DATA WAREHOUSE

INTRODUÇÃO

Informações importantes em uma organização, armazenadas em grandes bancos de dados, geralmente heterogêneas e distribuídas, são pouco aproveitadas para dar suporte à decisão. Tentando minimizar problemas de distribuição e heterogeneidade, no centro deste ambiente está o conceito de Data Warehouse.

A tecnologia de Data Warehouse surgiu principalmente devido às dificuldades que muitas organizações começaram a passar pela quantidade de dados que suas aplicações estavam gerando e à dificuldade de reunir estes dados de forma integrada para uma análise mais eficiente. A idéia, então, foi reunir em um único local, somente os dados considerados úteis no processo decisório.

Em um exemplo prático, suponhamos uma empresa de transporte aéreo. Através da tecnologia Data Warehouse pode-se obter a informação sobre qual mês do ano há uma maior procura por vôos para o Rio de Janeiro, ou ainda, para qual local os jovens com menos de vinte e cinco anos estão viajando através dos meios aéreos. Tendo em mãos essas informações em tempo hábil – em outras palavras, antes da concorrência – os executivos dessa organização podem dispor mais vôos para o Rio de Janeiro no mês de maior procura e, a respeito dos jovens, talvez fosse interessante disponibilizar algum tipo de lazer diferenciado durante a viagem.

De posse destas informações, os executivos/usuários do Data Warehouse dispõem de mecanismos que permitem, a partir de seu velho e volumoso banco

de dados, extrair dados que serão de grande utilidade e que darão maior lucratividade a médio-longo prazo.

O nosso exemplo se aplica a empresas privadas, mas o Data Warehouse também pode ser aplicada em organizações governamentais públicas. Tendo em mãos um Data Warehouse, o Secretário da Saúde, por exemplo, pode obter a informação de qual região da cidade ocorreram mais casos de dengue nos últimos cinco anos e, em quais meses desses anos, houve uma maior incidência desse vírus.

Os avanços da tecnologia de informação vieram garantir a possibilidade das organizações manipularem grandes volumes de dados e atingirem um alto índice de integração. Dados de todos os departamentos de uma organização podem estar em uma única base de dados, integrados, padronizados e resumidos para serem analisados pelos tomadores de decisões.

DESENVOLVIMENTO DO PROJETO DE DATA WAREHOUSE

EVOLUÇÃO DOS SISTEMAS DE APOIO À DECISÃO

Segundo Inmon (1997), a evolução dos sistemas de apoio a decisão pode ser dividida em cinco fases entre 1960 e 1980. No início da década de 1960 o mundo da computação consistia na criação de aplicações individuais que eram executadas sobre arquivos mestres, caracterizadas por programas e relatórios.

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

Por volta de 1970, surgiu a tecnologia DASD, substituindo as fitas magnéticas pelo armazenamento em disco. Com o DASD surgiu um novo tipo de software conhecido como SGBD ou sistema de gerenciamento de banco de dados, que tinha o objetivo de tornar o armazenamento e o acesso a dados no DASD mais fáceis para o programador.

Examinando a confusão criada pelos arquivos mestres e as enormes quantidades de dados redundantes ligadas a eles, não é de admirar que banco de dados seja definido como: *uma única fonte de dados para todo o processamento.* (Inmon, 1997).

Aproximadamente em 1975 surgiu o processamento de transações online. Com o processamento de transações online de alta performance, o computador pôde ser usado para tarefas que antes não eram viáveis como controlar sistemas de reservas, sistemas de caixas bancários, sistemas de controle de produção e outros.

Até o início da década de 1980, novas tecnologias, como os PCs e as L4Gs, começaram a aparecer. O usuário final passou a controlar diretamente os sistemas e os dados, descobrindo que era possível utiliza-los para outros objetivos além de atender ao processamento de transações online de alta performance. Foi nesse período também que se tornou viável a construção dos SIGs. Hoje conhecidos como SAD, os SIGs consistiam em processamento utilizado para direcionar decisões gerenciais.

A TEIA DE ARANHA

Após o advento das transações online de alta performance, começaram a surgir os programas de “extração”. Esses programas varrem arquivos de banco de dados usando alguns critérios, e, ao encontrar esses dados, transporta-os para outro arquivo de banco de dados.

Com a difusão do programa de extração, começou a formar-se a chamada “arquitetura de desenvolvimento espontâneo” ou “teia de aranha”, conforme mostrado na **Figura 3**. Primeiro havia extrações. Depois, extrações das extrações, e, então, extrações das extrações das extrações, e assim por diante.

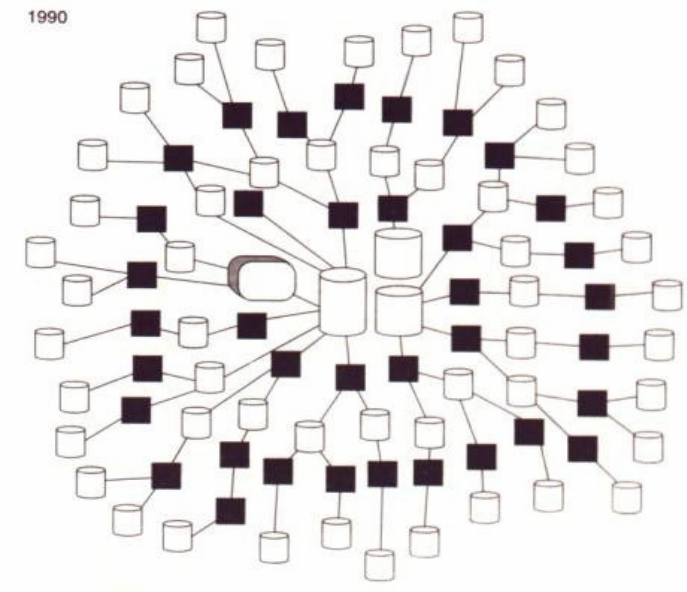

Figura 1 - A Teia de Aranha

Devido à arquitetura de desenvolvimento espontâneo, surgiram problemas com a credibilidade dos dados, a produtividade e a dificuldade de transformar dados puros em informações.

O AMBIENTE PROJETADO

A arquitetura de desenvolvimento espontâneo não era suficiente para atender as necessidades do futuro das empresas, fazendo-se necessário uma mudança de arquitetura, surgindo o ambiente projetado de Data Warehouse.

No cerne do ambiente projetado está a percepção de que há fundamentalmente duas espécies de dados – dados primitivos e dados derivados. A **Tabela 1** mostra algumas das principais diferenças entre dados primitivos e derivados.

Dados primitivos / Dados operacionais	Dados derivados / dados SAD
Baseado em aplicações	Baseados em assunto ou negócio
Detalhados	Resumidos ou refinados
Podem ser atualizados	Não são atualizados
São processados repetitivamente	Processados de forma heurística
Requisitos de processamento conhecidos com antecedência	Requisitos de processamento não são conhecidos com antecedência.
A performance é fundamental	Performance não é fundamental
Voltados para transação	Voltados para análise
Alta disponibilidade	Não é necessária alta disponibilidade
Atendem as necessidades cotidianas	Atendem as necessidades gerenciais
Alta taxa de acesso	Baixa ou média taxa de acesso

Tabela 1 - dados operacionais versus dados derivados

Dados primitivos e dados derivados devem estar fisicamente separados. *Há uma grande quantidade de diferenças entre dados primitivos e dados derivados. É espantoso que a comunidade de processamento de informações tenha pensado*

que dados primitivos e dados derivados pudessem se encaixar em um único banco de dados (Inmon, 1997).

Há quatro níveis no ambiente projetado – o operacional, a atômico ou Data Warehouse, o departamental e o individual, como representado na **Figura 4**. O nível operacional de dados contém apenas dados primitivos e atende à comunidade de processamento de transações de alta performance. O Data Warehouse contém dados primitivos que não são atualizados e dados derivados. O nível departamental de dados praticamente só contém dados derivados. E o nível individual de dados é onde o maior parte das análises heurísticas é feito.

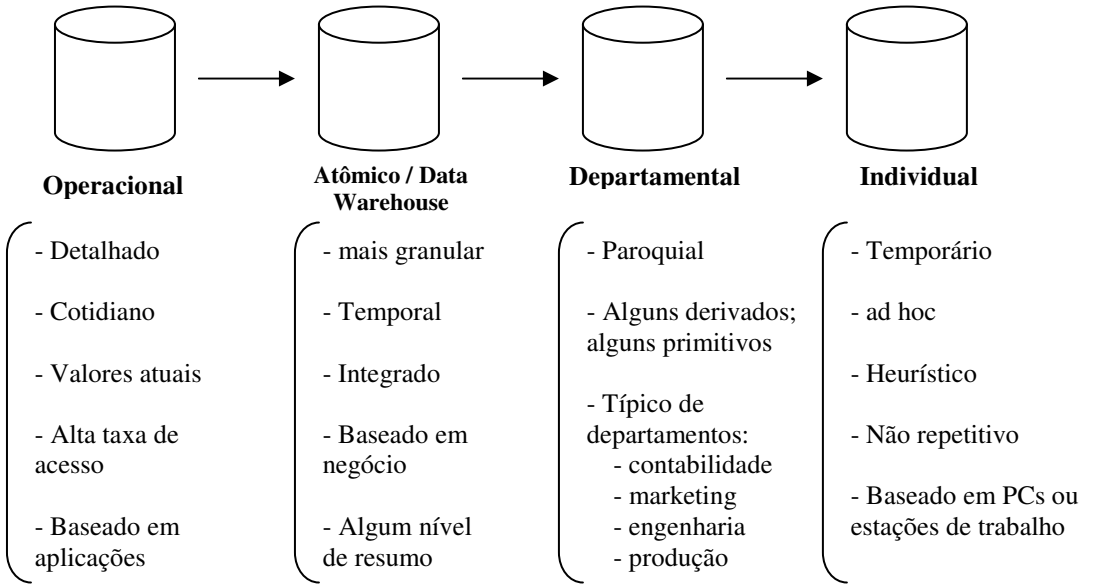

Figura 2 - Níveis do Ambiente Projetado

Um importante aspecto do ambiente projetado é a integração dos dados que ocorre ao longo da arquitetura. Se os dados chegarem ao Data Warehouse em um estado não integrado, não poderão ser utilizados como base para uma visão

corporativa dos dados. *A existência desta visão é um dos fundamentos do ambiente projetado* (Kimball, 1998).

O QUE É UM DATA WAREHOUSE

William H. Inmon foi um dos pioneiros no assunto Data Warehouse. Sua definição é a mais objetiva sobre o que é um Data Warehouse: *uma coleção de dados orientados por assunto, integrado, variável com o tempo e não-volátil, que tem por objetivo dar suporte aos processos de tomada de decisão* (Inmon, 1997).

Em outras palavras, um Data Warehouse é um banco de dados contendo dados extraídos do ambiente de produção da empresa, que foram selecionados e depurados, tendo sido otimizados para processamento de consulta e não para processamento de transações. Em geral, um Data Warehouse requer a consolidação de outros recursos de dados além dos armazenados em banco de dados relacionais, incluindo informações provenientes de planilhas eletrônicas, documentos textuais, etc.

Para Campos (1999), é importante considerar, no entanto, que um Data Warehouse não contem apenas dados resumidos, podendo conter também dados primitivos. É desejável prover ao usuário a capacidade de aprofundar-se num determinado tópico, investigando níveis de agregação menores ou mesmo dados primitivos, permitindo também a geração de novas agregações ou correlações com outras variáveis. Além do mais, é extremamente difícil prever todos os

possíveis dados resumidos que serão necessários: limitar o conteúdo de um Data Warehouse apenas a dados resumidos significa limitar os usuários apenas às consultas e análises que eles puderem antecipar frente a seus requisitos atuais, não deixando qualquer flexibilidade para novas necessidades.

Para ficar mais clara a concepção de Data Warehouse examina a **tabela 2** que contém uma comparação entre as características dos bancos de dados operacionais com um 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

Tabela 2 - Comparação entre banco de dados operacionais e Data Warehouse

O Data Warehouse é o alicerce do processamento dos SADs. Em virtude de haver uma única fonte de dados integrados, e uma vez que os dados apresentam

condições facilitadas de acesso e interpretação, a tarefa do analista de SAD no ambiente Data Warehouse fica incomensuravelmente mais fácil do que no ambiente clássico.

CARACTERÍSTICAS DE UM DATA WAREHOUSE

Quatro características principais regem o conceito de Data Warehouse.

Orientado por temas: R*efere-se ao fato do Data Warehouse armazenar informações sobre temas específicos importantes para o negócio da empresa. Exemplos típicos de temas são: produtos, atividades, contas, clientes, etc. Em contrapartida, o ambiente operacional é organizado por aplicações funcionais. Por exemplo, em uma organização bancária, estas aplicações incluem empréstimos, investimentos e seguros* (Campos, 1999).

Integrado: *Refere-se à consistência de nomes, das unidades das variáveis, etc, no sentido de que os dados foram transformados até um estado uniforme. Por exemplo, considere-se sexo como um elemento de dado. Uma aplicação pode codificar sexo como M/F, outra como 1/0 e uma terceira como H/M. Conforme os dados são inseirdos para o Data Warehouse, eles são convertidos para um estado uniforme, ou seja, sexo é codificado apenas de uma forma. Da mesma maneira, se um elemento de dado é medido em centímetros em uma aplicação, em polegadas em outra, ele será convertido para uma representação única ao ser colocado no Data Warehouse* (Campos, 1999).

Variante no tempo: *refere-se ao fato do dado em um Data Warehouse referir-se a algum momento específico, significando que ele não é atualizável, enquanto que o dado de produção é atualizado de acordo com mudanças de estado do objeto em questão, refletindo, em geral, o estado do objeto no momento do acesso. Em um Data Warehouse, a cada ocorrência de uma mudança, uma nova entrada é criada, para marcar esta mudança. O tratamento de séries temporais apresenta características específicas, que adicionam complexidade ao ambiente do Data Warehouse. Processamentos mensais ou anuais são simples, mas dias e meses oferecem dificuldades pelas variações encontradas no número de dias em um mês ou em um ano, ou ainda no início das semanas dentro de um mês. Além disso, deve-se considerar que não apenas os dados têm uma característica temporal, mas também os metadados, que incluem definições dos itens de dados, rotinas de validação, algoritmos de derivação, etc. Sem a manutenção do histórico dos metadados, as mudanças das regras de negócio que afetam os dados no Data Warehouse são perdidas, invalidando dados históricos* (Campos, 1999).

Não Volátil: *Significa que o Data Warehouse permite apenas a carga inicial dos dados e consultas a estes dados. 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. Esta volatilidade requer um trabalho considerável para assegurar integridade e consistência através de atividades de rollback, recuperação de falhas, commits e*

bloqueios. Um Data Warehouse não requer este grau de controle típico dos sistemas orientados a transações (Campos, 1999).

Granularidade: *diz respeito ao nível de detalhe ou de resumo contido nas unidades de dados existentes no Data Warehouse. Quanto maior o nível de detalhes, menor o nível de granularidade. O nível de granularidade afeta diretamente o volume de dados armazenado no Data Warehouse e ao mesmo tempo o tipo de consulta que pode ser respondida* (Campos, 1999).

USUÁRIOS TÍPICOS DE UM DATA WAREHOUSE

Inmon, Welch e Glassey (1999) identificaram três usuários típicos de um Data Warehouse: os fazendeiros, os exploradores e os turistas. A **Figura 5** ilustra os tipos de usuários.

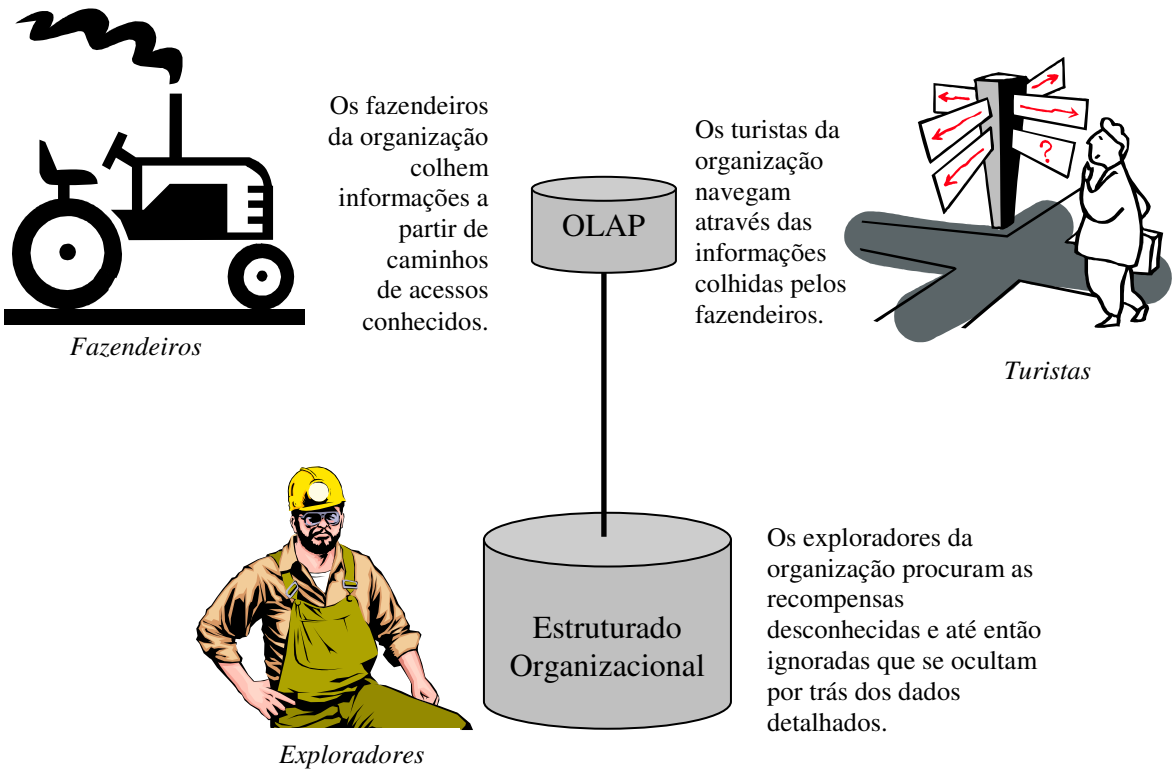

Figura 3 - Usuários do Data Warehouse

Como regra, os dados estruturados organizacional servem aos usuários fazendeiros e turistas. O dados detalhados servem aos usuários exploradores porque são orientados corporativamente, suportam acesso aleatório e são

completos e históricos. O ambiente OLAP (explicado mais adiante*) suporta os usuários fazendeiros porque os dados são personalizados antes de serem enviados ao ambiente OLAP. A fim de personalizar os dados, é necessário saber como os dados serão usados.; os fazendeiros tomam essas decisões com base no como os turistas consomem seus produtos. Em outras palavras, fornecimento e demanda aplicam-se à arquitetura do Data Warehouse na determinação do que deve ser populado no ambiente OLAP (*Inmon, Welch e Glassey, 1999).

Há diversas exceções a essa regra de diferentes usuários. Devido à quantidade limitada de dados lá encontrados, o grande número de índices e a elegância da interface, é possível executar explorações no ambiente OLAP. Contudo, a exploração no nível OLAP é superficial, e encontra uma visão geral e não detalhada. Na maioria das vezes, o ambiente OLAP existe e é perfeito para os usuários fazendeiros e turistas, mas não para a comunidade de exploradores.

ARQUITETURA DO DATA WAREHOUSE

Para ser útil o Data Warehouse 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 Data Warehouse 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 Data Warehouse é necessário aplicar filtros sobre eles.

O estudo de uma arquitetura permite compreender como o Data Warehouse faz para armazenar, integrar, comunicar, processar e apresentar os dados que os usuários utilizarão em suas decisões. Um Data Warehouse pode variar sua arquitetura conforme o tipo de assunto abordado, pois as necessidades também variam de empresa para empresa.

A **Figura 6** mostra os principais componentes da arquitetura de um Data Warehouse.

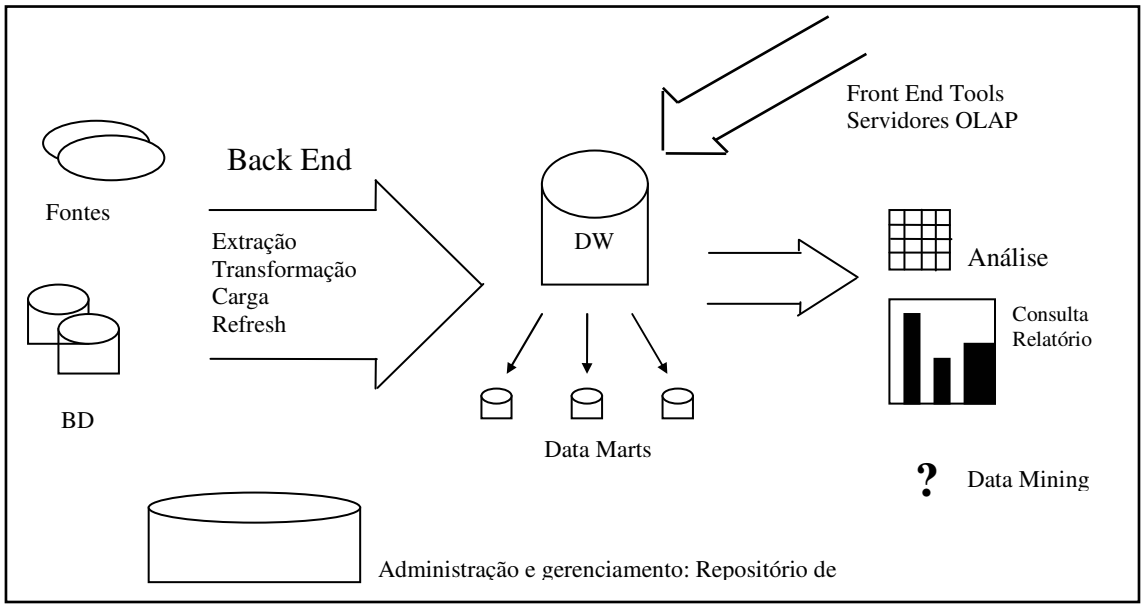

Figura 6 - Arquitetura do Data Warehouse

A arquitetura de um Data Warehouse inclui ferramentas para extrair dados de múltiplas bases de dados operacionais e fontes externas; limpar, transformar e integrar estes dados, carregá-los até o Data Warehouse e periodicamente fazer o refresh, isto é, propagar as atualizações ocorridas nas múltiplas base de dados operacionais. Em adição ao Data Warehouse principal, pode haver vários Data Warehouses departamentais, que são denominados Data Marts.

Dados no Data Warehouse e Data Marts são armazenados e gerenciados por um ou mais servidores de Data Warehouse, os quais apresentam visões multidimensionais de dados para uma variedade de ferramentas front end. Finalmente, há um repositório para armazenar e gerenciar metadados.

FERRAMENTAS BACK END

Sistemas de Data Warehouse usam uma variedade de ferramentas para extração, limpeza de dados, carga e refresh para “povoar” o banco de dados. Estas ferramentas são chamadas Back End e as principais funções desempenhadas por elas são:

Limpeza de dados: Já que o Data Warehouse é usado para tomada de decisão, é importante que os seus dados estejam corretos. Entretanto, uma vez que grandes volumes de dados estão envolvidos, há uma alta probabilidade de erros e anomalias nos dados. Tamanhos inconsistentes de campo, descrições inconsistentes, atribuição inconsistente de valores, entradas erradas e violação de restrições de integridade são alguns exemplos onde a limpeza de dados torna-se necessária.

Carga: Depois de extrair, limpar e transformar, os dados devem ser carregados para o Data Warehouse. Um pré-processamento adicional pode ser requerido, como por exemplo, checagem de restrições de integridade, sumarização, agregação, dentre outros mais. Tipicamente, batch load é usado para este propósito, isto é, o processo de carga é feito em lotes. A carga do Data Warehouse tem que lidar com volumes de dados muito maiores que os banco de dados operacionais.

Refresh: Fazer o refresh de um Data Warehouse consiste em propagar as atualizações ocorridas nos banco de dados operacionais para o banco de dados derivado do Data Warehouse.

FERRAMENTAS FRONT END

Segundo Moraes (1998), o componente *front end* de um sistema de Data Warehouse é o responsável por fornecer uma solução de acesso aos dados que atenda as necessidades por informações dos trabalhadores do conhecimento.

As ferramentas *front end* são utilizadas para análise, ajudando a interpretar o que ocorreu e a decidir sobre estratégias futuras. Neste tipo de aplicação, somente a operação de consulta se faz necessária.

As ferramentas Front End executam:

- Seleção do conjunto de dados necessários;
- Cálculo e manipulação dos dados;
- Apresentação das informações;

Os geradores de consultas e relatórios são considerados a primeira geração de ferramentas para o acesso a dados, as quais permitem a realização de consultas *ad-hoc*. Atualmente, as ferramentas de OLAP são as principais aplicações de suporte à decisão utilizadas em sistemas de Data Warehouse, sendo consideradas a segunda geração de ferramentas para acesso a dados. Ao

contrário dos geradores de consultas e relatórios, que apenas permitem uma visualização estática dos dados que não podem mais ser manipulados, as aplicações de OLAP possibilita que a partir de uma resposta se façam outros questionamentos, ou seja, o usuário consegue analisar o porquê dos resultados obtidos.

Moraes (1998), compilou a lista abaixo de características que possuem eficientes ferramentas de Front End.

- facilidades para acesso aos dados, manipulação e apresentação;
- capacidade de especificar consultas e relatórios com facilidade;
- suporte para a indústria de padrões de interface, incluindo Microsoft Windows GUI, ODBC, etc.
- suporte para o desenvolvimento de interfaces amigáveis;
- habilidade para acessar a funcionalidade nativa de uma variedade de BD e outras origens de dados;
- habilidade para suportar uma variedade de plataformas servidoras e SGBDs.

DATA MARTS

Um Data Mart é um sistema de suporte a decisão que incorpora um subconjunto de dados da empresa focalizado em funções ou atividades específicas da organização. Os Data Marts têm propósitos específicos relacionados ao negócio, como medida do impacto de promoções de marketing, medida ou previsão de vendas, medida do impacto da introdução de novos produtos, etc.

Data Marts podem incorporar dados substanciais, mas eles contém muito menos dados que teria um Data Warehouse desenvolvido para a mesma organização. *Uma vez que Data Marts são focalizados em propósitos específicos do negócio, o planejamento do sistema e a análise dos requerimentos são mais facilmente gerenciáveis, e o projeto, implementação, fase de testes e instalação são bem mais baratos que para um Data Warehouses* (Inmon, Welch e Glassey, 1999). Por esse motivo, os Data Marts estão se tornando uma alternativa bastante popular nos últimos anos.

Os projetos de Data Marts devem ser inicialmente simples e úteis para que possam atingir seus objetivos de forma rápida e clara. Não é desejável para uma empresa investir uma quantia em dinheiro e tempo de seus funcionários em um projeto que pode levar meses para ser concluído e que durante o processo de implantação possa terminar por gerar controvérsias e até mesmos problemas para os setores (Kimball, 1998).

DATA MINING

Data Mining é uma ferramenta de extração de dados. O Data Mining engloba um número de diferentes abordagens técnicas, como clustering (agrupamento), sumarização de dados, regras de classificação, detecção de anomalias, etc.

Data Mining é uma categoria de ferramentas de análise. Em vez de se fazerem perguntas, entrega-se grandes quantidades de dados e pergunta-se se existe algo de interessante (uma tendência ou um agrupamento, por exemplo). O processo de mineração de dados pode extrair conhecimento que está escondido ou informações de prognóstico do Data Warehouse sem a necessidade de consultas específicas ou requisições.

Esse processo de mineração usa técnicas avançadas como redes neurais, heurísticas, descoberta por regra e detecção de desvio. Ao contrário de relatórios e consultas cujos relacionamentos já se conhece, o trabalho do Data Mining é descobrir o que não se sabe que existe no banco de dados.

Alguns exemplos de aplicações de Data Mining:

- identificar padrões de compra dos clientes;
- identificar correlações escondidas entre diferentes indicadores financeiros;
- identificar superfaturamento em grandes obras públicas.

SISTEMAS GERENCIADORES DE BANCOS DE DADOS

SGBDs têm como função fornecer acesso e manipulação eficientes aos dados armazenados no banco, proteger estes dados contra acessos indevidos e manter sua consistência e integridade (Moraes, 1998).

Os SGBDs em sistemas de Data Warehouse devem suportar processamento analítico on-line (OLAP), ao contrário do já tradicional processamento de transações on-line (OLTP). Os SGBDs voltados ao processamento de transações têm como principal característica dar suporte para atualizações concorrentes de centenas de usuários. Já os SGBDs voltados para sistemas de Data Warehouse devem ser otimizados para o processamento de consultas complexas e ad-hoc.

Três classes de SGDBs devem ser citadas:

a) SGBDs relacionais tradicionais:

A tecnologia relacional vem sendo amplamente reconhecida como a melhor alternativa para a hospedagem de dados em sistemas de Data Warehouse. Rapidamente, as melhorias dos SGBDs na área de suporte à decisão vêm atendendo as necessidades impostas pelo ambiente de Data Warehouse. Isto se deve, principalmente, a dois principais pontos fracos dos SGBDs multidimensionais: inflexibilidade (estrutura de arquivos proprietária) e limitado volume de dados que podem gerenciar.

b) SGBDs multidimensionais (MOLAP):

Em um banco de dados multidimensional, em vez de armazenar registros em tabelas, eles armazenam os dados em matrizes. São projetados com o objetivo de permitir uma eficiente e conveniente armazenagem e recuperação de dados que estão intimamente relacionados. Estes dados são armazenados, visualizados e analisados segundo diferentes dimensões.

O grande problema dos SGBDs multidimensinais é a sua capacidade de armazenamento ainda limitada para as necessidades de um Data Warehouse. Desta forma, estes produtos são mais utilizados no mercado como gerenciadores de Data Marts.

c) SGBDs relacionais especializados para sistemas de Data Warehouse:

São otimizados para atender ambientes de somente leitura (read only), onde o processamento eficiente de consultas é importantíssimo. A idéia nestes produtos é abandonar os requisitos necessários ao processamento de transações (OLTP) e se concentrar nos requisitos necessários ao OLAP. Desta forma, estes SGBDs fornecem novas técnicas de otimização de consultas sobre estruturas do tipo “star scheme”, utilizam novos métodos de indexação e interpretam a sintaxe SQL para dar suporte a consultas que são importantes no ambiente de Data Warehouse.

MODELO DE DADOS

Obter respostas a questões típicas de análise dos negócios de uma empresa geralmente requer a visualização dos dados segundo diferentes perspectivas.

Suponhamos uma grande rede de hotéis que deseja melhorar o desempenho de seu negócio. Para isso, necessita examinar os dados sobre as reservas e seus clientes. Uma avaliação deste tipo requer uma visão histórica do volume de reservas informações sobre seus clientes sob múltiplas perspectivas, como por exemplo: qual a idade média de seus clientes, qual o período médio que os mesmos se hospedam no hotel. Uma análise da idade média de seus clientes utilizando uma ou mais destas perspectivas, permitiria responder questões do tipo:

Qual a idade média dos hóspedes na temporada de final de ano?

Tendo em mãos a resposta para essa questão, a gerência do hotel poderia investir no marketing para um cliente-alvo mais preciso. A capacidade de responder a este tipo de questão em tempo hábil é o que permite aos gerentes e altos executivos das empresas formular estratégias efetivas, identificar tendências e melhorar sua habilidade de tomar decisões de negócio. O ambiente tradicional de bancos de dados relacional certamente pode atender a este tipo de consulta. No entanto, *usuários finais que necessitam de consultas deste tipo, via acesso interativo aos bancos de dados, mostram-se frustrados por tempos de resposta ruins e pela falta de flexibilidade oferecida por ferramentas de consulta baseadas no SQL* (Kimball, 1998).

Daí a necessidade de utilizar abordagens específicas para atender a estas consultas.

A mais importante diferença entre sistemas OLTP e Data Warehouse está no modelo de dados. O tradicional modelo Entidade-Relacionamento divide os dados em várias entidades distintas, cada uma transformada em uma tabela do Banco de Dados OLTP. Há algumas observações a fazer sobre o diagrama entidade-relacionamento.

Em primeiro lugar, ele é muito simétrico. *Todas as tabelas parecem iguais; esses diagramas são difíceis de visualizar e memorizar tanto pelo usuário final quanto pelos projetistas* (Kimball, 1998). Segundo, quando duas tabelas do diagrama são necessárias para uma consulta, há um número imenso de conexões possíveis entre as duas tabelas. Em consultas que abrangem muitas tabelas e registros, os diagramas Entidade-Relacionamento tornam-se muito complexos tanto para o usuário entender quanto para o software navegar. Dito isto, pode-se concluir que modelos Entidade-Relacionamento são um desastre para ambientes *read only* (somente consulta) e não são propícios para serem utilizados como base para o Data Warehouse.

A representação dos dados em um Data Warehouse é estruturada como um “cubo de dados”. Essa estrutura é chamada modelo dimensional, também conhecida como *star scheme*. Ao contrário do modelo Entidade-Relacionamento, o modelo dimensional é muito assimétrico. Há uma tabela dominante no centro do diagrama com múltiplas junções a conectando nas outras tabelas. Cada uma das tabelas

secundárias possui apenas uma junção com a tabela central. A tabela central é chamada de “tabela de fatos” e as outras tabelas de “tabelas de dimensão”, como mostra a **Figura 7**:

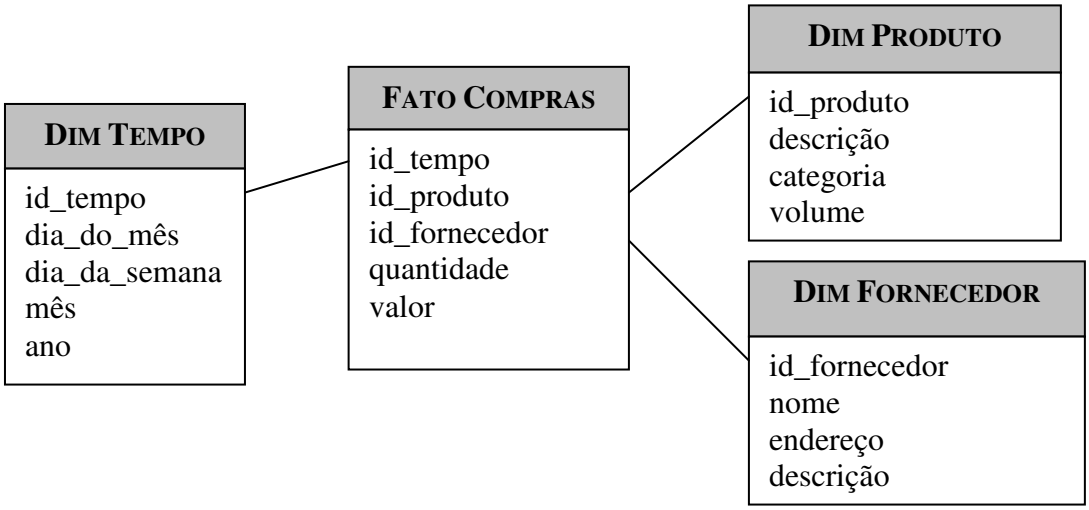

Figura 5 - Star Scheme

A tabela de fatos armazena medições numéricas do negócio. Cada uma das medições é obtida na interseção de todas as dimensões. Os fatos melhores e mais úteis são numéricos, valorados (diferentes a cada medida) e aditivos (podem ser adicionados ao longo das dimensões). O motivo para utilizar fatos valorados e aditivos é que em praticamente todas as consultas feitas à tabela de fatos, são solicitados centenas ou milhares de registros para construir o conjunto de resposta. Esse grande número de registros será compactado em algumas dezenas de linhas para produzir o conjunto de resposta do usuário. A única forma viável de compactá-los no conjunto de resposta será adicioná-los. Portanto, se as medições forem números e se forem aditivas, pode-se construir facilmente o conjunto de resposta.,

As tabelas dimensionais armazenam as descrições textuais das dimensões. Esses atributos textuais são usados como restrições e cabeçalhos de linha no conjunto de resposta.

Ao se projetar um banco de dados, pode-se ficar na dúvida se um campo de dados será modelado como um fato ou um atributo. Segundo Kimball (1998), se o dado for numérico e variar continuamente a cada amostragem, ele será considerado um fato. Do contrário, se for uma descrição praticamente constante de um item, será considerada um atributo de dimensão.

O *Star Scheme* tem uma série de vantagens que são descritas abaixo:

- O Star Scheme 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 dimensões 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 Data Warehouse.

DESENVOLVIMENTO DE UM DATA WAREHOUSE

O sucesso do desenvolvimento de um Data Warehouse depende fundamentalmente de uma escolha correta da estratégia a ser adotada, de forma que seja adequada às características e necessidades específicas do ambiente onde será implementado. Existe uma variedade de abordagens para o desenvolvimento de Data Warehouses, devendo-se fazer uma escolha fundamentada em pelo menos três dimensões: escopo (departamental, empresarial, etc), grau de redundância de dados, tipo de usuário alvo.

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

Quanto à redundância de dados, há essencialmente três níveis de redundância: o Data Warehouse virtual, o Data Warehouse centralizado e o data warehouse distribuído.

O Data Warehouse virtual consiste em simplesmente prover os usuários finais com facilidades adequadas para extração das informações diretamente dos

bancos de produção, não havendo assim redundância, mas podendo sobrecarregar o ambiente operacional.

O Data Warehouse central constitui-se em um único banco de dados físico contendo todos os dados para uma área funcional específica, um departamento ou uma empresa, sendo usados onde existe uma necessidade comum de informações. Um Data Warehouse central normalmente contém dados oriundos de diversos bancos operacionais, devendo ser carregado e mantido em intervalos regulares.

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

Os padrões de uso de um Data Warehouse também constituem um fator importante na escolha de alternativas para o ambiente. Relatórios e consultas pré-estruturadas podem satisfazer o usuário final, e geram pouca demanda sobre o SGBD e sobre o ambiente servidor. Análises complexas, por sua vez, típicas de ambientes de suporte à decisão, exigem mais de todo o ambiente.

Ambientes dinâmicos, com necessidades em constante mudança, são mais bem atendidos por uma arquitetura simples e de fácil alteração, ao invés de uma estrutura mais complexa que necessite de reconstrução a cada mudança. A freqüência da necessidade de atualização também é determinante: grandes

volumes de dados que são atualizados em intervalos regulares favorecem uma arquitetura centralizada.

ESTRATÉGIA EVOLUCIONÁRIA

Data Warehouses, em geral, são projetados e carregados passo a passo, seguindo, portanto uma abordagem evolucionária. Os custos de uma implementação "por inteiro", em termos de recursos consumidos e impactos no ambiente operacional da empresa justificam esta estratégia.

Muitas empresas iniciam o processo a partir de uma área específica da empresa, que normalmente é uma área carente de informação e cujo trabalho seja relevante para os negócios da empresa, criando os chamados Data Marts, para depois ir crescendo aos poucos, seguindo uma estratégia "botton-up" ou assunto-por-assunto.

Outra alternativa é selecionar um grupo de usuários, prover ferramentas adequadas, construir um protótipo do Data Warehouse, deixando que os usuários experimentem com pequenas amostras de dados. Somente após a concordância do grupo quanto aos requisitos e funcionamento, o Data Warehouse será de fato alimentado com dados dos sistemas operacionais na empresa e dados externos.

Data Marts também pode ser criados como subconjunto de um Data Warehouse maior, em busca de autonomia, melhor desempenho e simplicidade de compreensão.

ASPECTOS DE MODELAGEM

A especificação de requisitos do ambiente de suporte à decisão associado a um Data Warehouse é fundamentalmente diferente da especificação de requisitos dos sistemas que sustentam os processos usuais do ambiente operacional de uma empresa.

Os requisitos dos sistemas do ambiente operacional são claramente identificáveis a partir das funções a serem executadas pelo sistema. Requisitos de sistemas de suporte à decisão são, por sua vez, indeterminados.

O objetivo por trás de um Data Warehouse é prover dados com qualidade; os requisitos dependem das necessidades de informação individuais de seus usuários. Ao mesmo tempo, os requisitos dos sistemas do ambiente operacional são relativamente estáveis ao longo do tempo, enquanto que os dos sistemas de suporte à decisão são instáveis. No entanto, embora as necessidades por informações específicas mudem com freqüência, os dados associados não mudam. Imaginando-se que os processos de negócio de uma empresa permaneçam relativamente constantes, existe apenas um número finito de objetos e eventos com as quais uma organização está envolvida.

Por esta razão, o modelo de dados é uma base sólida para identificar requisitos para um Data Warehouse.

ETAPAS DO DESENVOLVIMENTO DE UM DATA WAREHOUSE

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

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

Desenvolver um Data Warehouse é uma questão de casar as necessidades dos seus usuários com a realidade dos dados disponíveis. Abaixo podemos analisar os chamados pontos de decisão, que constituem definições a serem feitas e correspondem a etapas do projeto:

1. Os processos, e por conseqüência, a identidade das tabelas de fatos;
2. A granularidade de cada tabela de fatos;
3. As dimensões de cada tabela de fatos;
4. Aos fatos, incluindo fatos pré-calculados;
5. Os atributos das dimensões;
6. Como acompanhar mudanças graduais em dimensões;

7. As agregações, dimensões heterogêneas, minidimensões e outras decisões de projeto físico;

8. Duração histórica do banco de dados;

9. A urgência com que se dá a extração e carga para o Data Warehouse.

Esta metodologia segue a linha top-down, pois começa identificando os grandes processos da empresa.

EXTRAINDO INFORMAÇÕES DE UM DATA WAREHOUSE

Existem várias maneiras de recuperar informações de um data Warehouse. As formas de extração mais comuns no mercado hoje são:

- Ferramentas de consulta e emissão de relatórios;

- EIS (Executive Information Systems);

- Ferramentas OLAP;

- Ferramentas Data mining.

A nova tendência dessas soluções é a integração com o ambiente Web, permitindo maior agilidade em consultas estáticas e dinâmicas.

A seguir veremos de forma básica e separadamente os conceitos das tecnologias OLAP e Data Mining. A diferença básica entre ferramentas OLAP e Data Mining está na maneira como a exploração dos dados é abordada.

Com ferramentas OLAP a exploração é feita na base da verificação, isto é, o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta para confirmá-la.

Com Data Mining, a questão é total ou parcialmente desconhecida e a ferramenta é utilizada para a busca de conhecimento.

FERRAMENTAS OLAP

OLAP – On-Line Analytical Processing – representa um conjunto de tecnologias projetadas para suportar análise e consultas ad hoc. Sistemas OLAP ajudam analistas e executivos a sintetizarem informações sobre a empresa, através de comparações, visões personalizadas, análise histórica e projeção de dados em vários cenários de "e se...".

Os sistemas OLAP são implementados para ambientes multi-usuário, arquitetura cliente-servidor e oferecem respostas rápidas e consistentes às consultas iterativas executadas pelos analistas, independente do tamanho e complexidade do banco de dados.

A característica principal dos sistemas OLAP é permitir uma visão conceitual multi-dimensional dos dados de uma empresa. *A visão multi-dimensional é muito mais útil para os analistas do que a tradicional visão tabular utilizada nos sistemas de processamento de transação. Ela é mais natural, fácil e intuitiva, permitindo a visão em diferentes perspectivas dos negócios da empresa e desta maneira tornando o analista um explorador da informação* (Bispo e Cazarini, 1999).

A modelagem dimensional é a técnica utilizada para se ter uma visão multi-dimensional dos dados. Nesta técnica os dados são modelados em uma estrutura dimensional conhecida por cubo. As dimensões do cubo representam os componentes dos negócios da empresa tais como "cliente", "produto", "fornecedor" e "tempo". A célula resultante da interseção das dimensões é chamada de medida e geralmente representa dados numéricos tais como "unidades vendidas", "lucro" e "total de venda". Além dos componentes dimensão e medida outro importante aspecto do modelo multi-dimensional é a consolidação dos dados uma vez que para a tarefa de análise são mais úteis e significativas as agregações (ou sumarização) dos valores indicativas dos negócios.

Além da visão multi-dimensional dos dados da empresa, a tecnologia OLAP tem uma série de outras características importantes relacionadas abaixo:

- Análise de tendências. A tecnologia OLAP é mais do que uma forma de visualizar a história dos dados. Deve, também, ajudar os usuários a tomar decisões sobre o futuro, permitindo a construção de cenários ("e

se...") a partir de suposições e fórmulas aplicadas, pelos analistas, aos dados históricos disponíveis;

- Busca automática (reach-through) de dados mais detalhados que não estão disponíveis no servidor OLAP. Detalhes não são normalmente importantes na tarefa de análise, mas quando necessários, o servidor OLAP deve ser capaz de buscá-los;

- Dimensionalidade genérica;

- Operação trans-dimensional. Possibilidade de fazer cálculos e manipulação de dados através diferentes dimensões;

- Possibilidade de ver os dados de diferentes pontos de vista (slice and dice), mediante a rotação (pivoting) do cubo e a navegação (drill-up/drill-down) entre os níveis de agregação;

- Conjunto de funções de análise e cálculos não triviais com os dados.

Segundo Inmon, Welch e Glassey (1999), existe também um conjunto de regras que servem para avaliar as ferramentas OLAP):

- Visão conceitual multidimensional;

- Transparência;

- Acessibilidade;

- Performance de Relatório consistente;
- Arquitetura cliente-servidor;
- Dimensionalidade genérica;
- Operação dimensional cruzada irrestrita;
- Manipulação de dados intuitiva;
- Flexibilidade quanto a relatórios;
- Dimensão e níveis de agregamentos ilimitados;
- Pesquisa de detalhes (drill down);
- Atualização incremental do banco de dados;
- Arrays múltiplos;
- Seleção de subconjuntos;
- Suporte a dados locais.

Uma arquitetura OLAP possui três componentes principais: um modelo de negócios para análises interativas, implementado numa linguagem gráfica que permita diversas visões e níveis de detalhes dos dados; um motor OLAP para processar consultas multidimensionais contra o dado-alvo; e um mecanismo para armazenar os dados a serem analisados.

MOLAP x ROLAP

Multidimensional OLAP (MOLAP) é uma classe de sistemas que permite a execução de análises sofisticadas usando como gerenciador de dados um banco de dados multidimensional. Em um banco de dados MOLAP os dados são mantidos em arranjos e indexados de maneira a prover uma ótima performance no acesso a qualquer elemento. O indexamento, a antecipação da maneira como os dados serão acessados e o alto grau de agregação dos dados faz com que sistemas MOLAP tenham uma excelente performance. Além de serem rápidos, outra grande vantagem destes sistemas é o rico e complexo conjunto de funções de análise que oferecem.

A maneira de se implementar os arranjos de dados pode variar entre fornecedores de soluções MOLAP. Existem as arquiteturas hiper-cubos e multi-cubos. Na arquitetura hiper-cubo existe um único cubo onde cada medida é referenciada por todas as outras dimensões. Por exemplo, um cubo onde a medida "compras" é referenciada pelas dimensões "produto", "ano", "mes", "estado" e "cidade".

Na arquitetura multi-cubos uma medida é referenciada por dimensões selecionadas. Em um cubo, a medida "vendas" é referenciada pelas dimensões "semestre", "estado" e "produto" e em outro cubo, a medida "custo" é referenciada pelas dimensões "mês" e "departamento". Esta arquitetura é escalável e utiliza menos espaço em disco. A performance é melhor em cada cubo individualmente, no entanto, consultas que requerem acesso a mais de um cubo podem exigir processamentos complexos para garantir a consistência do tempo de resposta.

Sistemas ROLAP fornecem análise multidimensional de dados armazenados em uma base de dados relacional. Existem duas maneiras de se fazer este trabalho:

- Fazer todo o processamento dos dados no servidor da base de dados. O servidor OLAP gera os comandos SQL em múltiplos passos e as tabelas temporárias necessárias para o processamento das consultas;

- Ou executar comandos SQL para recuperar os dados, mas fazer todo o processamento (incluindo joins e agregações) no servidor OLAP.

A principal vantagem de se adotar uma solução ROLAP reside na utilização de uma tecnologia estabelecida, de arquitetura aberta e padronizada como é a relacional, beneficiando-se da diversidade de plataformas, escalabilidade e paralelismo de hardware.

FERRAMENTAS DATA MINING

Segundo Pinheiros (1999), nos primórdios do Data Warehouse, o Data Mining era visto como um subconjunto das atividades associadas com o Data Warehouse. Mas atualmente os caminhos do Data warehouse e do Data Mining estão divergindo. Enquanto o Data Warehouse pode ser uma boa fonte de dados para minerar, o Data Mining foi reconhecido como uma tarefa genuína, e não mais como uma colônia do Data Warehouse.

Apesar do termo Data Mining ter se tornado bastante popular nos últimos anos, existe ainda uma certa confusão quanto à sua definição. Data Mining (ou

mineração de dados) é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados, usando-as para efetuar decisões cruciais. Data Mining vai muito além da simples consulta a uma banco de dados, no sentido de que permite aos usuários explorar e inferir informação útil a partir dos dados, descobrindo relacionamentos escondidos no banco de dados. Pode ser considerada uma forma de descobrimento de conhecimento em bancos de dados (KDD - Knowledge Discovery in Databases), área de pesquisa de bastante evidência no momento, envolvendo Inteligência Artificial e Banco de Dados (Campos, 1999).

Um ambiente de apoio à tomada de decisões, integrando técnicas de Data Mining sobre um ambiente de Data Warehousing, possibilita um grande número de aplicações, que já vêm sendo implementadas em diversos segmentos de negócios, como manufatura, automação de pedido de remessas, varejo, gerenciamento de inventários, financeiro, análise de risco, transporte, gerenciamento de frotas, telecomunicação, análise de chamadas, saúde, analise de resultados, markenting, estabelecimento do perfil dos consumidores, seguros, detecção de fraude, dentre outros (Pinheiros, 1999).

O Data Mining pode ser utilizado com os seguintes objetivos:

- Explanatório: explicar algum evento ou medida observada, tal como porque a venda de sorvetes caiu no Rio de Janeiro;

- Confirmatório: confirmar uma hipótese. Uma companhia de seguros, por exemplo, pode querer examinar os registros de seus clientes para determinar se famílias de duas rendas tem mais probabilidade de adquirir um plano de saúde do que famílias de uma renda;

- Exploratório: analisar os dados buscando relacionamentos novos e não previstos. Uma companhia de cartão de crédito pode analisar seus registros históricos para determinar que fatores estão associados a pessoas que representam risco para créditos.

O diferencial do Data Mining está no fato de que as descobertas de padrões de consumo se dão por uma lógica de algoritmos com base em uma rede neural de raciocínios. São ferramentas de descobertas matemáticas feitas sobre os registros corporativos já processados contra descobertas empíricas.

CARACTERÍSTICAS DE UM DATA WAREHOUSE BEM-SUCEDIDO

O que pode ser feito para criar um ambiente de análise de dados moderno no qual os usuários possam embarcar numa viagem aleatória e direta? Segundo Inmon, Welch e Glassey (1999) há quatro objetivos-chave que devem ser alcançados para um Data Warehouse ser considerado bem-sucedido.

- Fornecer modos melhores e mais rápidos para que os usuários descubram as respostas a questões complexas e imprevisíveis.

- Colocar os usuários em contado direto com os dados de que precisam para tomar decisões melhores.

- Permitir que os usuários tornem-se responsáveis pela especificação, criação e geração repetida dos relatórios e análises que necessitem.

- Contar com uma manutenção apropriada e responsável dos recursos de dados corporativos.

O sistema que satisfaz esses objetivos é um sistema de suporte a decisões moderno. Os projetos de Data Warehouse obtêm sucesso quando os usuários são mais independentes. Data Warehouses bem-sucedidos colocam os usuários no centro do projeto. Quando todos reconhecem isso, uma nova atitude e abordagem são os ingredientes mais bem-sucedidos nessa mistura. As organizações que entendem esses fatores fundamentais que estão conduzindo a alterações no paradigma terão sucesso em estabelecer Data Warehouses bem-sucedidos (Inmon, Welch e Glassey, 1999).