MATERIAL DE APOIO

DATA WAREHOUSE

CONCEITOS E APLICAÇÕES

SENAI

Prof. FERNANDO FIGUEIREDO DOS SANTOS

06/10/2020

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 da Data Warehouse dispõem de mecanismos que permitem, a partir de seu velho e volumoso banco de dados, extraírem 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 aplicado 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.

Surgimento do Data Warehouse

A arquitetura warehousing surgiu da noção de que deveria haver uma divisão entre tipos diferentes de bancos de dados. O princípio da tecnologia de banco de dados afirmava que deve haver um único banco de dados para todos os tipos de processo, mas a realidade mostrou, por uma série de razões, que é necessário desenvolver mais de um tipo de banco de dados para satisfazer necessidades distintas nas organizações. Esta divisão entre tipos de banco de dados foi classificada em Banco de Dados Operacional e Bancos de Dados Warehousing e aconteceu por diversas razões, sendo que as principais foram:

- Um banco de dados operacional requer, para ser eficiente, o que se chama de tempo de resposta, que significa agilidade na recuperação das informações. Um DW, que trata de processamento do tipo DSS (Decision Support System) não precisa desse recurso;
- Usuários em geral utilizam transações orientadas para bancos de dados operacionais; já os gestores devem utilizar um DW;

- Decisões de curto prazo advêm de Sistemas de Informação convencionais; já decisões de longo prazo são geradas a partir de informações extraídas de um DW;
- Um banco de dados operacional contém informação muito atual; um DW contém informação histórica;
- Um banco de dados operacional não integra uma informação através do conhecimento gerado pelos sistemas de informação da organização; um DW contém dados que são integrados e fazem parte do mesmo todo;
- Um banco de dados operacional é projetado para dados detalhados; um DW é projetado para armazenar, tanto dados detalhados quanto dados sumarizados;
- As exigências do processamento de um sistema convencional, usando um banco de dados operacional, são conhecidas antes desse sistema ser construído; no caso de um DW, as exigências são descobertas durante o processo de desenvolvimento;
- As exigências de processamento em um ambiente operacional convencional são estáticas; já as exigências de processamento em um ambiente DW são heurísticas, ou seja, são descobertas pela interatividade no processo de desenvolvimento.

Como pôde ser notado, da necessidade de se atender a novos requisitos dentro de uma organização surgiu a divisão entre os dois tipos de bancos de dados. O DW surgiu, então, a partir das limitações de um banco de dados operacional (convencional) e para atender a uma nova necessidade, a de suprir as empresas de informações para sistemas DSS (Decision Support System) e CRM (Costumer Relationship Management).

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 1. Primeiro havia extrações. Depois, extrações das extrações, e, então, extrações das extrações das extrações, e assim por diante.

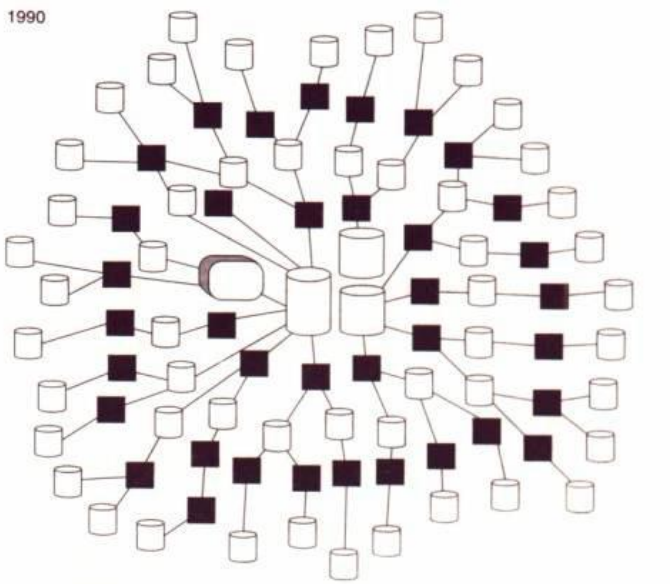

Figura 1 - A Teia de Aranha

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 2**. 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.

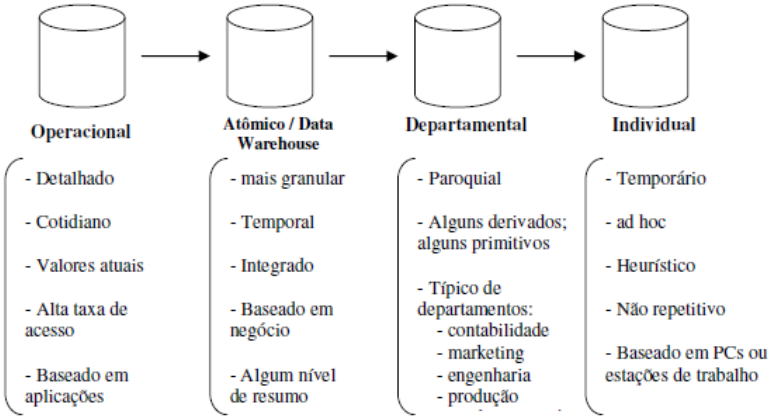

Figura 2 - Níveis do Ambiente Projetado

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

Cinco características principais regem o conceito de Data Warehouse.

Orientado por temas: Refere-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).

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 3** mostra os principais componentes da arquitetura de um Data Warehouse.

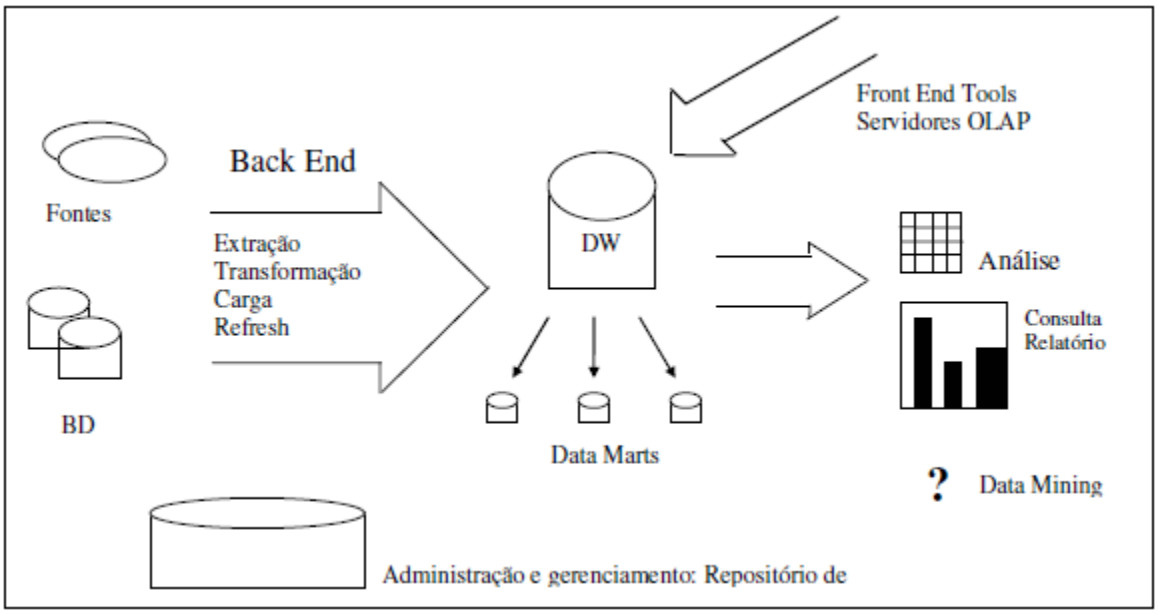

Figura 3 - Arquitetura do Data Warehouse

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.

Vídeo de apoio: https://youtu.be/UYqqGcMKFW8

Exercício de fixação

Pesquise e descreva um exemplo de aplicação do Data Warehouse, elencando todos os seus conceitos.