitamente, ao passo que no modelo multidimensional estes relacionamentos são representados pela existência de fatos no cruzamento entre as dimensões [RAD96].

4.1.2.5 Mapeamento do modelo E-R para o modelo Estrela

É de fundamental importância para o DW, que os dados da empresa sejam integrados, de forma que este possa ser construído facilmente e de forma consistente. O modelo de dados corporativo normalmente expressa esta integração, por isso alguns autores sugerem que se utilize o modelo de dados corporativo para se chegar ao modelo

dimensional. Entretanto o modelo multidimensional mais utilizado para a construção de DW é o modelo Estrela que possui uma estrutura totalmente diferente do modelo E-R.

Como os dois modelos podem representar dados de uma mesma empresa de perspectivas diferentes é normal que exista uma ligação entre eles. Baseando-se nisso, propostas de mapeamento entre os dois modelos começaram a surgir, de forma que os modelos E-R corporativos pudessem ser aproveitados para gerar os modelos Estrela.

Como os modelos E-R representam todos os possíveis processos de negócio da empresa, enquanto cada modelo dimensional focaliza em único assunto de interesse, é possível que de um modelo E-R surjam vários modelos dimensionais que normalmente estarão interligados.

Uma das abordagens, feita por Kimball detalha os seguintes passos para o mapeamento [KIM97]:

1. Separar o diagrama E-R nos diferentes processos de negócio e modelar cada um separadamente;

2. Selecionar os relacionamentos muitos-para-muitos que contenham fatos aditivos e numéricos, no modelo E-R, para transformá-los em tabelas de fato;

3. Por fim, deve-se desnormalizar todas as tabelas remanescentes em tabelas com chaves simples, que se conectam diretamente as tabelas de fato. Estas tabelas se tornam as tabelas de dimensão.

Nos casos em que a tabela de dimensão se conecta a mais de uma tabela de fatos, esta deve ser representada em ambos os esquemas, e estas são ditas então serem conforme entre os dois modelos.

O modelo dimensional global resultante, para uma grande empresa, deve consistir de cerca de dez a vinte e cinco modelos Estrela conectados. Cada um tendo de quatro a doze dimensões. Se o projeto foi feito de forma correta, muitas destas dimensões serão compartilhadas entre as tabelas de fatos.

4.1.3 Modelo Físico

Cada um dos modelos apresentado tem um propósito particular. O propósito do modelo físico é alcançar objetivos de desempenho. Se os computadores não tivessem restrições de velocidade e de recursos de armazenamento, não seria necessária a preocupação com o modelo físico, mas como este não é o caso a elaboração de um bom modelo físico e outro não tão bom pode ser a diferença entre sucesso e fracasso do DW. O modelo físico também é dependente do SGBD e da configuração de hardware. As diretrizes encontradas na definição de Kimball [KIM96] procuram não enfocar nenhum SGBD ou hardware específico.

Com raras exceções o desempenho de um DW será limitado nos processo de entrada e saída dos dados, portanto são problemas que estão relacionados com o SGBD ou com o próprio DW. Caso o DW tenha significantes problemas de desempenho é importante que o modelo físico sofra uma revisão[MCG98].

4.2 Modelo de dados segundo W.H.Imon

Este modelo que será resumidamente apresentado neste trabalho, pode ser pesquisado mais detalhadamente em [INM93]. Este modelo é composto por três níveis de modelagem: a modelagem de alto nível, a modelagem de nível intermediário e a modelagem de baixo nível.

4.2.1 Modelo de dados de alto nível

O alto nível de modelagem apresenta as entidades e seus relacionamentos. As entidades exibidas neste nível encontram-se no nível mais alto de abstração. Para determinar quais entidades participam deste nível é necessário estabelecer o "escopo de integração" que é quem define as fronteiras do modelo de dados e deve ser definido antes do início do processo de modelagem. O escopo de integração pode ser definido pelos analistas, pela gerência e pelo usuário final e deve ser redigido em não mais do que cinco páginas e em uma linguagem de fácil entendimento para as pessoas de negócios.

4.2.2 Modelo de dados de nível intermediário

O modelo de dados de nível intermediário é criado a partir das áreas de interesse ou entidades identificadas no nível alto de modelagem, para cada uma destas áreas ou entidades é desenvolvido um nível intermediário próprio. Este nível é composto por quatro elementos.

1. Um agrupamento primário de dados, o qual é composto pelos atributos que aparecem uma única vez em cada área de interesse. Como todos os agrupamentos de dados o agrupamento primário contem atributos e chaves;
2. Um agrupamento secundário de dados que engloba os atributos que podem aparecer mais de uma vez na mesma área de interesse;
3. Um conector que representa os relacionamentos dos dados entre as áreas de interesse;
4. O tipo dos dados.

Esses quatro elementos de modelagem são usados para identificar os atributos de dados de um modelo e os relacionamentos entre tais atributos. Normalmente cada agrupamento de dados existente no modelo de dados resulta na definição de uma tabela durante o processo de projeto do banco de dados.

4.2.3 Modelo de dados de baixo nível

O modelo de baixo nível, também chamado de modelo físico, de dados é criado a partir do modelo de nível intermediário simplesmente expandindo este de forma que ele passe a apresentar chaves e características físicas. Neste ponto o modelo físico de dados necessita ser alterado para receber as características de performance.

No caso do DW, o primeiro passo para a inclusão dos fatores de performance consiste em decidir sobre a granularidade e o particionamento dos dados, além da alteração da estrutura da chave para a inclusão do elemento de tempo. Depois que a granularidade e o particionamento tiverem sido incluídos, várias outras atividades de projeto físico são embutidas no projeto.

Uma das principais atividades no projeto físico se refere a otimização das operações de E/S (entrada/saída) física. E/S física é a atividade que introduz os dados no computador a partir do meio de armazenamento ou envia os dados do computador para o meio de armazenamento. A tarefa do projetista do DW consiste em organizar os dados fisicamente de forma que durante a execução de uma E/S seja transferida uma massa de dados que apresente alta probabilidade de ser acessada. Por exemplo, se um programador precisar acessar e recuperar cinco registros e estes registros estiverem dispostos em blocos de dados diferentes, serão necessárias cinco E/S. Mas se o projetista tiver condições de prever a necessidade de agrupar os registros e coloca-los no mesmo bloco físico, somente uma E/S será necessária, fazendo com que o programa seja executado com mais eficiência.

5.0 Estudo de Caso para o desenvolvimento de um Data warehouse Para Intranet

5.1 Introdução ao Data warehouse Para Intranet

Um data warehouse para intranet se baseia nas tecnologias Internet dentro dos limites corporativos para assegurar um acesso confiável as informações essenciais da empresa. O desenvolvimento de aplicativos comerciais para a instalação de intranet apresenta novas abordagens para distribuição de tarefas computadorizadas entre clientes e servidores da rede. Um data warehouse para intranet requer consultas sofisticadas a banco de dados, processamento analítico e lógica de formatação para transformar dados brutos em relatórios e imagens que possam ser exibidas em um navegador. Mas o objetivo chave de um data warehouse é proporcionar amplo acesso à informação na empresa para apoio à tomada de decisões em cada nível de gerenciamento . A intranet proporciona as tecnologias para atingir essa meta tornando disponíveis informações em toda a empresa para apoiar decisões eficazes.

5.2 Aplicativos de data warehouse baseados em Cliente/Servidor versus Intranet

Utilizar a abordagem intranet significa abrir o data warehouse aos usuários em toda a empresa assim como a usuários ligados indiretamente à empresa, como fornecedores, atacadistas, varejistas, franqueados e assim por diante. As diferenças entre a organização cliente/servidor e a intranet começa a crescer em virtude da necessidade de suporte a uma comunidade muito grande de usuários.

As diferenças entre um data warehouse para intranet e um data warehouse cliente/servidor são filosóficas em vez de técnicas. As diferenças dependem das necessidades de informação da comunidade de usuários e do método usado para organizar essa informação de acordo com essas necessidades.

Um data warehouse para intranet enfatiza mais o apoio à tomada de decisões e táticas do que um data warehouse cliente/servidor.

O que diferencia um data warehouse para intranet de um data warehouse para cliente/servidos são:

- Necessidade do Usuário;
- Conteúdo do data warehouse;
- Escalabilidade;
- Segurança;

5.2.1 Necessidades do Usuário

Os data warehouse cliente/servidor enfocam principalmente a necessidade dos usuários experientes - possuidores de habilidades técnicas e analíticas, mas sem acesso as fontes apropriadas. Esses usuários precisam executar uma ampla análise de dados para apoiar a tomada de decisões estratégicas.

A comunidade de usuários do data warehouse para intranet, por outro lado, consiste freqüentemente em usuários que desejam informações relacionadas a tópicos comerciais ao invés de ferramentas de análise, ou seja, precisa de apoio a decisões operacionais e táticas.

5.2.2 O Conteúdo do data warehouse

Os fatores desempenho e segurança podem desempenhar papéis cruciais na determinação do conteúdo de um data warehouse para intranet. Esses fatores influenciam decisões de conteúdo e o nível de detalhes a ser mantido, assim como o projeto do data warehouse.

O tamanho da comunidade de usuários do data warehouse para intranet e a diversidade dos aplicativos organizados em uma intranet afetam igualmente o desempenho de um data warehouse. Em muitos casos as organizações percebem que um projeto descentralizado de data warehouse (por exemplo, data marts) oferecerá a

segurança e o desempenho desejado, além de proporcionar acesso eficiente às fontes de informação e aplicativos que o usuários precisam para apoiar suas decisões.

5.2 Implementando o Data warehouse

Para iniciarmos a implementação do Data Warehouse utilizamos como base , um sistema transacional utilizado pelas escolas Yázigi do Brasil. Este sistema se chama SAEL – Sistema de Administração de Escolas Livres e tem como principal objetivo fazer o gerenciamento diário de todos os processos utilizados pela escola.

5.2.1 Especificando o Sistema (OLTP)

O Sistema de Administração de Escola Livres - SAEL é um aplicativo multi-empresa, totalmente integrado que gerencia as atividades ligadas a secretaria da escola, passando pelo controle e acompanhamento pedagógico dos alunos, até os controles de administrativos.

Possui um alto nível de parametrização, evitando assim a dependência do usuário quanto as modificações dinâmicas, inerentes ao ambiente escolar.

Atualmente o sistema está portado para 3 plataformas distintas que são windows 9x, para ambientes mono processados, windows nt para redes locais e Red Hat Linux para redes locais independentes e integradas.

O sistema está estruturado em 4 módulos operacionais. Secretaria, Administrativo, Financeiro e Pedagógico.

5.2.1.2 Secretaria

Este módulo contempla os eventos que são tratados diretamente com o público alvo. Oferece aos profissionais de vendas e marketing, ferramentas para auxilio na garimpagem de pessoas. Auxilia todas as atividades geradas pela movimentação de alunos, tais como transferências, registros, históricos, controle de pontuação de ficha score, controle individual de movimento de caixa, entrega de material didático, etc.

Neste módulo, há uma tabela de prospects[3], onde são cadastradas as pessoas interessadas em estudar na escola, assim que as matriculas destes prospects são efetuadas, automaticamente eles são lançados no cadastro de alunos.

O Sistema controla todos os passo do aluno dentro da escola, desde a matricula até o pagamento das mensalidades. São emitidos relatórios mensais de controle de alunos inadimplentes, controle de pagamentos, controle de mensalidades, controle de saídas e entradas de alunos, desistências.

5.2.1.3 Administrativo

Um dos mais completos módulos do aplicativo. O módulo administrativo possui uma sub-divisão que torna a estrutura empresarial, existente em cada escola, bastante objetiva.

O controle de estoque dispõe dos controles básicos para manipulação de produtos e/ou serviços. Permite a inclusão de itens compostos, ou seja, itens que são formados pôr conjuntos de itens. A definição de valores é altamente flexível, no que diz respeito a utilização de até 99 moedas diferentes.

O controle de contas a pagar provê, mediante ao cadastramento de todos os fornecedores e demais credores financeiros, um visão ampla da capacidade de endividamento da empresa. É totalmente integrando ao módulo de contabilidade gerencial. A contabilidade gerencial permite ao diretor da escola a perfeita planificação de todos os ativos e também das contas de compromissos e obrigações para com terceiros. Permite, na sua estrutura básica, um plano de contas que pode ter até 9 níveis de detalhamento.

Outra importante característica encontrada ao módulo administrativo, é a atribuição de níveis funcionais, com diferentes permissões. Assim, mesmo em um ambiente mono-usuário, cada colaborador possui suas seções personalizadas, podendo controlar desde movimento individual de caixa até o trabalho de marketing em seus próprios prospects.

---

[3] Prospects são pessoas interessadas em estudar na escola, mas que ainda não fizeram a sua matricula.

5.2.1.4 Financeiro

Apesar de ser compacto, o módulo financeiro oferece ao diretor as ferramentas necessárias para o controle de faturamento, inadimplência e totalização de impostos.

Seguindo a mesma filosofia de operação, o sistema permite ao usuário a definição de qualquer tipo de plano de pagamento, da maneira mais flexível possível. Assim na venda de produtos e/ou serviços, condições de pagamentos diferentes podem ser oferecidas para atender situações especiais.

É também no módulo financeiro que o usuário poderá, a seu exclusivo critério, controlar até 99 moedas diferentes com cotações diárias.

5.2.1.5 Pedagógico

Sendo um dos mais completos módulos do SAEL – Sistema de Administração de Escolas Livres, o pedagógico constitui-se numa das mais completas base de informações sobre avaliações, estatísticas e movimentações de alunos.

Permite ao usuário a definição de esquemas individuais de avaliações de cursos com ampla flexibilidade.

Controla a evasão de alunos emitindo relatórios pôr motivos e em datas especiais. Gerencia também a emissão de pedidos de certificados.

Uma grande variedade de relatórios pode ser gerada, tais como diário de classe, mapa de movimentação, campanha de rematrícula, avaliações para convênios, etc.

Ainda contamos com a parte de Utilitários do sistema, é neste módulo que são parametrizadas várias informações sobre o sistema, entre eles os caixas das secretárias. É neste módulo que temos as opções de cadastramento da empresa, senhas e backups.

5.3 Especificando as Tabelas do Sistema (OLTP)

Os sistemas transacionais são tradicionalmente implementados como uma parte de suas rotinas desenvolvidas para dar suporte à entrada de dados, outra parte destas rotinas serve para o processamento destes dados e a terceira parte que implementa as saídas de dados.

No sistema que serviu de base para a implementação do protótipo são utilizadas as informações oriundas das bases transacionais de gerenciamento dos alunos onde são feitas as entradas de dados a partir de enumeras tabela. Entre as principais temos :

Tabela	Descrição
EMPRESA	Cadastramento da Empresa (Escola) com seus dados
PARAMETROS	Armazena os parâmetros utilizados no sistema
FUNCIONARIO	Cadastramento dos funcionários da escola com os seus acessos ao sistema.
PRODUTOS	Cadastramento dos Produtos Comercializados na escola (Material e Curso)
PLANO DE PAGTO	Cadastramento dos vários planos de pagamentos que a escola poderá aplicar para o pagamento dos alunos
CONVENIO	Cadastramento dos Convênios que poderão ser aplicados aos alunos
CURSO	Cadastramento dos cursos
TURMA	Cadastramento das Turmas do Semestre
CRONOGRAMA	Cadastramento do Cronograma de Aula dos Cursos
ITEM AVALIAÇÃO	Cadastramento dos Itens de Avaliação dos Cursos
ALUNO	Cadastramento dos Alunos matriculados na Escola
SEMESTRE	Cadastramento dos Alunos Ligados as suas Turmas
ITEMSEMESTRE	Cadastramento dos Produtos Vendidos aos Alunos
FINANCEIRO	Armazena as Parcelas dos Alunos

RESERVA	Armazena os Materiais Comprados pelos Alunos para a entrega no inicio das aulas
BANCO	Cadastramento do Banco para a geração de Boletos Bancários
ESTOQUE	Guarda a Quantidade em estoque dos produtos cadastrados
CAIXA	Guarda os recebimentos feitos pelo caixa
MOVFINANCEIRO	Guarda as movimentações financeiras feitas pelos funcionários
NOTAS	Armazena as notas dos alunos por semestre, por curso e por turma
PROSPECTS	Atendimento Pré-Venda de Pessoas interessadas em estudar na escola
ACOES P/ PROSPECTS	Guarda as ações que vão ser aplicadas para os prospects.

Tabela 5: Tabelas do sistema de gerenciamento escolar

5.4 Justificativa

Os relatórios obtidos através do sistema de administração escolar descrito anteriormente são pré-definidos e foram programados para fornecer um padrão estático de informação. Neste sistema, os tomadores de decisão podem apenas consultar os dados neste formato, ou seja, com linhas e colunas imutáveis.

Num sistema de Data warehouse os usuários podem decidir a qualquer momento quais são as colunas e quais são as linhas da grade, fazer gráficos a partir destes dados, totalizar por esta ou aquela dimensão, tudo isto sem que seja necessário programar uma linha sequer de código.

A modelagem dimensional, permite que, com uma simples operação de arrastar e soltar com o mouse, o usuário mude totalmente o panorama da planilha, fornecendo ângulos diferentes com um mínimo de esforço e gasto de tempo.

Os programas de administração da escola servem para analisar de forma estática as informações, sendo que para incorporar uma nova consulta são necessárias várias horas de programação e de definições de consultas complexas para atingir resultados satisfatórios.

Com os sistemas de apoio à decisão pode-se realizar inclusive projeções com os dados, além de acessar informações que foram coletadas há muitos anos, pois a base de dados transacional, que normalmente armazena informações de três meses a um ano, no máximo, não comportaria o armazenamento, e conseqüente gerenciamento, de terabytes de informações com performance aceitável.

Neste contexto, faz-se necessária a implementação de um sistema de Data Warehouse no arquivo de gerenciamento da escola, pela carência de flexibilidade e falta de autonomia dos usuários que o sistema estatístico possui.

A intenção não é a de substituir o sistema de administração da escola mas sim oferecer novas possibilidades de consulta aos dados disponíveis.

5.5 Seleção do Modelo do Data warehouse

Dentro do Sistema de administração da Escola Yázigi existem diversos tipos de informação, tais como controle do faturamento, de cobrança, de caixa, contabilidade, o controle de atendimentos a clientes, cadastro do aluno e outros. Todos estes controles foram implementados dentro dos quatro módulos do sistema, Secretaria, Administrativo, Financeiro e Pedagógico.

A seleção de uma área específica para desenvolver um foi tomada após uma série de análises, sendo que o escolhido foi o modulo da Secretaria que resulta na estatística mensal de atendimentos a prospects, matriculas e rematrículas feitas na escola.

A análise das áreas de negócio da escola pode ser feita buscando-se subsídios em cada um dos setores citados acima, detectando-se as vantagens da implantação de um sistema de apoio à decisão:

1. Os sistemas de controle de recebimento e cobrança dos alunos, localizados Módulo financeiro, poderiam ter sido utilizados pelo fato de que, segundo [KIM97], pode-se cativar os usuários finais do projeto global com respostas à pergunta: "Quanto Dinheiro Eles nos Devem?".

Com um sistema de apoio à decisão é possível ter o controle dos valores arrecadados e ainda investigar em que momentos e por quê estes valores estão acima ou abaixo dos totais de despesas em diversos períodos. Esta área é bastante utilizada para a definição de protótipos, pois é aquela que mede as entradas efetivas de capital na empresa e os montantes a receber que podem não estar sendo pagos pelos devedores;

2. O sistema de controle de caixa está localizado em uma área que monitora as quantias arrecadadas e gastas todos os dias pela empresa, e é controlado no Módulo Secretaria. Poderia ser feita a implementação com um modelo estrela formado pela tabela de fatos que representaria cada um dos lançamentos diários e que permitiria controlar, entre outras coisas, a relação entre a arrecadação e os gastos diários, mensais, trimestrais, anuais ou em qualquer período definido, auxiliando os tomadores de decisão

a analisar as diversas épocas do ano para saberem quando devem reter maiores quantias em caixa para suprirem eventuais baixas no faturamento líquido da empresa;

3. A área de contabilidade registra todos os fatos financeiros ocorridos na empresa através do lançamento de documentos. Um Data warehouse nesta área disponibiliza um grande volume de informações que geralmente estão disponíveis apenas em relatórios de papel. Se os altos executivos da empresa puderem ter à mão as informações e ainda poder analisá-las sob diversos ângulos, poderão ser determinados pontos positivos e negativos nas atividades da organização, sem que o executivo tenha que solicitar para o departamento financeiro estas informações, eliminando os intermediários e fornecendo os dados relativos a períodos de tempo que poderiam nem estar mais à mão dos funcionários do setor, uma vez que a legislação determina os prazos legais para a guarda dos livros;

4. O controle do atendimento a clientes gerencia todas as operações ligadas à atividade-fim da escola, uma vez que o produto que estes estabelecimentos oferecem é o serviço prestado aos alunos. Podemos dividir os atendimentos a clientes em dois grandes grupos: os prospects e os alunos matriculados. A diferença básica entre os dois grupos é que o primeiro é representado pelas pessoas interessadas em estudar na escola. Já os representantes do outro grupo são alunos efetivamente matriculados na escola. Relatórios em função de alunos matriculados em vários semestre são de suma importância para uma escola livre, pois as vendas de cursos de inglês são trabalhados em função de campanhas de rematrículas, em função de números é que são avaliadas se uma campanha foi bem ou mau sucedida.

Analisando cada uma das possibilidades chegou-se à conclusão de que um Data Warehouse na área de controle de atendimento seria interessante para a fase inicial do projeto, uma vez que já se tem um sistema de Administração em uso nesta área e, portanto, existe uma cultura de análise de informações, o que, em tese, quebraria as possíveis barreiras iniciais à implantação dos sistemas de apoio à decisão.

5.6 Seleção do Modelo de Dados

Foi escolhida a modelagem multidimensional porque como foi dito anteriormente , 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 [KIM96].

Com este modelo de dados podemos utilizar ferramentas baseadas em SQL. Podemos construir uma aplicação em um espaço menor de tempo.

5.7 Projeto do Sistema

O Projeto do Sistema é construir um protótipo de Data Warehouse para o setor de Secretaria das Escolas Yázigi de Florianópolis, onde serão definidas as diversas fases do projeto, conforme abaixo descritas [MOR98]:

a. definição do banco de dados;

b. construção do componente back end;

c. construção do componente front end;

d. definição do repositório de metadados.

A seguir são apresentados os passos para a construção do protótipo que permite realizar consultas com a ferramenta OLAP. Como a base de dados do sistema transacional (OLTP) de origem, não está definida neste mesmo padrão, será necessário desenvolver uma ferramenta de conversão destes dados para possibilitar a sua leitura para a base analítica.

5.8 Identificando as Origens dos Dados

O sistema transacional que está em uso na escola Yázigi que será utilizado como fonte foi desenvolvido em ZIM para Unix, mais precisamente em um ambiente Linux.

Neste sistema temos diversas tabelas relacionadas entre si e que servem basicamente para armazenar os dados relativos a administração da escola.

5.9 Identificando as Necessidades de Informações para Análise

Uma das finalidades principais da implantação de um sistema de apoio à decisão é permitir que os usuários finais possam realizar consultas, de maneira que algumas de suas dúvidas sejam esclarecidas pelo sistema na forma de respostas às consultas, bem como fornecer subsídios para que novas informações e situações sejam trazidas à tona.

Para que estes sistemas possam operar adequadamente é necessária uma definição detalhada dos dados disponíveis que serão carregados a partir da base transacional, além da definição das necessidades dos usuários finais, uma vez que um sistema somente poderá responder com base nas informações que estiverem armazenados no seu Data warehouse.

5.10 Definindo a Duração dos Dados

A duração dos dados refere-se ao período de tempo que serão mantidos os dados no Data warehouse. Existem sistemas analíticos que guardam informações de décadas, sendo que não é necessário guardar todos os dados de diversos anos.

Uma prática bastante usada é a de se guardar os últimos cinco anos na granularidade original dos dados, tal qual foram carregados originalmente, e, posteriormente, passando-os para dados mais resumidos, com uma granularidade menor, desde que atendam à maioria das consultas desejadas.

6.0 Conclusões

O presente trabalho seguiu os conceitos que cercam a teoria de Data Warehouse. Foram abordados diversos fatores, entre eles conceitos, objetivos, características e arquiteturas de sistemas de Data Warehouse.

Foi traçado um histórico dos sistemas de informação, foram apresentados os componentes fundamentais de Data Warehouse, os principais tipos de Data Warehouse, os objetivos, as suas diversas características, os diversos modelos e as suas diversas arquiteturas que podem ser utilizadas nos projetos de DW, além de fazer um comparativo entre os dois tipos de processamento de transações on-line, o processamento transacional e o processamento analítico, ficando bem claro que é bastante interessante separar as aplicações dedicadas ao controle das operações do dia-a-dia da empresa das aplicações destinadas a prover a análise dos dados obtidos destas operações.

Este trabalho serviu também para despertar o interesse para a área de Data Warehouse, uma área ainda carente de profissionais que projetem e implementem sistemas e, conseqüentemente, disponibilizem informações com grande qualidade às altas direções das empresas que cada vez mais necessitam desvincular-se da relação direta de dependência de seus subordinados para obterem as informações que norteiam suas tomadas de decisão. Este trabalho precisa de continuidade, ou seja, precisa que novos profissionais, possam a partir das especificações apresentadas, partir para uma implementação de um DW para Intranet, assim este projeto estaria realmente concluído.

7.0 Bibliografia

[ADE97] ADELMAN, S. e LEBARON, M. - Meta Data Standards, Review Magazine, Dezembro 1997.

[BAR96] BARQUINI, RAMON - Planning and designing the Warehouse, New Jersey, Prentice-Hall, 1996, 311 pg.

[BAT86] BATINI, C. e LENZERINI, M. - Comparative Analysis Of Methodologies For Database Schema

Integration, ACM Computing Surveys, New York, v.18, nº 4, pg.323-364, Dezembro 1986.

[BAU97] BAUER, A. e LEHNER, W. - The Cube-Query-Language (CQL) for Multidimensional Statistical and Scientific Database Systems, International Conference On Database Systems For Advanced Applications,Melbourne, Australia, pg.263-272, Maio 1997.

[VAL96] VALENTE, DAPHNIS LOPES – Estudo sobre Armazém de Dados, CPGCC da UFRGS, Porto Alegre, 1996.56 pg.

[BOH97] BOHN, K. – Convering Data for Warehouses DBMS, Califórnia, Junho 1997.

[CAM97] CAMPOS, MARIA LUIZA e ROCHA, ARNALDO V. – Data Warehouse, XVII Congresso da Sociedade Brasileira de Computação, XVI Jornada de Atualização em Informática, Rio de Janeiro, 1997, 261 pg.

[CEL95] CELKO, J. e MCDONAL, J. – Don't Warehouse dirty data. Datamation, New York, Outubro 1997.

[CAM97] Campos, Maria Luiza e Rocha Fº, Arnaldo V. Data Warehouse. XVII Congresso da Sociedade Brasileira de Computação. 1997

[CHA97] CHAUDHURI, S. e DAYAL, U. – An Overview of Data Warehousing and Olap Tecnology, SIGMODRecord, New York, v.26, nº 1, pg.65-74, Março 1997.

[COD95] CODD, E. F. Twelve Rules for On Line Analytical Processing, Computerworld, Abril 1995.

[DAL99] DalAlba, Adriano pela Internet em 26/06/1999.

www.geocities.com/SiliconValley/Port/5072/Index.htm

[DAT86] Date, C. Introdução a sistemas de bancos de dados.

3.ed. Rio de Janeiro: Campus, 1986. 513 pág.

[DAR96] DARLING, C. – How to integrate your Data Warehouse. Datamation, New York, v.42, nº 10, pg.40-51,Maio 1996.

[FLO97] FLOHR, U. – OLAP by Web. Byte, Peterborough, v.22, nº 9, pg.81-84, Setembro 1997.

[GEL96] GELMAN, S. e PECK, D. – Bringing Business Information to AT&T Network Systems Through a Data Warehouse. AT&T Technical Journal, New York, v.75, nº 2, pg.60-78, Abril 1996.

[HAC93] Hackathorn, R. D.

Enterprise database connectivity: the key to enterprise aplications on the desktop. New York: John Wiley & Sons, 1993. V.1, p.251-267.

[HAC95] HACKATHORN, R. – Data Warehousing Energizes Your Enterprise. Datamation, New York, v.41, nº 02,pg.38-43, Fevereiro 1997.

[HAR96] HARJINDER, G. e RAO, P. C. – The Officil Guide to Data Warehousing, Que Corporation, 1996.

[HYP98] Hyper Consultoria, site sobre a empresa Hyper Consultoria que presta serviços de consultoria em Sistemas de Apoio à Decisão, disponível na Internet via protocolo http, no endereço http://www.hyperinf.com.br/maestro.htm.

[INM93] INMON, W.H. – Information System Arquiteture: Development in 90's., John Wiley & Sons Inc., New York,1993.

[INM96] INMON, W.H. – Building the Data Warehouse, John Wiley & Sons Inc., New York, 1996.

[INM97] INMON, W.H. – Como Construir o Data Warehouse, Campus, Rio de Janeiro, 1997, 387 pg.

[INM97a] INMON, W.H. & RICHARD D. HACKATHORN – Como Usar o Data Warehouse, Infobook, Rio de Janeiro,1997.

[KIM95] KIMBALL, RALPH – Is E-R Modeling Hazardous to DSS?, DBMS, Outubro 1995.

[KIM96] KIMBALL, RALPH – The Data Warehouse Toolkit, John Wiley & Sons Inc., New York, 1996.

[KIM97] KIMBALL, RALPH – Dimensional Modeling Manifesto, DBMS, Agosto 1997.

[MCG98] McGUFF, FRANK, Designing the Perfect Data Warehouse – Hitchhiker's Guide to Decision Support, disponível na Internet via protocolo http, no endereço http://members.aol.com/fmcguff/ dwmodel, data do último acesso 22/11/1998.

[ONE97] ONEIL, B. – Oracle Data Warehousing. Indianapolis, Sams Publishing, 1997.

[TAN98] TANLER, RICK – Intranet Data Warehouse , Infobook, Rio de Janeiro, 1998.

[ORL96] ORLI, R. J. – Data Extraction, Transformation and Migration Tools, disponível na Internet via protocolo http, no endereço http://www.kismeta.com/ ex2.html, data do último acesso 21/11/1998.

[ORR96] ORR, KEN. – Data Warehouse Technology, The Ken Orr Institute – 1996, disponível na Internet via protocolo http, no endereço http://www. kenorrinst.com/datawh.html, data do último acesso 2/10/1998.

[RAD96] RADEN, N. – Technology Tutorial – Modeling a Data warehouse Value to organization means turning data into actionable information, Information Week, Issue 564, Janeiro 1996.

[RIB95] RIBEIRO, C. – Bancos de Dados Heterogêneos - Mapeamento dos Esquemas Conceituais em um modelo Orientado a Objetos, Tese de Doutorado, Porto Alegre, CPGCC da UFRGS, 1995.

[SPR91] SPRAGUE, H. e WATSON, J. – Sistema de apoio a decisão: colocando a teoria em prática. Ed.Campus, Rio de Janeiro, 1991, 497 pg.