2004

Paper colaborativo

DW & DM – DATA WAREHOUSE e DATA MINING

Autoria: Adriana, Fernando, Josip, Ricardo e Pedro
São Paulo, 30/06/2004

Sumário

Datawarehouse

Introdução

DATAWAREHOUSE (ou *Data Warehouse*) é um imenso banco de dados que integra diversos outros bancos de dados, daí a sua tradução literária ser "Fábrica" ou "Armazém" de dados.

As diversas informações dos bancos de dados de um *Datawarehouse* podem servir para dinamizar e trazer muitas vantagens para qualquer empresa, agilizando seus negócios, diminuindo custos e otimizando processos decisórios devido à riqueza de informações que agrega.

Um *Datawarehouse* pode ser definido também como um super banco de dados que trabalha por traz de outros bancos de dados de ferramentas de *Content Management*, *e-Business*, *Business Inteligence*, *Workflow*, *CRM*, Gestão Empresarial etc, integrando informações que são de extrema relevância em processos decisórios e de estratégia dentro de uma empresa.

Data Mining

Os *Data Mining* (DM) são as ferramentas de busca de dados em um sistema amplo de banco de dados, de um *Datawarehouse*. De que adianta uma empresa agregar dados de todas suas sedes, seus departamentos e clientes se não souber buscar as informações necessárias para otimizar seus serviços e agilizar suas mudanças e inovações? *Data Mining* é a ferramenta, ou são as diversas ferramentas que buscam informações relevantes dentro de um *Datawarehouse*: estatísticas, números, relatórios e diversos dados necessários para aperfeiçoar diversos processos empresariais.

Características do Datawarehouse

Comparação entre ele e os bancos de dados operacionais:

Características	Bancos de dados Operacionais	Datawarehouse
Objetivo	Operações diárias do negócio	Analisar o negócio
Uso	Operacional	Informativo
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 – Terabytes	Terabytes – petabytes
Histórico	60 a 90 dias	5 a 10 anos
Granularidade	Detalhados	Detalhados e resumidos
Estrutura	Estática	Variável
Manutenção desejada	Mínima	Constante
Acesso a registros	Dezenas	Milhares
Atualização	Contínua (tempo real)	Periódica (em *batch*)
Número de índices	Poucos/simples	Muitos/complexos
Intenção dos índices	Localizar um registro	Aperfeiçoamento de consultas

- É um banco de dados que extrai informações de toda a empresa, dos setores de produção, vendas, recursos humanos etc;
- Seus dados são otimizados para extrair informações, e não para processamento de informações;

- Utiliza ferramentas de *mining* desenvolvidas para buscar as informações mais relevantes, mas que podem também acessar os dados primitivos de outros bancos de dados em caso de necessidade de se aprofundar em algum nível específico de informação, as ferramentas *Data Mining* devem ser maleáveis para permitir tais consultas;
- Um *Datawarehouse* contém informações de diversos outros bancos de dados relacionais, e também de arquivos e documentos diversos.

Características peculiares do Datawarehouse

- **Orientado por temas:** armazena informações dentro de temas específicos e de relevância para a empresa, tais como suas atividades, clientes etc;

- **Integrado:** o *Datawarehouse* uniformiza informações provenientes de diferentes base de dados;

- **Variante no tempo:** o *Datawarehouse* tem a características de não atualizar seus dados, e sim acumulá-los, e sem perder a sua referência de tempo;

- **Não volátil:** os dados carregados dentro de um *Datawarehouse* são imutáveis, ficam lá disponíveis para consulta, e não para processos transacionais;

- **Granulidade:** diz respeito ao nível de organização dos dados dentro de um *Datawarehouse*, compactando os dados e organizando-os por índices, se torna mais fácil o acesso aos dados relevantes e economiza-se espaço de armazenamento, obtendo-se assim uma baixa granulidade;

- **Particionamento dos dados:** é mais uma maneira de se organizar os dados dentro de um *Datawarehouse*. Dispõem-se os dados mais detalhados em unidades físicas menores, classificando-os por data, área de negócio ou área geográfica, unidades ou outros critérios.

Arquitetura do Datawarehouse

Para que um *Datawarehouse* seja eficiente, não basta agregar informações provenientes de vários sistemas, é preciso filtrar informações relevantes para um processo decisório e estratégico de uma empresa. Isso depende muito da arquitetura do *Datawarehouse* e das necessidades da empresa em questão, mas, de um modo geral, a arquitetura de um sistema *Datawarehouse* segue o padrão genérico como veremos a seguir.

Estrutura Genérica de um Datawarehouse[1]

Um *Datawarehouse* é composto por diversas camadas:

- **Banco de dados operacionais/ fontes externas**: são os banco de dados operacionais e fontes de arquivos externos que irão povoar os dados do *Datawarehouse;*
- **Acesso à informação**: *softwares* e *hardware* necessários para se extrair as informações do *Datawarehouse*, camada pela qual os usuários interagem por meio das ferramentas de *Data Mining*;
- **Acesso aos dados**: camada de ligação entre as ferramentas de *Data Mining* aos dados primitivos do *Datawarehouse*;
- **Metadados**: são os índices, dicionários de dados e outras ferramentas que classificam e descrevem os dados e onde estão armazenados;
- **Gerenciamento de processos**: camada que faz a atualização do *Datawarehouse*, que tem a função de coletar dados, classificá-los e resumi-los;
- **Transporte**: camada que gerencia o transporte de dados através da rede;

[1] Adriano Dal'alba.

- **Datawarehouse:** local físico e virtual onde estão os dados e informações que compõem do *Datawarehouse* propriamente dito. Essa camada pode ser, muitas vezes, composta somente de índices e dicionário de dados.

Expostas suas características e vantagens em comparação aos sistemas de bancos de dados relacionais e particionados, o *Datawarehouse* se apresenta como a melhor solução para agregar e organizar essas estruturas de dados para as empresas, especialmente as de grande porte e transnacionais. Contudo, para obter vantagem de negócio por meio de uma estrutura de *Datawarehouse*, é preciso lançar mão de ferramentas capazes de extrair inteligência dessas largas bases de dados integradas. Na ponta dessas ferramentas está o *Data Mining* a qual abordaremos a seguir.

Data Mining

Definição

DATA MINING (ou *Data Mining*) significa "garimpagem" automática de dados, recurso ou recursos que buscam informações armazenadas em compartimentos diversos de um sistema e, dependendo do caso, as cruzando, resultando na obtenção de relatórios, freqüentemente com o objetivo de apoiar decisões corporativas.

Qualquer sistema de *Datawarehouse* (DW) só funciona e pode ser utilizado plenamente, com boas ferramentas de exploração. Com o surgimento do *Datawarehouse*, a tecnologia de *Data Mining* (mineração de dados) também ganhou a atenção do mercado.

Como o *Datawarehouse* possui bases de dados bem organizadas e consolidadas, as ferramentas de *Data Mining* ganharam grande importância e utilidade. Essa técnica, orientada à mineração de dados, oferece uma poderosa alternativa para as empresas descobrirem novas oportunidades de negócio e, acima de tudo, traçarem novas estratégias para o futuro.

O propósito da análise de dados é descobrir previamente características dos dados, sejam relacionamentos, dependências ou tendências desconhecidas. Tais descobertas tornam-se parte da estrutura informacional em que decisões são formadas. Uma típica ferramenta de análise de dados ajuda os usuários finais na definição do problema, na seleção de dados e a iniciar uma apropriada análise para geração da informação, que ajudará a resolver demandas descobertas nessas análises. Em outras palavras, o usuário final reage a um estímulo externo, a descoberta do problema por ele mesmo. Se o usuário falhar na detecção do problema, nenhuma ação é tomada. A premissa do *Data Mining* é uma argumentação ativa, isto é, em vez do usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do *Data Mining* pesquisam automaticamente os

mesmos à procura de anomalias e possíveis relacionamentos, assim detectando problemas que não tinham sido identificados pelo usuário. Ou seja, as ferramentas de *Data Mining* analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos desses dados, então diagnosticam o comportamento de inúmeros negócios, requerendo a mínima intervenção do usuário, que poderá se dedicar somente à busca por conhecimento assim angariando vantagens competitivas.

Assim compreendido os recursos potenciais da tecnologia, entende-se que as ferramentas de *Data Mining* são baseadas em algoritmos que formam blocos de inteligência artificial, redes neurais, regras de indução e lógica de predicados, com isso facilitam e auxiliam o trabalho dos analistas de negócio das empresas, ajudando as mesmas a serem mais competitivas e maximizarem seus lucros.

A estatística

O *Data Mining* descende fundamentalmente de três linhagens. A mais antiga delas é a *Estatística Clássica*. Sem a estatística não seria possível operar o *Data Mining*, é a estatística que forma a base da maioria das tecnologias a partir das quais o *Data Mining* é construído.

A estatística clássica envolve conceitos como distribuição normal, variância, análise de regressão, desvio simples, análise de conjuntos, análises de discriminantes e intervalos de confiança, todos usados para estudar dados e os relacionamentos entre eles.

Esses são as pedras fundamentais nas quais as mais avançadas análises estatísticas se apóiam. E, sem dúvida, no coração das mais atuais ferramentas e técnicas de *Data Mining*, a análise estatística clássica desempenha um papel fundamental.

Inteligência Artificial

A segunda linhagem do *Data Mining* é a *Inteligência Artificial*, ou IA. Essa disciplina, que é construída a partir dos fundamentos da heurística, em oposto à estatística, tenta

imitar a maneira como o homem pensa na resolução dos problemas estatísticos.

Em função desse "approach", ela requer um impressionante poder de processamento, que era impraticável até os anos 80, quando os computadores começaram a oferecer um bom poder de processamento a preços mais acessíveis.

No início, a IA desenvolveu algumas aplicações restritas para o alto escalão do governo norte-americano e de cientistas, mas os altos preços não permitiram que ficasse ao alcance de todos. As notáveis exceções foram certamente alguns conceitos de IA adotados por alguns produtos de ponta, como módulos de otimização de consultas para SGBD – Sistema de Gerenciamento de Banco de Dados.

Machine Learning

A terceira e última linhagem do *Data Mining* é a chamada *Machine Learning*, que pode ser mais bem descrita como o casamento entre a Estatística Clássica e a Inteligência Artificial. Enquanto a IA ainda não era um sucesso comercial, suas técnicas foram sendo largamente cooptadas pela linhagem *Machine Learning*, que se valeu das sempre crescentes taxas de preço/performance oferecidas pelos computadores nos anos 80 e 90, para conseguir mais e mais aplicações devido às suas combinações entre heurística e análise estatística. A *Machine Learning* tenta fazer com que os programas de computador "aprendam" com os dados que lêem e analisam, tal que esses programas tomem decisões diferentes baseadas nas características dos dados estudados, usando a estatística para os conceitos fundamentais, adicionando mais heurística avançada e algoritmos de IA para alcançar os seus objetivos. De muitas formas, o *Data Mining* é fundamentalmente a adaptação das técnicas da *Machine Learning* voltada para aplicações de negócios. Desse modo, podemos descrevê-lo como a união dos históricos e dos recentes desenvolvimentos em estatística, em IA e *Machine Learning*. Essas técnicas são usadas juntas para estudar dados relacionais e achar tendências ou padrões antes não contabilizados. Hoje, o *Data Mining* tem experimentado uma crescente aceitação nas ciências e nos negócios que precisam analisar grandes volumes de dados

e achar tendências que seriam impossíveis de serem identificadas de outra forma. Para atender a novas necessidades, as ferramentas de SAD – Sistema de Apoio a Decisão – têm sido incrementadas com sofisticadas funções, tais como a análise OLAP (*On Line Analitical Processing*), formatações de relatórios cada vez mais flexíveis, visualização tridimensional, filtros, classificações e alertas entre outros.

De todas essas funções, a OLAP é, sem discussão, a mais sofisticada, na medida em que possibilita aos usuários estudar os dados de maneira multidimensional, os mesmos podem "perfurar" os dados em detalhes (função comumente chamada de "drill down"), ou ainda ver porções sumarizadas desses dados (função "slice-and-dice") do ponto de vista que desejarem, enquanto "perseguem" as respostas que procuram. Assim, essa função permite que o usuário enxergue dados de diferentes perspectivas em numerosos níveis de detalhe ou agregação. Comparativamente, o *Data Mining* apresenta um método alternativo e automático de descobrir padrões nos dados. Alternativo por que trabalha diretamente com todos os dados de um grupo ao invés de se ater a determinado caminho e "perfurar" (ou executar um *drill down*) na busca por maiores detalhes. E é automático em sua execução devido ao fato de a ferramenta estudar os dados e apresentar seus "achados" aos usuários, apesar deste ter que tomar o cuidado de fornecer dados úteis à ferramenta para que ela "triture" o grupo de dados a seu modo.

Devido essa dupla característica, o *Data Mining* é extremamente adequado para analisar grupos de dados que seriam difíceis ou dispendiosos apenas com funções OLAP, visto que esses grupos são grandes demais para serem "navegados" ou explorados manualmente, ou por que contêm dados muito densos e/ou não-intuitivos para serem compreendidos.

O *Data Mining* não requer que o usuário "pilote" a ferramenta durante o processo de análise dos dados, ou que administre o processo ao longo de seu andamento. Fornecidos dados adequados no início do processo, o *Data Mining* traz sentido aos grupos de dados ao relacionar tendências ou padrões "escondidos" nos mesmos, e apresentá-los ao usuário em um formato compreensível.

Dessa forma, a diferença básica entre OLAP e *Data Mining* está na maneira como a exploração dos dados é realizada. Na análise OLAP a exploração é feita através da verificação, isto é, o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta pra refutá-la ou confirmá-la. Com o *Data Mining*, a questão é total ou parcialmente desconhecida e a ferramenta é útil para a busca de conhecimento.

Essa capacidade claramente agrega valor às soluções de apoio à decisão. Em muitos casos, os resultados apresentados pelo *Data Mining* fazem surgir interessantes questões sobre os dados originais ou que tenham alguma relação com os mesmos no *Datawarehouse*, ou seja, é o melhor exemplo de como o *Data Mining* agrega valor aos SAD.

Quando os resultados do *Data Mining* propõem questões adicionais, os usuários podem procurar por mais respostas em tempo real simplesmente rodando uma nova consulta na base de dados e minerando de novo. Ou usar os resultados de uma mineração como guia para pesquisar mais dados por meio da análise OLAP. Usando um processo interativo de consultas, *Data Mining* e OLAP, os usuários conseguem obter as informações que mais lhes interessam, ao mesmo tempo em que podem formatar os relatórios da maneira mais conveniente.

Dessa forma, o *Data Mining* está se tornando um componente essencial para análise em sistemas de apoio à decisão, complementando as funções já existentes.

Algumas barreiras ao uso do Data Mining

Nem sempre o *Data Mining* pode agregar valor aos SAD. De fato, no passado existiam (e ainda existem, de certa forma) muitas barreiras para o *Data Mining* se tornar uma função essencial dos SAD. As mais importantes têm sido ultrapassadas, outras ainda se mantêm.

Fundamentalmente, as mais importantes foram: alto custo das soluções; necessidade de grandes volumes de dados armazenados em poderosos servidores; e a pouca facilidade de uso das ferramentas de *Data Mining* para pessoas que não fossem altamente especializadas.

Outras que se podem citar seriam: o desafio de preparar os dados para mineração; as dificuldades em se obter uma

análise de custo/benefício bem fundamentada antes de iniciar o projeto e a preocupação quanto à viabilidade de muitos dos fornecedores dessas ferramentas.

Altos Custos

O alto custo da maioria das ferramentas dificulta a disseminação das mesmas entre as corporações. Uma simples operação matemática mostra que um projeto, com o custo de US$ 20.000 por usuário, pode atender não mais que um seleto grupo de usuários.

Certos fornecedores têm introduzido produtos com custo mais baixo, mesmo assim, o preço continua limitando para uso em larga escala. Evidentemente, os custos por usuário precisam ser reduzidos antes que os benefícios desta tecnologia possam atingir uma massa maior de usuários.

Necessidade de grandes volumes de dados

O maior obstáculo ao *Data Mining* no passado foi a necessidade de armazenar e administrar montanhas de dados requerendo amplos servidores. Isso por si só já dificultava bastante o crescimento do mercado de *Data Mining*. No entanto, a maioria dos fornecedores dessa tecnologia continua insistindo no discurso de que o *Data Mining* requer terabytes ou petabytes de dados e poderosos servidores, mas soluções mais acessíveis já têm aparecido no mercado e criado condições para a tecnologia deslanchar. É unânime no mercado a afirmação de que 80% do valor de um determinado grupo de dados pode ser encontrado em 20% dos dados que o compõem, então é claro que os fornecedores vão fazer de tudo para achar esses 20% e minerá-los.

As ferramentas que procuram tornar pessoalmente manejáveis grupos de dados dão aos usuários a possibilidade de minerar porções de dados em seu próprio computador, o que permite, efetivamente, contornar essa barreira. Embora não tenham a intenção de substituir aplicações que necessitam de grandes volumes de dados, essas ferramentas podem prover um

acesso alternativo e serem usadas em conjunto com as ferramentas tidas como "pesadas".

Complexidade das Ferramentas

Mesmo com essa nova geração de ferramentas que permitem a mineração em moderados grupos de dados, uma terceira barreira ainda permanece: a grande maioria das ferramentas continua incompreensível para os usuários comuns. De fato, muitas ferramentas ainda fazem o seu trabalho em uma "caixa-preta", não permitindo que se saiba como alcançaram os seus resultados.

Isso significa que o *Data Mining* serve ao contexto da área de sistemas, útil para usuários que têm de submeter solicitações, esperar por dias ou semanas enquanto um *expert* processa os dados, para então receberem e examinarem a saída sumarizada. Se os resultados não satisfizerem, todo processo tem que ser recomeçado.

Felizmente, há técnicas de *Data Mining* mais compreensíveis, como, por exemplo, as árvores de decisão, que permitem a qualquer um com conhecimentos básicos de computação utilizar e compreender o processo.

Contudo, o poder dessas ferramentas não chega nem perto daquelas que utilizam técnicas mais sofisticadas.

O desafio da preparação dos dados para a mineração

A preparação dos dados para se realizar a mineração envolve muitas e trabalhosas tarefas em um projeto de *Data Mining*, sendo considerada como 80% do trabalho.

Os dados devem ser relevantes para as necessidades dos usuários, "limpos" (livres de erros lógicos ou de entrada de dados), consistentes e sem excessivas nulidades.

Mesmo que haja um projeto de Datawarehouse anterior, no qual os dados são normalmente limpos e centralizados em um único local, continua havendo a necessidade de prepará-los para a mineração, assim como é crítica a escolha dos dados certos para se minerar.

As dificuldades de se realizar a análise custo/benefício do projeto de Data Mining

Como o objetivo da tecnologia de *Data Mining* é descobrir tendências (em dados) que não seriam visíveis de outra maneira, torna-se virtualmente impossível estimar a taxa de retorno a partir de algo que é desconhecido, assim, igualmente é difícil estimar o investimento necessário de um projeto de *Data Mining*. Visto que normalmente um projeto de *Data Mining* é razoavelmente caro, pode ser um tanto arriscado se decidir por em favor de sua implementação.

Viabilidade dos fornecedores de ferramentas de Data Mining

Finalmente, a viabilidade de mercado da maioria das ferramentas é uma preocupação das empresas que procuram uma ferramenta confiável para uso em curto, médio ou longo prazo. O mercado está abarrotado de fornecedores, desde pequenas empresas que comercializam sistemas até grandes companhias em que as ferramentas dos clientes são apenas mais uma das inúmeras que produz, não fornecendo uma atenção exclusiva para todos. Assim como qualquer nova tecnologia, a escolha do fornecedor é tão importante quanto à escolha da ferramenta.

Implementação de projetos de Data Mining: Algumas questões importantes

A regra 80/20

Os analistas concordam que, por dentro de qualquer grande grupo de dados, 80% da informação podem ser encontrada por meio de 20% dos dados que aglutina. Essa regra força a redução do tamanho do grupo de dados a ser analisado. O particionamento dos grupos de dados pode ser usado para conseguir esse intento, o que significa que apenas dados relevantes são usados para fins de atividade *Data Mining*, concentrando os dados úteis dentro dos 20% selecionados para análise. Mesmo se os grupos de dados particionados

continuarem grandes, pode-se retirar amostras para se ter uma idéia do que acontece no todo.

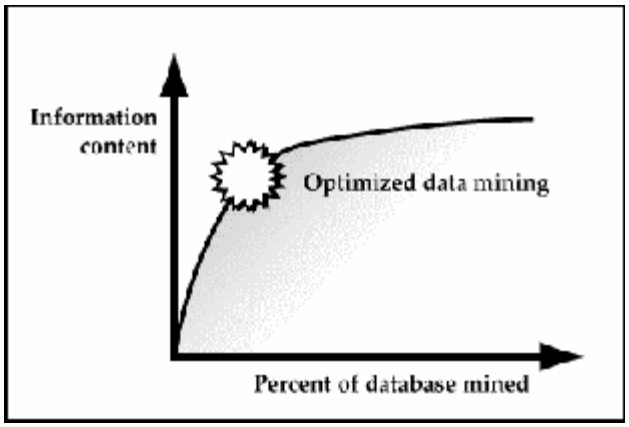

Gráfico da relação 80/20

Viabilidade estatística mínima

Quando se tenta minimizar o tamanho dos grupos de dados, é importante ter em mente que um número mínimo de registros é necessário para se ter viabilidade estatística. Em geral, um mínimo de 200 registros pode ser analisado de maneira a se obter resultados estatisticamente viáveis, pois é um bom tamanho dentro do escopo de dados de negócio.

As necessidades específicas de negócio

No contexto dos negócios, o *Data Mining* encontra aplicações adequadas para estudar aspectos específicos do negócio, o que se chama *Escopo de Análise*. Por exemplo, um gerente de uma agência bancária quase certamente está mais preocupado com os seus clientes da casa do que com os clientes estaduais ou nacionais do seu banco. Então, o escopo de análise deste gerente certamente é a sua agência, ou possivelmente a sua região, de modo a poder comparar os dados da sua agência com os das agências de sua região. Grandes quantidades de dados, neste caso, não são necessários para se chegar a bons resultados.

Porque dados de negócio são mais fáceis de minerar

Os dados de negócio apresentam oportunidades únicas para minerar. Comparados com outros tipos de dados como os científicos, atuariais[2] ou estatísticos, eles são mais homogêneos e intuitivos, além de proporcionar mais facilidades de agregação, que podem muitas vezes reduzir o volume de dados "crus" necessários para uma dada operação. Além disso, os dados de negócio normalmente são mantidos por pessoas de negócio, que conhecem o seu significado. Os outros dados são mais freqüentemente recolhidos por um processo remoto, e aí transferidos para outros analistas para processamento posterior, reduzindo as possibilidades de esses últimos entenderem o significado de cada aspecto dos dados.

Os dados de negócios são previsíveis

A possibilidade de se fazer previsões é o fator chave para que os dados de negócio sejam mais "mineráveis". Eles são coletados dentro do sistema de um negócio particular, descrevendo, por exemplo, os clientes daquele negócio. Dados de negócio "limpos" tenderão a conter valores que se incluirão em certas escalas cabíveis. Ë completamente diferente, por exemplo, um vendedor de carros vender um a R$ 40.000,00 ou a R$40.000.000,00. Os preços dos carros tendem a se enquadrar dentro de uma faixa razoável. Igualmente, não é razoável o mesmo vendedor de carros vende-los a pessoas que residem em países distantes ou que pagam com moedas estrangeiras. Os dados de negócio sobre vendas de carros tendem a descrever vendas a consumidores locais, pagando na moeda do país local. Devido a terem menos exceções, é mais fácil ao usuário do *Data Mining* reconhecer tendências ou padrões. Valores fora do escopo de um grupo de dados fazem com que seja mais difícil minerar. Usando os termos de *Data Mining*, esses valores excepcionais são chamados de "ruído".

[2] Atuarial – termo que designa o profissional técnico especialista em mensurar e administrar riscos.

Os Dados de Negócio são intuitivos

A natureza intuitiva dos dados de negócio é outro facilitador para o *Data Mining*. Enquanto o dado científico normalmente contém valores "impenetráveis" e muitas minúcias, o dado de negócio se encontra do lado diametralmente oposto.

Dados de negócio descrevem negócios, são nomeados e guardados pelas pessoas de negócio. Termos como receita, despesa, taxa de resposta, nível do estoque e limites de crédito são intuitivos, e os dados armazenados para estes termos fazem sentido em um contexto de negócios. As pessoas de negócio saberão intuitivamente que os valores para limite de crédito serão certamente quantias de dinheiro, e que um valor para taxa de resposta será em percentual. O fato dos dados de negócio serem intuitivamente compreensíveis às pessoas de negócio é uma grande vantagem para realizar *Data Mining*, pois consegue levantar muitos novos conhecimentos a partir de pequenos ou médios grupos de informações. Na linguagem do *Data Mining*, essa característica de dados intuitivos é chamada de "percepção nativa dos dados".

As agregações podem acelerar o Data Mining

Os dados de negócio são freqüentemente armazenados em formatos agregados como, por exemplo, receitas por trimestre, vendas por região ou respostas de promoções por CEP. Esses formatos podem ser muito mais fáceis de minerar do que dados crus. Por outro lado, dados que não são de negócio suportam muito menos sumarizações desse modo. Essas agregações permitem uma mineração proveitosa em um grupo de dados muito menor do que seria possível com dados científicos. Minerando agregações, os usuários de negócio podem descobrir tendências em seus negócios em qualquer nível que eles desejem. Minerando em receitas regionalmente agregadas por gastos com propaganda nos vários meios de comunicação poderia descrever que tipo de propaganda é mais efetiva em cada região. Neste caso, não haveria a necessidade de minerar as vendas em todo o território nacional.

Os usuários dos dados são os donos dos dados

Por fim, a última grande vantagem importante dos dados de negócio é que estes são, quase sempre, coletados, mantidos e apropriados pelas mesmas pessoas que os usam, isto é, pelas pessoas que conduzem o negócio. Em contraste, os dados científicos são freqüentemente compostos de fontes de dados estratificadas, como, por exemplo, amostras que são coletadas em pesquisas de campo por algum agente e que são enviadas ao centro de operações para serem compiladas por um time de analistas.

Considerações Finais

Por todos esses motivos expostos e descritos, o *Data Mining*, inserido em uma bem estruturada base de *Datawarehouse*, é uma ferramenta que tende a se tornar vital para as empresas em um cenário globalizado de negócios como se desenha a atualidade capitalista neoliberal.

Assim, pode-se afirmar que DW e DM são a ponta de organização, estratificação, análise e inteligência de dados que se tornará indispensável para as empresas gerirem seus negócios e competirem no mercado mundial.

Referências

ANDREATTO, Ricardo. *Construindo um Datawarehouse e Analisando suas Informações com Data Mining e Olap*. www.datawarehouses.hpg.com.br

DAL'ALBA, Adriano. *Um estudo sobre Datawarehouse*. http://www.geocities.com/siliconvalley/port/5072/

GATES, Bill. *A Empresa na Velocidade do Pensamento*. São Paulo: Cia das Letras, 1999.

Sites

www.ibm.com.br
www.datawarehouse.inf.br
www.b2bmagazine.com.br