Mineração de Dados usando o software WizRule em Base de Dados de Compras de TI

Denise Chaves Carvalho Barbosa, Maria Augusta Machado

IBMEC/RJ

RESUMO

Esta pesquisa tem como objetivo validar a hipótese de que a Mineração de Dados pode ser aplicada em base de dados de compra, gerando a descoberta do conhecimento oculto, como uma grande contribuição ao processo decisório da gestão de compras. Para melhor compreensão desse trabalho abordamos, na Revisão de Literatura, primeiramente, um histórico sobre o processo decisório, bem como a evolução dos estudos deste tema e da relação entre a Tomada de Decisão e os Sistemas de Informação. Posteriormente, ainda na Revisão da Literatura, nos voltamos para a abordagem sobre as ferramentas, objeto deste estudo, *Data Warehouse* e *Data mining*, passando pelo processo KDD, por constituírem uma recente geração de Sistemas de Apoio à Decisão. O *Data Warehouse* por tratar-se de um banco de dados apropriado para objetivos gerenciais e o *Data Mining* por permitir a análise dos dados armazenados para a descoberta das relações ocultas, revelando informações valiosas sobre as compras já efetuadas. Por fim, demonstraremos, no último capítulo deste trabalho, a aplicação prática do *Data Mining* em base de dados de compras de TI de uma empresa de grande porte, incluindo a análise dos resultados gerados e comprovando ser de grande utilidade o uso dessas ferramentas na obtenção de informação útil sustentando o processo decisório e a estratégia de negócios na área de compras de produtos.
Palavras Chave: Mineração de Dados, Compras, Base de Dados.

ABSTRACT

This research has as objective validates the hypothesis that the Data Mining can be applied in purchase database, generating the discovery of the occult knowledge, as a contribution to the decisive process of the administration of purchases. For better understanding of that work we approached, in the Revision of Literature, firstly, a report on the decisive process, as well as the evolution of the studies of this theme and of the relationship between the Decision Making and the Information Systems. Later, still in the Revision of the Literature, we went back to the approach on the tools, object of this study, Date Warehouse and Date Mining, going by the process KDD, for they constitute a recent generation of Decision Support Systems. The Date Warehouse for treating of an appropriate database for managerial objectives and Date Mining for allowing the analysis of the data stored for the discovery of the occult relationships, revealing valuable information on the purchases already made. Finally, we will demonstrate, in the last chapter of this work, Data Mining practical application in IT purchase database of a big company, including the analysis of the generated results and proving to be of great usefulness the use of those tools in the obtaining of useful information sustaining the decisive process and the strategy of businesses in the area of purchases of products.
Key Words: Date Mining, Purchases, Dates Warehouse.

1 Introdução

Os avanços obtidos nas áreas de software e hardware possibilitaram a criação de aplicações comerciais e científicas capazes de processar grandes volumes de dados. Por exemplo, o sistema que é usado por uma grande empresa do setor petroquímico para processar as compras de suprimentos diversos processa milhões de transações diariamente, produzindo um volume de dados que pode chegar a mais de uma dezena de Gigabytes.

De acordo Send e Jacob (1998), cada vez mais, as empresas vêm fazendo grandes investimentos em aplicativos e equipamentos usados para o armazenamento, integração, análise e gerenciamento dos seus dados. Isto se deve a uma mudança de filosofia, pois, atualmente, as bases de dados não são mais consideradas simples repositórios de informações, mas sim, um importante patrimônio da organização.

Os dados gerados pelas organizações de médio e grande porte superam a capacidade humana

de interpretar, analisar e compreender tanta informação. Por isso, são necessárias novas ferramentas e técnicas capazes de analisar automaticamente o volume de dados produzidos, fornecendo o conhecimento necessário para auxiliar nos processos decisórios.

A área conhecida por Extração de Conhecimento de Base de Dados ou *Knowledge Discovery in Databases* (KDD) surgiu para auxiliar a análise de grande volume de dados. Os trabalhos neste segmento objetivam o estudo da aplicação de novas metodologias, ferramentas e técnicas capazes de extrair conhecimento contido em grandes volumes de dados.
O processo de KDD é feito a partir dos conceitos de Bases de Dados, ferramentas de visualização, métodos estatísticos e técnicas de Inteligência Artificial (IA).

1.1 Cenário

Para a demonstração prática da utilização do *Data Mining,* fomos buscar os dados para a aplicação da técnica em uma empresa multinacional de grande porte, com sede no Brasil, com o sistema ERP (*Enterprise Resource Planning*) do fornecedor SAP em produção há aproximadamente dois anos, bem como o *software* de *Data Warehouse*, SAP *Business Warehouse* - BW, cujo ambiente específico, separado do ambiente transacional[*1], armazena informações que são estruturadas para facilitar a consulta e análise, suportando assim o processo decisório e a gestão da empresa conforme demonstra a figura 1 abaixo:

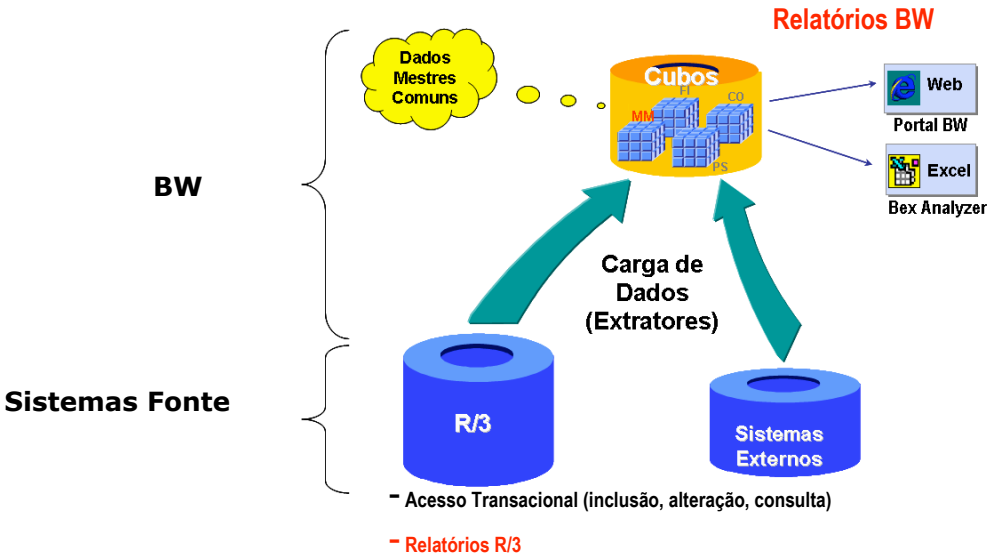

Figura 1: SAP BW --→ Ambiente Integrado

Nesta empresa a estrutura de compras de produtos encontra-se centralizada para valores superiores aos estabelecidos periodicamente pelos órgãos internos competentes e descentralizada para os valores inferiores aos anteriormente citados.

O estudo foi realizado na estrutura descentralizada de compras de pequenos itens de TI de baixo valor (*pendrives*, impressoras, *scanners*, *mouses* etc).

A estrutura de compras, objeto do estudo, é composta por cinco compradores que realizam para cada pedido de compra o mínimo de 03 (três) cotações junto a 77 (setenta e sete) fornecedores.

O período analisado está compreendido entre outubro de 2004 a março de 2006.

1.2 Objetivo

Buscando contribuir com uma solução de otimização de tarefas capazes de auxiliar a tomada de decisão na gestão de compras de suprimentos de TI em uma empresa de grande porte, o objetivo geral deste trabalho é a aplicação da técnica de *Data Mining,* posteriormente ao processo de KDD, no auxílio a tomada de decisão. Será realizado, como objetivo secundário, um estudo referente aos principais conceitos das áreas de Processo Decisório, Armazenamento Dados e Extração de Conhecimento de Base de Dados para, finalmente, chegarmos ao nosso objetivo específico que é a aplicação das técnicas estudadas em base de dados de compras de itens de TI de uma grande empresa utilizando o *software* WizRule verificando a possibilidade do uso efetivo desta técnica a fim de contribuir para a tomada de decisão na área de Gestão de Materiais.

1.3 Relevância

Há alguns anos, o uso da tecnologia de informática vem sofrendo várias mudanças no que tange ao

acesso e à análise de dados. Estas transformações estão moldando um novo paradigma baseado no armazenamento, tratamento e análise de imenso volume de dados. Conseqüentemente, as grandes empresas estão começando a explorar as possibilidades oferecidas pelas diversas técnicas e ferramentas atualmente disponíveis para aprimorar o processo de tomada de decisão.

As implicações destas mudanças para o mundo dos negócios são enormes. Assim, este novo paradigma envolve o uso de sistemas especificamente projetados para o tratamento dos dados e geração de conhecimento de forma flexível em tempo hábil para análises pelos gestores das empresas.

Aplicar as técnicas de KDD na tentativa de se encontrar conhecimento nesta nova realidade nas diversas áreas da organização é de interesse das grandes empresas. É bastante comum encontrarmos a aplicação dessas técnicas nas áreas financeiras, vendas e, principalmente, de marketing. Como a área de compras de uma organização também possui uma imensa massa de dados, acreditamos ser de grande relevância um estudo sobre a aplicação das técnicas de Data Mining para a geração e análise de informações valiosas em tomadas de decisões específicas dessa área.

1.4 Limitações

Este estudo estará limitado à tarefa de descobrir afinidades entre os dados a serem analisados através da técnica de *Data Mining* denominada "regras de associação". Do resultado deste trabalho serão obtidas diversas conclusões, aumentando-se o conhecimento extraído das bases de dados.

Outras limitações dizem respeito à restrições impostas por código de ética e normas da empresa pesquisada no uso da base de dados de compras em sua totalidade por questões estratégicas de segurança da informação, o que nos levou a utilizar o *WizRule Demo 4.05* que, por ser uma versão demo, possui uma limitação de uso relacionada ao número máximo de linhas (1 000) que podem ser analisadas.

2 Revisão de Literatura

Para uma melhor compreensão deste estudo, abordamos neste capítulo, primeiramente, um histórico sobre o processo decisório, bem como a evolução dos estudos deste tema e da relação entre a Tomada de Decisão e os Sistemas de Informação. Posteriormente, nos voltamos para uma abordagem sobre as ferramentas, objeto deste estudo, *Data Warehouse* e *Data Mining*, passando pelo processo KDD, por constituírem uma recente geração de Sistemas de Apoio à Decisão.

2.1 Processo Decisório

2.1.1 Histórico

Por tratar-se de tema longo e muito abrangente, o surgimento do processo decisório e sua evolução serão apresentados de forma resumida com base em BISPO e CAZARINI (1998) destacando os mais importantes pontos de sua evolução.

Segundo Bispo e Cazarini (1998), o homem sempre procurou alguma ajuda para seu processo decisório desde o começo da civilização. Considerava-se que pessoas com "místicos poderes" teriam livre e direto contato com os seres considerados divinos e que todas as orientações dadas por essas pessoas eram, também, divinas. Dessa forma, as decisões eram consideradas sábias e se, no entanto, os resultados não fossem os esperados, tais erros significavam que as divindades estariam insatisfeitas. Nesta época, as entidades divinas e as pessoas que as representavam tinham forte influência nas decisões.

Depois outras divindades mais populares, como Maomé, Buda e Cristo, sugiram, bem como, líderes religiosos passando esses novos personagens a influenciar diretamente nas decisões pessoais através de preceitos religiosos e da mesma forma, como no passado, resultados insatisfatórios significavam falta de fé dos decisores. Desta forma, continua a grande influência das divindades nas tomadas de decisão.

No presente, vários outros fatores influenciam o processo decisório. No passado, esses outros fatores também existiam, mas somente num passado recente foram ganhando importância significativa.

Segundo Pereira e Fonseca (1997), no início do século XX, os critérios usados para tomada de decisão se concentravam no executivo maior, que muitas vezes além de ser o dono do negócio, possuía o privilégio da escolha que acreditasse ser a melhor para a empresa e seus trabalhadores. Isso devido o entendimento existente na época de que os trabalhadores eram pessoas sem capacidade e não estavam preparados para tomarem decisões, sendo avaliados por sua produção e descartados quando não produziam o esperado pela empresa. Acreditava-se que apenas os executivos de alto escalão tivessem capacidade para sábias decisões devido ao amplo conhecimento a eles atribuído sobre todas as alternativas possíveis e suas conseqüências.

Somente no início dos anos 60, essa perspectiva mudou com o surgimento do movimento conhecido como Escola de Relações Humanas, oriundo da contribuição da Psicologia Social à Teoria da Administração. A partir desse movimento os trabalhadores passam a ser reconhecidos como alguém capaz de pensar, de decidir e de ser motivado (PEREIRA e FONSECA, 1997), ou seja, não mais se restringia ao alto escalão a capacidade de decidir.

2.1.2 Evolução dos Estudos

Segundo Simon (1986), o estudo do processo decisório, principalmente após a Segunda Guerra, ganhou muita força, mas especificamente centrada no modelo racional, seguindo uma teoria prescritiva.

Na Teoria da Administração, o processo decisório foi esquecido até por volta da metade do século passado devido a ciência administrativa ter nascido tendo como base um conjunto de valores funcionais e mecanicistas, e as organizações foram concebidas somente como instrumentos técnicos, tendo como objetivo principal a maximização dos lucros e dos resultados.

Assim sendo, não havia dificuldade em perceber que os fatores determinantes das escolhas ou dos critérios de avaliação das opções se baseassem apenas na relação custo-benefício.

O mais relevante, no entanto, nesse fato era que se acreditava que a melhoria na relação custo-benefício e a maximização dos resultados aconteceriam de forma natural e que a decisão tomada seria a melhor diante dos instrumentos apresentados pela Teoria Administrativa. Esse era o motivo real de não haver grandes preocupações com o processo decisório. Supõe-se que essa negligência não era uma opção totalmente consciente em relação a não se valorizar o estudo da tomada de decisão, mas era uma conseqüência natural da maneira como fora construída a teoria até o momento.

Os modelos de tomada de decisão e sua classificação surgem pela divisão do estudo e/ou abordagem do processo decisório pelas diferentes escolas de Administração. A teoria da decisão, hoje, assume um privilegiado lugar no pensamento administrativo, contemplando os níveis estratégico, tático e operacional. A partir de Simon, a teoria da decisão vem conquistando sua relevância e sua especificidade, deixando, ao longo do tempo, a abordagem simplesmente quantitativa e adaptando-se a nova realidade decorrente das complexas mudanças pelas quais vêm passando, nas últimas décadas, as organizações.

Hall (2004) afirma que o processo decisório está envolvido de pressões imediatas sobre o tomador das decisões, a análise do tipo do problema e de suas dimensões básicas, da busca de soluções variadas e do exame minucioso de suas conseqüências, inclusive a antecipação dos diversos tipos de conflito pós-decisório e a escolha final.

Segundo Simon (1979), as organizações encontram-se envoltas em decisões, onde o processo decisório abrange ações inconscientes ou conscientes que são inerentes a um planejado sistema de esforço cooperativo.

Segundo esta ótica, o processo decisório pode ser feito de acordo com a percepção das situações, onde cada elemento da estrutura organizacional desempenha um papel definido com deveres e atividades a executar.

A organização pode então ser entendida como um sistema de decisões, onde cada elemento atua, escolhendo e decidindo entre as alternativas mais ou menos racionais, de acordo com sua motivação e personalidade. As decisões, portanto, são um processo de análises e escolhas, entre as várias alternativas apresentadas, durante a ação, que o indivíduo deverá seguir. É primordial ressaltar que o gestor, é quem decide sobre uma situação, onde possui opções e arbítrio, escolhendo a melhor opção entre estas.

O trabalhador que participa do processo decisório geralmente define a situação envolto num complexo de processos cognitivos e afetivos. Destaca-se o fato das organizações plurais buscarem a evolução do processo decisório baseadas na melhor direção do fluxo de informações gerenciais, onde é priorizada a qualidade dos processos sobre a dos produtos.

Procura-se, através de uma curta retrospectiva histórica, apresentar com a sistematização das características principais das diferentes etapas ou fases da evolução do processo decisório e a forma como foi discutido ao longo dos anos, alguns destaques pinçados pelas escolas e pelos ideólogos em busca do aperfeiçoamento da gestão empresarial.

Estes destaques podem ser identificados em alguns dos principais estudiosos que se destacaram em abordagens gerenciais, tais como: Friedrich Wislow Taylor, Peter Drucker, Earnest Archer, Joseph Newman e Herbert Simon, que, certamente, encontram-se entre os precursores dessa sistêmica abordagem.

O entendimento de Archer (1980), adaptado a seguir, vislumbra oportunidades que nos conduzem a maiores entendimentos sobre as abordagens de Herbert Simon, em que é desejável que a forma prática do envolvimento cognitivo do ser humano ultrapasse os desvios que podem estar escondidos em maior número de etapas na tomada de decisão, como também na reprodução de várias complexidades da subjetividade humana, nem sempre necessárias ao sucesso do processo decisório.

Fases	Método Científico	Earnest Archer	Peter Drucker	Herbert Simon	Joseph Newman	Abordagem Sistêmica

1	Observação	Monitoração do ambiente decisório		Inteligência (busca de condições que pedem por solução)		Escolha do problema
2	Formulação do problema	Conceituação de problemas ou situações	Definição de problema		Reconhecimento do sistema que requer ação de decisão	Definição e quantificação do problema
3	Estabelecimento dos objetivos	Objetivos de decisão	Definição de expectativas			
4	Determinação das causas	Diagnóstico do problema ou situação				Determinação de relações causais entre os fatos para soluções
5	Formulação de hipóteses	Desenvolvimento de Soluções alternativas	Desenvolvimento de soluções alternativas	Invenção, desenvolvimento e análise de curso de ação	Identificação e desenvolvimento de caminhos alternativos de ação	Determinação de tentativas opcionais de solução
6	Metodologia	Definição de metodologia ou critério para avaliar alternativas				
7	Teste de hipóteses	Avaliação das soluções alternativas			Avaliação de alternativas	Teste das possíveis soluções
8	Formulação de conclusões	Escolha da melhor alternativa		Seleção de um caminho de ação	Escolha de uma das alternativas	
9	Comunicação de Resultados	Implementação da melhor alternativa	Saber o que fazer com a decisão	Implementação do caminho de ação selecionado	Implementação do caminho de ação selecionado	Documentação dos procedimentos

Tabela 1: Evolução do processo decisório - Adapt de: Earnest R. Archer, How to Make a Business Decision: An Analysis of Theory and Practice, Management Review, AMACOM, vol. 69, no. 2, fev. 1980, p.54-61.

2.1.3 A tomada de decisão e os sistemas de apoio

A sobrevivência das empresas e a situação das pessoas que direta ou indiretamente estão a ela ligadas, empregados, fornecedores, clientes ou acionistas, são afetadas diretamente pelas decisões gerenciais. Assim, o tomador de decisões é atingido por vários fatores de influência, inclusive por cobranças das pessoas atingidas para obtenção de um resultado de sucesso. Cada uma dessas pessoas solicita soluções diferenciadas e, possivelmente antagônicas, como solução de um problema, e é preciso que prioridades sejam estabelecidas quando estamos diante de posições e objetivos diferentes, antagônicos ou disputas de informações e recursos. É preciso transformar os objetivos da organização em objetivos gerais para todos os membros da empresa, buscando o compartilhamento da participação e da visão do futuro, buscando a satisfação dos usuários e clientes, não se descuidando, no entanto, dos demais grupos de interesses - acionistas e empregados.

Segundo Pereira e Fonseca (1997), no dia a dia, a viabilização desse processo envolto em conflitos de interesses, exige liderança, habilidade de negociação permanente, objetivos compartilhados e comunicação efetiva.

Nesse contexto, segundo Gates (1997), a informação é desejada e existe quem esteja disposto a pagar por ela, não é mensurável em tangível, porém é um valioso produto no mundo moderno porque proporciona poder.

É pela informação que se tem a possibilidade de suportar melhor o processo decisório, sendo função das diversas ferramentas que darão esse suporte ao processo, obter as informações necessárias de forma confiável, rápida e mostrá-las de maneira compreensível.

Segundo Power (2002), a conceituação de suporte computacional à decisão surge com o desenvolvimento de duas vertentes de pesquisa: estudos teóricos sobre Processo de Tomada de Decisão Organizacional, feitos durantes as décadas de 50 e 60 no *Carnegie Institute of Technology* e os trabalhos feito com Sistemas Computacionais Interativos, durante a década de 60 no *Massachuttes Institute of Technology*.

Segundo Fisher (2006), Costa (1997) e Pearson e Shim (1995), os pioneiros SADs - Sistemas de Apoio à Decisão - surgiram em 60 e

70 para suporte aos tomadores de decisão em soluções de problemas gerenciais que não fossem estruturados.
Durante esse período, os sistemas de computador que davam suporte à decisão eram desenvolvidos, primeiramente, para auxílio na resolução de problemas gerenciais específicos e, depois, aperfeiçoados a fim de incorporar outros problemas gerenciais. No entanto, não houve possibilidade de que, com um sistema desses, se chegasse a um bom suporte ao processo de tomada de decisão por tratar-se de um processo dinâmico, onde o fornecimento das informações tem que ocorrer no momento certo.

Apenas nos anos 80, com o crescimento na utilização dos Sistemas de Gerenciamento de Banco de Dados - SGDB - é que foi possível um acesso melhor aos dados disponíveis, à formatação desses dados e a construção de consultas e relatórios de maneira mais rápida, prática e barata. No entanto, quando se fazia necessária uma análise mais profunda dos dados essas eram feitas fora de um sistema computacional, ou seja, ainda faltava o desenvolvimento de uma ferramenta que auxiliasse realmente os tomadores de decisão.

Apesar dos avanços obtidos, o grande problema era que a modelagem dos dados se baseava na estrutura de processos quando deveria se basear na estrutura de negócios. Começam então a surgir os primeiros sistemas desenvolvidos especialmente para gerentes: os Sistemas de Informações para Executivos - *Executive Information Systems* (EIS) - , mas as empresas e os negócios evoluem mais rápido do que esses sistemas.

Segundo Weldon (1998), no início da década de 80 surgem as ferramentas *CASE* e Linguagens de 4ª geração, que prometem resolver problemas de usuários que necessitavam de rápidas informações e não podiam lidar com perdas de tempo em desenvolvimentos específicos para suas necessidades e , ambas as ferramentas, não eram versáteis o bastante para atender a todas as necessidades dos gerentes.

Ao longo do tempo, com o crescimento das empresas e o aumento dos negócios, o volume de dados armazenados também aumentou e também surgiu a necessidade de aumento do número de gerentes ou de divisão de tarefas em diversos níveis gerenciais. Com isso ocorreu a necessidade de crescimento da análise de dados, de respostas rápidas, confiáveis e adaptáveis as novas formas de gerenciamento das empresas e negócios. Novos métodos de gestão empresarial foram elaborados, tais como: Reengenharia (HAMMER, 1994) e o Gerenciamento pela Qualidade Total (DEMING e SCHERKENBACH, 1992).

Segundo Fisher (1998), quando as necessidades de progresso tecnológico e as necessidades de mercado convergem, estes realizam mudanças primordiais na prática dos negócios e a evolução das Tecnologias da Informação possibilitou muitas empresas a encarar um ambiente cada vez mais competitivo.

Segundo Bispo e Cazarini (1998), nos anos 90, diversos sistemas para apoio e suporte às decisões nas empresas foram desenvolvidos. Dentre essas novas ferramentas está o ERP (*Enterprise Resource Planning*), como ferramenta de gestão integrada utilizada para gerenciamento no ambiente operacional e, também, uma nova geração de Sistemas: o *Data Warehouse*, o OLAP e o *Data Mining* que vêm sendo utilizadas para o gerenciamento no ambiente gerencial.
Com as ferramentas *Data Warehouse* e OLAP, os relatórios e as consultas passam a ser feitos pelos próprios usuários dos sistemas sem que haja a necessidade de um profundo conhecimento em tecnologias computacionais, sendo sua confecção barata, rápida, confiável e adaptável aos modelos diversos de negócios. Ao usarem essas ferramentas os gerentes gastam um tempo bem menor manipulando os dados e construindo modelos conforme suas necessidades, usando melhor o tempo para as necessárias análise e soluções de problemas.

O surgimento dessa nova geração de Sistemas de Apoio a Decisão não inutiliza nem substitui os tradicionais e antigos sistemas. Na maioria das vezes os antigos e os novos sistemas atuam em conjunto no auxílio a gerência dos negócios, na solução de problemas e na elaboração de estratégias novas. Informações obtidas através do *Data Mining* ou do OLAP podem alimentar qualquer sistema que trabalhe em otimização ou na linha de pesquisa operacional, por exemplo.

2.2 Data Warehouse

2.2.1 Definição

A surpreendente capacidade do homem em produzir dados vem aumentando de uma maneira crescente desde a última década do século XX e, ao analisarmos as perspectivas, observamos que a tendência de crescimento terá continuidade se transformando na tônica dominante do Século XXI.
Atualmente, os diversos recursos tecnológicos disponíveis facilitam muito o processo de coletar dados, a Internet é um exemplo, e indica o desenvolvimento de tecnologias novas com capacidade de tratar os dados transformando-os em úteis informações e extraindo, a partir deles, conhecimentos (*knowledge discovery*) (BRACHMAN e ANAND, 1996).

Vários pesquisadores das áreas de inteligência artificial, *machine learning* (WEISS e KULIKOWSKI, 1991), estatística, base de dados espaciais, aquisição de conhecimentos, visualização de dados, entre outras, consideram a possibilidade de obtenção de informações valiosas e extração de conhecimentos geradas por grandes massas de dados, como sendo um ponto chave de pesquisa, e devido a essa importância, têm demonstrado grande interesse do assunto, que é conhecido como *Data Mining*.

Visando facilitar o trabalho de Mineração de Dados, aponta-se como fundamental uma análise criteriosa dos dados armazenados nas várias bases.

Além disso, as empresas de grande porte possuem um enorme volume de dados que estão espalhados em vários sistemas diferentes, não possibilitando a busca de informações que permitam a tomada de decisão baseada em um histórico dos dados, o que possibilitaria a identificação de tendências e o posicionamento das empresas estrategicamente para a competitividade e para a maximização dos lucros.

Dessa forma, foi introduzido no mercado um novo conceito que permite reagrupar esses dados espalhados pelos diversos sistemas e reorganizá-los estrategicamente: o *Data Warehouse*.

Ferramenta definida, segundo Simon (1995), como "uma fonte de dados voltada para o suporte à decisão de usuários finais derivada de diversos bancos de dados operacionais".

Ou, conforme Inmon (2005), "um conjunto de dados baseado em assuntos, integrado, não-volátil e variável em relação ao tempo, de apoio às decisões gerenciais".

Ou ainda, segundo Singh (2001), "um ambiente de suporte a decisão que alavanca os dados armazenados em diferentes fontes e os organiza e entrega aos tomadores de decisões da empresa, independente de plataforma que utilizam ou de seu nível de qualificação técnica".

Enfim, um conceito que consiste em organizar dados corporativos da melhor forma, a fim de subsidiar os gerentes e diretores das empresas com informações para a tomada de decisão, num banco de dados paralelo aos sistemas operacionais.

A Figura 2 mostra uma base de dados operacional típica comparada ao *Data Warehouse*.

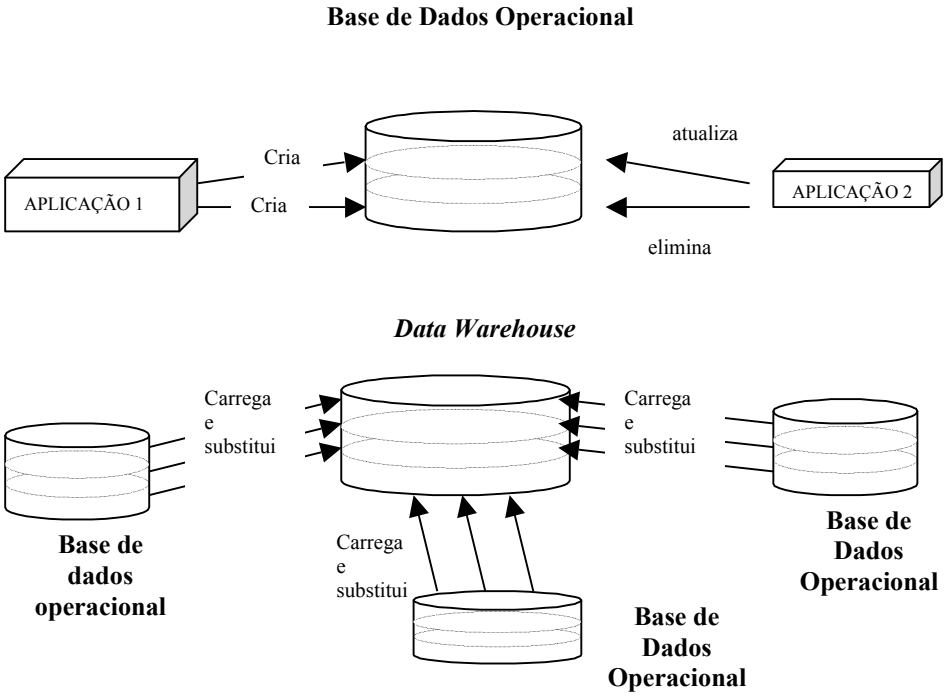

Figura 2: Modelos de update: Base de Dados x *Data Warehouse* (SIMON, 1995)

2.2.2 Princípios

O primeiro princípio básico que abordaremos será o da Não-volatilidade. Segundo Inmon (2005) e Simon (1995), o *Data Warehouse* é não volátil e isso significa que ele não está sujeito às operações de atualização, comuns nas bases de dados operacionais. Esse princípio deve ser considerado no seu projeto e na sua construção.

Se tomamos como exemplo o modelo relacional, verificamos que, freqüentemente, ocorre a substituição de valores dos atributos, a inclusão e a eliminação de registros e outras alterações. No entanto, no *Data Warehouse*, essas operações não são feitas, conforme pode ser observado na figura 1. O *Data Warehouse* recebe toda a carga de dados em regulares intervalos de tempo e obedece algumas regras de extração.

Segundo Kimball, Reeves, Ross e Thornthwaite (1998), o processo de carga, normalmente, "envolve um sofisticado tratamento para eliminação de inconsistências de tipos de dados, tamanhos, significado dos atributos,

codificação e outras propriedades intrínsicas dos dados que estão sendo recuperados".

Logo após essa carga, o *Data Warehouse* está preparado para as consultas dos sistemas de informações gerenciais e por sistemas de apoio à decisão.

O segundo princípio básico é a Orientação por assunto. Para Inmon (2005), utiliza-se sub conjuntos das bases de dados operacionais para organizar os *Data Warehouses* e sua construção é feita da extração de dados de diferentes aplicações, que podem estar em várias plataformas diferentes, o que requer capacidade de integração.

A capacidade de integração é um outro princípio fundamental, já que consiste na montagem de um esquema inequívoco e global, que parte de aplicações múltiplas e diferentes fontes de dados que não são uniformes e que usam critérios próprios (INMON, 2005).

Por fim, o princípio da sensibilidade ao tempo. Segundo Inmon (2005), esse princípio (*time variance*) é relevante porque o *Data Warehouse* refletirá sempre um momento no tempo.

2.2.3 Arquitetura

Um ambiente de *Data Warehouse* pode ser criado a partir de dados operacionais originados de sistemas *Operational Data Store* (ODS) legados ou de dados externos e, ainda, de maneira lógica com a união de *Data Marts* (DM). Os DMs são grupos de dados orientados por assuntos estratégicos do negócio e , também são oriundos de dados dos sistemas ODS ou de dados externos.

Cada *Data Warehouse* deve ser moldado conforme as necessidades dos usuários, alinhado às suas áreas funcionais na empresa e de acordo com as condições de negócio e as pressões de competitividade. No entanto, são quatro as arquiteturas mais utilizadas para seu desenvolvimento:

- Arquitetura *Top-down*;
- Arquitetura *Bottom-up*;
- Arquitetura Híbrida;
- Arquitetura Federada.

Segundo Kimball (2002), o plano de arquitetura é um importante fator na elaboração do projeto de um *Data Warehouse* como ferramenta de comunicação, flexibilidade e planejamento, facilitando o aprendizado e o aumento da produtividade.

Para o sucesso na obtenção de um *Data Warehouse* que atenda as necessidades de informação da Empresa, as sutilezas existentes entre cada uma das abordagens devem ser conhecidas.

2.2.4 Arquitetura *Top-Down*

Essa abordagem entende o *Data Warehouse* como o ponto central do ambiente analítico da empresa, possuindo atomicidade ou transação de dados que são purgados de um ou mais sistemas e integrados conforme a modelagem de dados normalizada da empresa. É neste ambiente que os dados são "enxugados", dimensionados e distribuídos para os *Data Marts* conforme demonstrado na figura 3 abaixo:

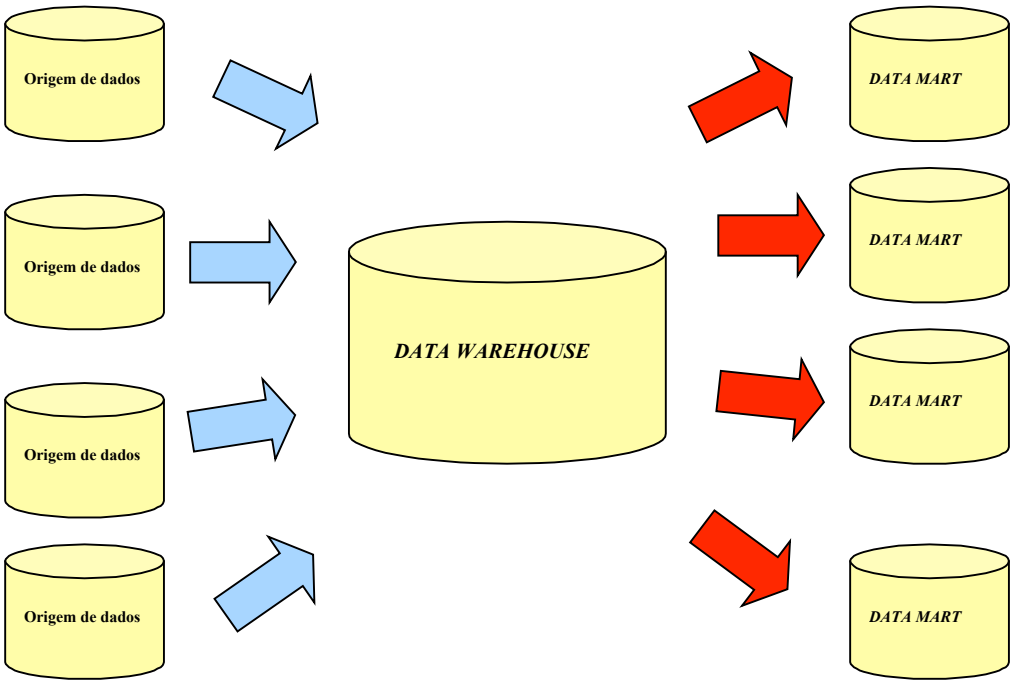

Figura 3: Arquitetura *Top Down*

2.2.5 Arquitetura *Bottom-up*

Segundo Kimball (2002), é uma abordagem que tem como característica principal o desenvolvimento inverso à abordagem *Top-dowm* proposto por Imnon. Primeiro, a partir de dados dos sistemas legados ou de dados externos, são criados os *Data Marts* e, a partir dos *Data Marts* é gerado um *Data Warehouse* ou ainda considerar como *Data Warehouse* a união de todos os *Data Marts*.

A figura 4 representa essa abordagem:

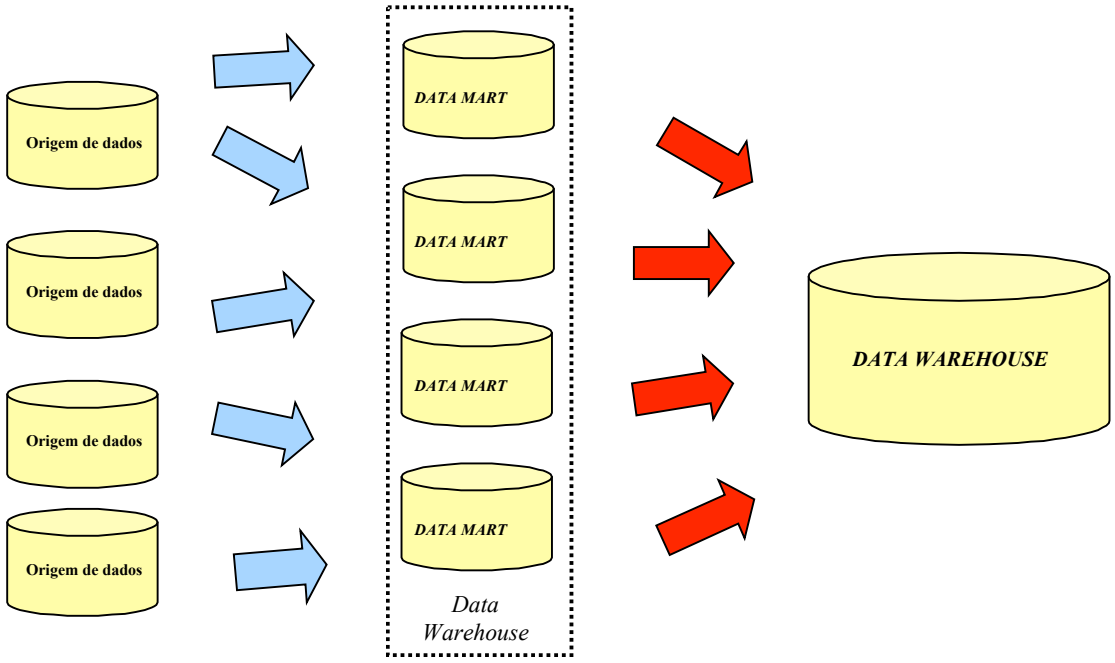

Figura 4: Arquitetura *Bottom-Up*

2.2.6 Arquitetura Híbrida

Essa abordagem, segundo Hackney (1998), tem como finalidade integrar as abordagens de *Top Down* e *Bottom up*, utilizando o que há de melhor nas duas abordagens e tirando proveito da orientação do usuário e da velocidade que existe na abordagem *Bottom-up* sem prejudicar a integração forçada por um *Data Warehouse* que existe na abordagem *Top-Down*. Primeiro, todo o *Data Warehouse* da organização é modelado para posterior implementação de partes desse modelo representando os assuntos da organização e que serão os *Data Marts*.

A seguir, a figura 5 mostra a Arquitetura Híbrida:

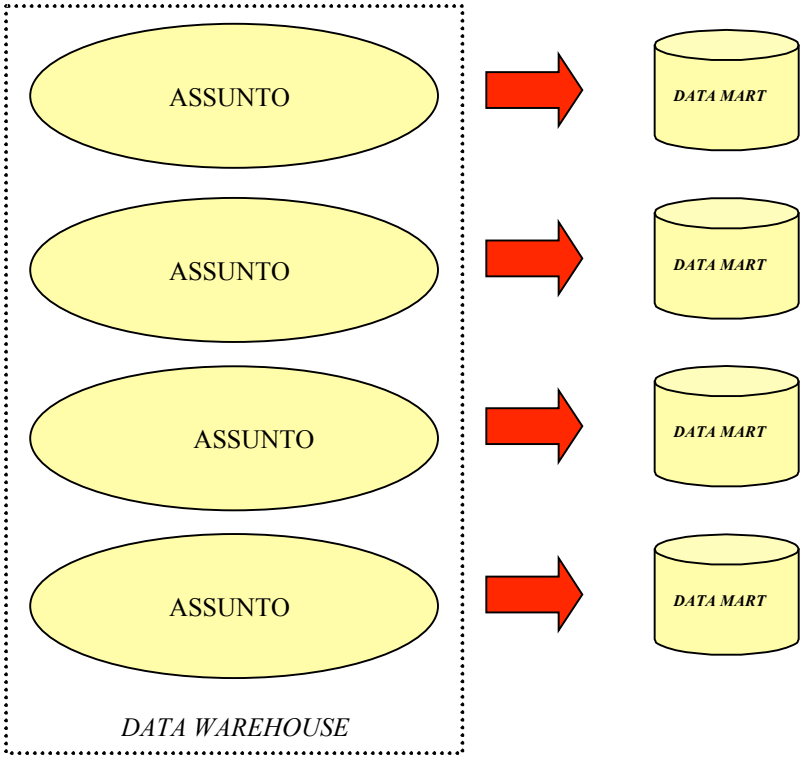

Figura 5: Arquitetura Híbrida

2.2.7 Arquitetura Federada

Segundo Huber (2001), o *Data Warehouse* federado provê uma interface que o torna parecido com um grande *Data Warehouse*, mas na verdade apenas soma camadas sobre os *Data Warehouses* existentes permitindo a execução de consultas sobre um enorme *Data Warehouse* virtual.

Um *Data Warehouse* federado é composto por camadas de *Data Warehouses* existentes, heterogênios e distribuídos, acessados através de uma camada de integração organizadora dos dados de forma homogênea para carregar o *Data Warehouse* federado, conforme demonstrado na figura

O *Data Warehouse* é uma ferramenta que está suprindo os anseios dos analistas de negócios e gerentes. Por intermédio do *Data Warehouse* , atualmente, esses usuários obtêm ganhos qualitativos e quantitativos no suporte ao processo decisório. Menos tempo é perdido no acesso e análise de dados, sobrando o suficiente para que se voltem as estratégias para os negócios, com base em fatos e informações comprovadas e analisadas. No entanto, o *Data Warehouse* não pode, sozinho, realizar todas as consultas e análises necessárias aos usuários. A seguir apresentaremos uma segunda ferramenta, o "OLAP", que proporciona uma maior sofisticação nos dados.

2.2.8 Ferramenta OLAP

As ferramentas de processamento analítico on-line ou OLAP (*on-line analytical processing*) são formadas por conjuntos de tecnologias projetadas, especialmente, para suportar o processo decisório e às estratégias de negócio com consultas, análise e cálculos mais aprimorados nos dados corporativos, que estejam ou não, armazenados. num *Data Warehouse*.

A HYPERION (2006) disponibiliza uma tabela em que compara algumas características das ferramentas OLTP, *Data Warehouse* e OLAP.

Sistema	**OLTP**	***DATA WAREHOUSE***	**OLAP**
Propósito	Operacional	Armazenamento e acesso aos dados histórico detalhados	Analítico
Tipo de acesso	Leitura e escrita	Somente leitura	Leitura e escrita

Modo de acesso	Atômico	Consultas e relatórios	Iterativo, comparativo e investigativo
Escopo	Aplicações específicas	Dados corporativos	Análise de dados
Nível de detalhe	Transação	Dados limpos e sumarizados	Sumarizados e calculados
Estrutura dos dados	Normalizados	Desnormalizados	Dimensional e hierárquicos
Implementação	Vários meses ou anos	Vários anos	Semanas

Tabela 2: Comparativo entre *OLTP*, *Data Warehouse* e *OLAP*

Segundo Kimball (2002) as estruturas OLAP são:

- ROLAP (OLAP Relacional) que é um conjunto de interfaces feitas para os usuários e aplicações que dão uma aparência dimensional os SGBDs.
- MOLAP (OLAP Multidimensional) que é o conjunto de tecnologias, aplicações e interfaces de SGBDs proprietários com aparência multidimensional e
- HOLAP (OLAP híbrido) que combina o ROLAP com o MOLAP.
- WOLAP (OLAP web) que é a tecnologia *OLAP* direcionada para a internet.

O termo OLAP foi definido por E.F.Codd (2006), também criador, em 1993, de um conjunto de 12 regras utilizadas para avaliar a ferramenta por desenvolvedores e usuários. São elas:

- Visão conceitual multidimensional, onde os usuários manipulem os modelos multidimensionais de dados com facilidade e intuitivamente.
- Transparência - interação fácil com os front-ends habituais dos usuários devendo permitir a inclusão de uma ferramenta analítica onde o usuário desejar sem que isso provoque impacto na funcionalidade.
- Acessibilidade - Tratamento dos dados heterogêneos de forma lógica que permita a conversão para apresentação aos usuários de forma única, coerente e consistente.
- Desempenho consistente de fornecimento de informações - apesar do tamanho do banco de dados, o usuário não deve observar redução significante no desempenho de fornecimento de informações.
- Arquitetura cliente/servidor - É importante que a ferramenta tenha capacidade de operar num ambiente cliente/servidor
- Dimensionalidade genérica - A dimensão dos dados não deve influenciar a estrutura dos dados e os formatos dos relatórios.
- Manipulação dinâmica da matriz esparsa - Com base na densidade dos dados, a ferramenta deve possibilitar o ajuste do esquema físico para o desempenho máximo.
- Suporte multiusuário - Prover acesso simultâneo sem prejudicar a segurança e a integridade dos dados.
- Operações irrestritas com dimensões cruzadas - Qualquer conjunto de dados deve poder ser acessado, a qualquer momento, para cáculos.
- Manipulação intuitiva de dados - A realização dos cálculos e a manipulação dos dados devem ocorrer da maneira mais intuitiva possível.
- Relatórios flexíveis - Os relatórios devem ter a capacidade de fazer a apresentação dos dados de forma lógica e sintetizada ou, ainda, informações que sejam o resultado de cálculos de um modelo criado, conforme qualquer visão.
- Dimensões e níveis de agregação ilimitados - deve ser possível a acomodação de pelo menos quinze e até vinte dimensões de dados, dentro de um modelo analítico comum e cada dimensão deve possibilitar um número sem limites de níveis de agregação definido pelo usuário.
-

2.3 O Processo KDD

2.3.1 Introdução

Nas últimas décadas, todo o mundo tem armazenado uma considerável quantidade de dados, superando consideravelmente as nossas habilidades de interpretação, gerando uma necessidade de criação de técnicas e ferramentas que automatizem e analisem a base de dados de maneira inteligente (FAYYAD, 1996).

Essas ferramentas e técnicas que procuram transformar os dados armazenados em conhecimento, são o objetivo do chamado *Knowledge Discovery in Databases - KDD* (descoberta de conhecimento em bases de dados).

Um número crescente de publicações vêm se dedicando ao tema.

Segundo Fayyad (1996), o termo *Knowledge Discovery in Databases* ou KDD, foi criado em 1989 como referência ao amplo processo para encontrar conhecimento nos dados e enfatizar uma aplicação em especial - o método *Data Mining* (Mineração de Dados).

KDD é todo processo de descoberta de conhecimento útil nos dados, enquanto *Data Mining* refere-se à aplicação de algoritmos para

extração de modelos dos dados. No entanto, os termos KDD e *Data Mining* foram considerados como sinônimos por muitos autores até 1995.
Dessa forma, cabe ressaltar que o processo KDD é dependente de uma nova geração de técnicas e ferramentas de análise de dados envolvendo diversas etapas. A principal etapa desse processo chama-se *Data Mining* ou Mineração de Dados, também conhecida como reconhecimento de padrões ou processo de arqueologia de dados (CHEN; HAN; YU, 1996).
Na Conferência Internacional de KDD, ocorrida em Montreal - 1995, foi apresentada uma definição para cada um dos dois termos. Para Adriaans e Zantinge (1996): "...KDD será empregado para todo o processo de extração de conhecimento dos dados. Neste contexto, conhecimento significa relacionamento e padrões entre elementos de dados. O termo Mineração de Dados deveria ser utilizado para os estágios de descoberta do processo de KDD".
Essa relação existente entre KDD e *Data Mining*, é retratada na Figura 7.

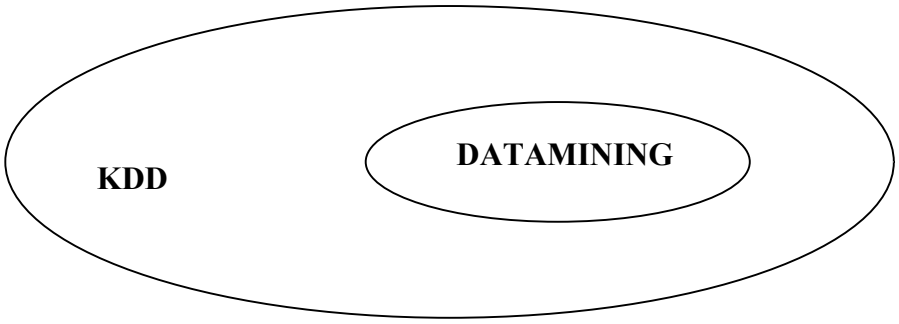

Figura 7 - Diferença entre KDD e Data Mining

Assim, o processo global para achar e interpretar modelos extraídos de dados é chamado de processo KDD, tipicamente iterativo e interativo, que envolve aplicações específicas repetidas de métodos ou algoritmos *Data Mining* e a interpretação dos padrões gerados por estes algoritmos (FAYYAD, 1996).

2.3.2 As Etapas do Processo

KDD é um processo de descoberta de conhecimento em bases de dados que envolvem uma diversificada abrangência, como: estatística, banco de dados, matemática, visualização de dados, inteligência artificial e reconhecimento de padrões. Este processo utiliza técnicas, métodos e algoritmos com origem dessas áreas, em que o principal objetivo é a extração do conhecimento partindo de grandes bases de dados.
Sendo o processo de KDD um conjunto de atividades contínuas para o compartilhamento do conhecimento descoberto a partir de bases de dados, segundo FAYYAD (1996), esse conjunto é composto de 5 (cinco) etapas:

- Seleção dos dados;
- Pré-processamento e limpeza dos dados;
- Transformação dos dados;
- Data Mining;
- Interpretação e Avaliação dos resultados.

Etapas que podem ser visualizadas através da Figura 8, a seguir:

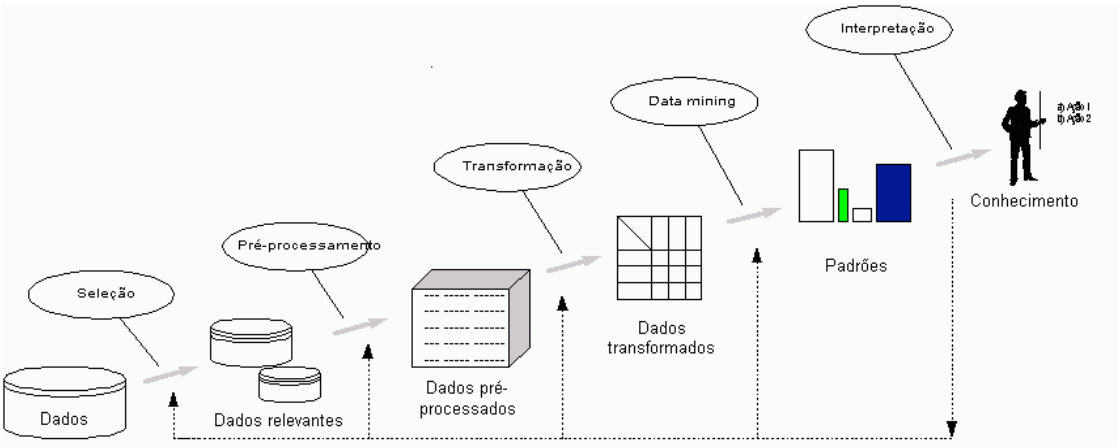

Figura 8: Processo KDD (FAYYAD, 1996)

O processo de KDD inicia-se, obviamente, com o entendimento do domínio da aplicação e dos resultados finais a serem atingidos. A seguir, é feito um agrupamento de uma massa de dados organizadamente, alvo da prospecção. A etapa seguinte, limpeza dos dados (*data cleaning),* é realizada através de um pré-processamento dos dados, com vistas a adequá-los aos algoritmos. Isso se faz com a integração de dados heterogêneos, com a eliminação de dados incompletos e outras. Essa etapa é uma das mais demoradas podendo tomar até 80% de todo o tempo necessário para o processo completo, devido às muito conhecidas dificuldades de integração de bases de dados heterogêneas (MANNILA, 1996).

Os dados pré-processados devem ainda sofrer uma transformação com o conseqüente armazenamento adequado, visando facilitar o uso das técnicas de *Data Mining.*

É aí que o uso de *Data Warehouse* (Armazenamento de Dados) se torna significativo, pois com essa tecnologia as informações estarão armazenadas de maneira bastante eficiente.

Segundo Inmon (2005), o *Data Warehouse* é um conjunto de dados, integrado, não volátil e variável em relação ao tempo, dando apoio às decisões gerenciais.

Dando continuidade ao processo, chega-se à etapa de *Data Mining* especificamente, que se inicia com a escolha das ferramentas (algoritmos) que serão utilizadas. Essa escolha depende basicamente do objetivo do processo de KDD: classificação, agrupamento, regras associativas, ou outra. De maneira geral, na fase de *Data Mining*, as ferramentas especializadas buscam padrões nos dados. Essa pesquisa pode ser efetuada pelo sistema automaticamente, de forma livre (*roams* - percorrer/vasculhar o banco de dados) ou interativamente com um analista responsável pela geração de hipóteses, chamada análise direcionada (*directed analysis*) ou também chamada aprendizado supervisionado (*supervised learning*), onde temos como que um "professor" que "ensina" o sistema indicando, por exemplo, quando uma premissa foi ou não correta.

Várias ferramentas distintas, como árvores de decisão, redes neurais, sistemas baseados em regras e programas estatísticos, tanto de forma isolada quanto em combinação, podem ser aplicadas ao problema. Geralmente, o processamento de busca é interativo, de maneira que os analistas revêem o resultado, formam um novo conjunto de questões para aprimorar a busca em um determinado aspecto das descobertas, e realimentam o sistema com novos parâmetros.

O sistema de *Data Mining* gera um relatório das descobertas, ao final do processo, que passa a ser interpretado pelos analistas de mineração. Somente após a interpretação das informações obtidas é encontrado conhecimento.

No passado, a aura futurológica que envolvia o *Data Mining* supunha que ele eliminaria a necessidade de analistas estatísticos para a construção de modelos preditivos. No entanto, analistas serão sempre necessários na avaliação dos modelos e validação da plausibilidade das predições realizadas. Considerando o fato do *software* de *Data Mining* não contar com a experiência e intuição humana para reconhecer a diferença entre uma correlação relevante e irrelevante, analistas estatísticos permanecerão em alta demanda (THEARLING, 2000).

Uma diferença marcante entre *Data Mining* e outras ferramentas de análise está na forma como exploram as inter-relações entre os dados. As várias ferramentas de análise disponíveis utilizam um método baseado na verificação, onde o usuário constrói hipóteses sobre inter-relações específicas e aí verifica ou refuta, através do sistema. Esse modelo torna-se dependente da habilidade e intuição do analista em propor interessantes hipóteses, na manipulação da complexidade do espaço de atributos, e no refinamento da análise baseado nos resultados de consultas potencialmente complexas ao banco de dados. Já o processo de *Data Mining* é o responsável pela geração de hipóteses, garantindo maior qualidade, rapidez e integridade aos resultados.

2.3.2.1 Seleção dos Dados

Após a definição do domínio sobre o qual se quer executar o processo de descoberta, o passo seguinte é a seleção e a coleta do conjunto de dados ou variáveis necessárias. A maioria das empresas já possui bases de dados. No entanto, nem sempre todos os dados que serão utilizados estão disponíveis em bases adequadas, o que torna necessário um trabalho de compatibilidade.

2.3.2.2 Limpeza dos Dados

É a atividade através da qual dados estranhos ou inconsistentes e ruídos, são tratados e onde são estabelecidas as estratégias para a resolução dos problemas de ausência de dados.

2.3.2.3 Transformação dos Dados

Nesta etapa, como já foi citado, o uso de *Data Warehouse* se expande consideravelmente, visto que são nessas estruturas que as informações estão alocadas da forma mais eficiente. Um *Data Warehouse* é um repositório de informações para suportar decisões, como já vimos no capítulo 3, que funciona coletando dados a partir de diversas aplicações de uma organização, integrando e organizando os dados em áreas lógicas de assuntos, armazenando as informações de maneira que elas se tornem acessíveis e compreensíveis a pessoas não técnicas disponibilizando os dados da melhor forma possível aos decisores, para que possam ser aplicadas técnicas de análise e extração de informações.

2.3.2.4 *Data Mining*

A descoberta do conhecimento é uma das atividades mais fascinantes, onde a maioria dos métodos de *Data Mining* são baseados em conceitos de aprendizagem de máquina, estatística,

reconhecimento de padrões, agrupamento, classificação e modelos gráficos.

2.3.2.5 Interpretação e Avaliação dos Resultados

Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas que devem possibilitar uma análise criteriosa na identificação da necessidade de retorno a qualquer um dos estágios anteriores do processo de KDD.

Em cada etapa do processo KDD pode ser identificada a necessidade de retorno para cada uma das etapas anteriores. Se, por exemplo, na etapa de codificação ou mesmo na etapa de *Data Mining*, é identificado que os dados não estejam totalmente consistentes ou se for identificada a necessidade de um dado que não previsto anteriormente, isso pode levar ao retorno para a etapa de consistência ou mesmo de seleção dos dados.

2.3.3 Áreas Relacionadas ao KDD

O processo KDD é interdisciplinar e envolve áreas relativas a estatística, banco de dados, matemática, visualização de dados, inteligência artificial, aprendizado de máquina e sistemas especialistas. Este processo utiliza métodos, técnicas e algoritmos oriundos destas áreas, com o principal objetivo de extrair conhecimento a partir de grandes bases de dados.

Na figura 9 visualizamos a relação das áreas no processo KDD:

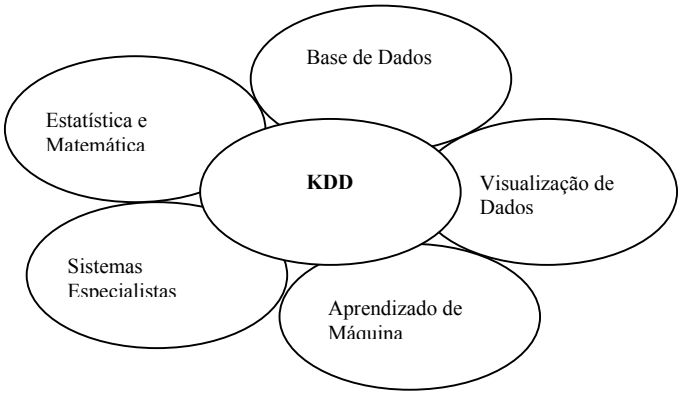

Figura 9: KDD é um campo multi-disciplinar (ADRIAANS e ZANTINGE, 1996)

2.3.3.1 Aprendizado de Máquina

Na área do aprendizado de Máquina são utilizados estratégias de aprendizado de máquina ou modelos cognitivos e paradigmas para a aquisição automática de conhecimento.

2.3.3.2 Bases de Dados

Nesta área existem tecnologias específicas e, também, uma série de pesquisas que têm como objetivo melhor explorar as diversas características dos dados a serem trabalhados.

2.3.3.3 Estatística e Matemática

É comum que modelos estatísticos ou matemáticos sejam construídos para que sejam criadas regras, padrões e regularidades.

Especificamente na Estatística é disponibilizado um grande número de procedimentos técnicos e resultados de testes para as tarefas de *Data Mining,* tais como, por exemplo, para a verificação se as estimativas e procedimentos de pesquisa estão consistentes sob determinados critérios de avaliação e para a identificação do grau de incerteza.

2.3.3.4 Sistemas Especialistas

Os sistemas especialistas são programas de Inteligência Artificial gerados para a resolução de problemas reais. Inicialmente, estes sistemas ofereciam somente mecanismos que representavam o conhecimento, o raciocínio e as explicações. Depois, foram sendo incorporadas ferramentas objetivando a aquisição do conhecimento.

2.3.3.5 Visualização de Dados

A Visualização de Dados tem um papel importante visto que em diversos momentos é necessária a interação entre o ser humano e o processo de descoberta. Como exemplo, podemos citar a prévia análise dos dados que vão ou não fazer parte do processo onde são realizadas várias consultas que usam ferramentas de análise ou até mesmo de visualização de dados.

Para a visualização, recorre-se a distintas formas, tais como: ícones, gráficos e figuras.

2.4 A Etapa DATA MINING

Para que se tenha uma compreensão ampla a respeito do *Data Mining,* apresentaremos nesta seção detalhadamente cada passo desta etapa.

2.4.1 Introdução ao *Data Mining*

Data mining é a parte mais interessante do processo, sendo que no contexto empresarial é a que mais impulsiona e auxilia o decisor a descobrir filões de mercado.
Segundo Possas et all (1998), está provado que o cérebro humano consegue fazer até 8 (oito) comparações simultaneamente. A funcionalidade do *Data Mining* está em ampliar esta comparação para "infinito" e tornar isso visível ao olho humano.
Muito conhecimento encontra-se escondido na vasta quantidade de dados disponíveis nos bancos de dados das empresas e é com o *Data Mining* que se pode transformar esses dados brutos em informação valiosa a fim de auxiliar a tomada de decisão.
A diferença entre o *Data Mining* e as técnicas estatísticas está na utilização dos próprios dados para a descoberta dos padrões e não na verificação de padrões hipotéticos.
As bases de dados armazenam conhecimento que podem nos auxiliar na melhoria dos negócios e as técnicas tradicionais permitem, apenas, verificar hipóteses que são, aproximadamente, apenas 5% de todas as relações encontradas por esses métodos. O *Data Mining* pode descobrir as outras relações desconhecidas: os 95% restantes. Ou seja, pode-se dizer que técnicas convencionais "falam" para a base de dados, enquanto *Data Mining* "ouve" a base de dados.
Para Thearling (2000), o *Data Mining* explora as bases de dados através de dezenas de centenas de pontos de vista diferentes.
O *Data Mining* não veio para substituir as técnicas estatísticas tradicionais, sendo uma extensão dos métodos estatísticos que, em parte, são o resultado de uma considerável mudança na comunidade estatística. O poder cada vez maior dos computadores aliado aos custos mais baixos e com a necessidade crescente da análise de enormes conjuntos de dados com milhões de linhas, permitiu o desenvolvimento de técnicas baseadas na exploração de soluções possíveis pela força bruta (THEARLING, 2000).

2.4.2 Definições de *Data Mining*

São muitas as definições e conceitos de *Data Mining* encontradas na literatura, sendo que a seguir listaremos algumas delas.
Para Silva, (2000), *"Data Mining* é uma técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão".
Segundo Rodrigues (2000), *"Data Mining* é um processo que encontra relações e modelos dentro de um grande volume de dados armazenados em um banco de dados".
Conforme Possas et all. (1998), *"Data Mining* é um conjunto de técnicas que envolve métodos matemáticos, algoritmos e heurísticas para descobrir padrões e regularidades em grandes conjuntos de dados.
Nimer e Spandri (1998) entendem que, "*Data Mining* é uma ferramenta utilizada para descobrir novas correlações, padrões e tendências entre as informações de uma empresa, através da análise de grandes quantidades de dados armazenados em *Data Warehouse* usando técnicas de reconhecimento de padrões, estatísticas e matemáticas".
Para uma das maiores autoridades em D*ata Mining* do mundo, o pesquisador Gregory Piatetsky - Shapiro, conforme relato a "NEGÓCIOS EXAME": *"Data Mining* é a extração de informações potencialmente úteis e previamente desconhecidas de grandes bancos de dados".

2.4.3 Objetivos do *Data Mining*

O principal objetivo do *Data Mining* é a extração de valiosas informações dos dados, para a descoberta do "ouro escondido". Esse "ouro" são as valiosas infomações que os dados contém. Pequeninas alterações nas estratégias oriundas das descobertas das ferramentas de *Data Mining*, podem transformar-se em significativas diferenças no caixa da empresa. Com a ampliação do uso dos *Data Warehouses*, as ferramentas de *Data Mining* tornaram-se primordiais. No entanto, é importante lembrar que o uso de um *Data Warehouse* não é necessário para a aplicação de uma ferramenta de *Data Mining*. Basta que se tenha dados.
Várias ferramentas de análise de dados, tipo geradores de relatórios ou análises estatísticas, usam o termo *Data Mining* nos seus *softwares* computacionais. Produtos com bases em inteligência artificial também se intitulam ferramentas de *Data Mining*. Porém, o que denomina-se um verdadeiro *Data Mining*? O principal objetivo do *Data Mining* é a descoberta do conhecimento, que com sua metodologia extrai informações preditivas das bases de dados.

2.4.4 A origem do *Data Mining*

Para Freitas (2000), "*Data Mining* é um campo interdisciplinar, que emergiu da interseção entre várias áreas, principalmente aprendizado de máquina" (uma subárea da inteligência artificial, estatística e banco de dados), conforme mostra a Figura 10, a seguir.

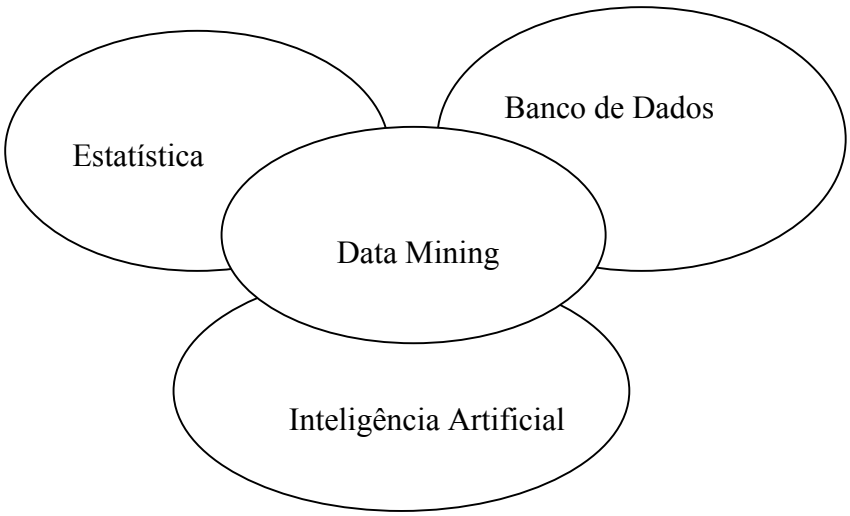

Figura 10: A origem do *Data Mining*

Então, *Data Mining* é a combinação de diferentes técnicas de sucesso comprovado, como estatística, inteligência artificial e bancos de dados.

2.4.4.1 Inteligência Artificial

Inteligência Artificial ou IA, é uma disciplina com base nos fundamentos da heurística, diferentemente da estatística, sua tentativa é a de imitar a maneira como o homem pensa na resolução dos problemas estatísticos.
Segundo Rodrigues (2000), em função dessa abordagem, a IA 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.
O aprendizado de máquina, que podemos descrever como a união entre a estatística e a IA, tenta fazer com que os programas de computador "aprendam" com os dados estudados por eles, de forma que esses programas tomem diferentes decisões com base nas características dos dados estudados.

2.4.4.2 A Estatística

Não seria possível termos o *Data Mining* sem a estatística, já que a mesma é a base de construção do *Data Mining*.
A Estatística Clássica envolve conceitos como variância, distribuição normal, desvio simples, análise de regressão, análise de conjuntos, intervalos de confiança e análise de discriminante, todos usados para o estudo dos dados e seus relacionamentos.
Essas são as bases fundamentais onde se apóiam as mais avançadas análises estatísticas. E, sem dúvida, a análise estatística clássica desempenha um papel fundamental na essência das atuais ferramentas e técnicas de *Data Mining*.

2.4.4.3 Banco de Dados

Uma das técnicas mais utilizadas para melhorar a base de dados é o *Data Warehouse,* como já vimos anteriormente.

Vale lembrar, no entanto, que o *Data Warehouse* também pode ser definido como um conjunto de tecnologias propiciam a conversão de uma grande quantidade de dados em informação útil, transformando um banco de dados operacional em um ambiente que permite o uso dos dados estrategicamente. É um ambiente e não um produto.

2.4.4.4 Características Desejáveis do Conhecimento a ser Descoberto por *Data Mining*

Segundo Freitas (2000), idealmente, o conhecimento a ser descoberto deve satisfazer três propriedades, a saber:

- compreensível por usuários humanos;
- correto (tanto quanto possível);
- interessante / útil / novo (surpreendente).

2.4.4.5 Características Esperadas do Método de Descoberta de Conhecimento por *Data Mining*

Ainda, conforme Freitas (2000), o método de descoberta do conhecimento deve apresentar as seguintes características:

- flexível (facilmente modificável).
- eficiente (rápido);
- genérico (aplicável a vários tipos de dados);
-

2.4.5 As Principais Técnicas de *Data Mining*

O *Data Mining* é um campo que, atualmente, compreende muitas ramificações importantes. Cada tipo de tecnologia tem suas próprias vantagens e desvantagens, do mesmo modo que nenhuma ferramenta consegue atender todas as necessidades em todas as aplicações.
Dentre as técnicas de *Data Mining*, destacam-se as apresentadas nas seções a seguir.

2.4.5.1 Árvores de Decisão

Árvore de Decisão é um método adequado quando o objetivo do *Data Mining* é classificação de dados ou predição de saídas. É conveniente usar árvore de decisão quando o objetivo for categorizar dados de arquivos. Também é uma boa escolha quando o objetivo é gerar regras que podem ser facilmente entendidas, explicadas e traduzidas para linguagem natural.

2.4.5.2 Redes Neurais

As Redes Neurais tentam construir representações internas de modelos ou padrões detectados nos dados, mas essas representações não são apresentadas para o usuário.

Estruturalmente, uma Rede Neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão que conectam as camadas.

Segundo Din (1998), as Redes Neurais Artificiais utilizam um conjunto de elementos de processamento (ou nós) análogos aos neurônios no cérebro. Estes elementos de processamento são interconectados em uma rede que pode identificar padrões nos dados uma vez expostos aos mesmos, ou seja, a rede aprende através da experiência, tais como as pessoas.

2.4.5.3 Análise de Agrupamento

Esta técnica agrupa informações homogêneas de grupos heterogêneos entre os demais e aponta o item que melhor representa cada grupo, permitindo desta forma que se consiga perceber a característica de cada grupo. Desse modo, objetos dentro do mesmo grupo são os mais semelhantes possíveis, enquanto que objetos de grupos diferentes são os mais diferentes possíveis.

Por exemplo, suponha que os objetos sejam clientes, e que se tenha vários atributos descrevendo cada cliente, tais como a idade, faixa de salário, sexo e outros. Analisando esses dados, um sistema de *Data Mining* pode, por exemplo, criar um grupo de clientes com idade baixa e faixa de salário baixa, outro grupo de clientes com idade alta e faixa de salário alto e assim por diante. Essa diferenciação dos clientes em grupos pode ser bastante útil, já que clientes de grupos diferentes, presumidamente, tendem a ter comportamentos de compra bem diferentes (FREITAS, 2000).

2.4.5.4 Indução de Regras

A Indução de Regras (*Rule Induction)* se refere à detecção de tendências dentro de grupos de dados ou de "regras" sobre os dados. As regras são, então, apresentadas aos usuários como uma lista "não encomendada", ou seja, sem que obedecam algum critério previamente estabelecido.

Indução de Regras é o processo de analisar uma série de dados e, a partir dela, gerar padrões. O processo é, em sua essência, semelhante àquilo que um analista humano faria em uma análise exploratória.

Consiste na descoberta de regras de previsão, do tipo SE...ENTÃO, onde a parte SE (a "condição") da regra especifica alguns valores de atributos previsores e a parte ENTÃO da regra prevê um valor para um determinado atributo cuja previsão é desejada. Por exemplo, suponha que se tenha um banco de dados de vendas de produtos, com dados sobre produtos vendidos e os clientes que compraram aqueles produtos. Assuma que os dados incluem atributos tais como a idade e sexo do cliente e o tipo do produto comprado. Analisando esses dados, um sistema de *Data Mining* poderia descobrir uma regra de previsão do tipo SE ... ENTÃO, tal como: SE (idade_cliente < 18) E (sexo_cliente = "M") ENTÃO (produto_comprado_videogame) (FREITAS, 2000).

Idealmente as regras descobertas deveriam satisfazer três propriedades, a saber:

(a) fazerem previsões corretas, ou seja, na maioria das vezes que a parte "SE" da regra é verdadeira, a parte "ENTÃO" da regra também é verdadeira;

(b) serem compreensíveis para o usuário, ou seja, as regras representam conhecimento em um alto nível de abstração, tal como a regra acima, ao invés de equações matemáticas complexas e não compreensíveis pelo usuário;

(c) serem úteis para a tomada de decisão, o que está relacionado ao fato da regra expressar conhecimento novo ou surpreendente para o usuário. No exemplo acima, o usuário poderia usar a regra descoberta para, por exemplo, fazer uma mala direta direcionada, enviando uma propaganda de um novo videogame apenas para clientes que têm menos de 18 anos e são do sexo masculino (FREITAS, 2000).

Vários algoritmos e índices são usados para executar esse processo. Na Indução de Regras, a grande maioria do processo é feito pela máquina e uma pequena parte é feita pelo usuário.

2.4.5.5 Análise Estatística de Séries Temporais

A estatística é a mais antiga tecnologia em *Data Mining*, e é parte da fundamentação básica de todas as outras tecnologias. Ela incorpora um envolvimento muito forte do usuário, exigindo engenheiros experientes, para construir modelos que descrevam o comportamento dos dados através dos métodos clássicos de matemática. Interpretar os resultados dos modelos requer especialistas (*expertises)*. O uso de técnicas estatísticas também

requer um trabalho muito forte de máquinas/engenheiros.

A análise de séries temporais é um exemplo disso, apesar de freqüentemente ser confundida como um gênero mais simples de *Data Mining* chamado previsão (*Forecasting)*.

Enquanto que a análise de séries temporais é um ramo altamente especializado da estatística, o *Forecasting* é, de fato, uma disciplina muito menos rigorosa, que pode ser satisfeita, embora com menos segurança, através da maioria das outras técnicas de *Data Mining.*

2.4.5.6 Visualização

As técnicas de Visualização são um pouco mais difíceis de definir, sendo que muitas pessoas a definem como "ferramentas complexas de visualização", enquanto outras como simplesmente a capacidade de geração de gráficos.

Nos dois casos, a Visualização mapeia o dado que está sendo minerado de acordo com dimensões especificadas. Nenhuma análise é executada pelo programa de *Data Mining* além da manipulação da estatística básica. O usuário, então, interpreta o dado através do monitor de vídeo.

2.4.6 As Etapas do *Data Mining*

A implementação de um sistema de *Data Mining* pode ser dividida em seis fases interdependentes para que o mesmo atinja seus objetivos finais, descritas a seguir.

2.4.6.1 Entendimento do Problema

A fase inicial do projeto deve ter como objetivo identificar as metas e necessidades partindo de uma perspectiva do problema, e então convertê-las para uma aplicação de *Data Mining* e um plano inicial de "ataque" ao problema.

2.4.6.2 Entendimento dos dados

Esta fase tem como principal atividade a extração de uma amostra dos dados a serem usados e avaliar o ambiente em que os mesmos se encontram.

2.4.6.3 Preparação dos dados

Criação de programas de extração, limpeza e transformação dos dados para utilização pelos algoritmos de *Data Mining.* É nessa etapa que os dados são adaptados para serem inseridos no algoritmo escolhido para processamento.

2.4.6.4 Modelagem do Problema

Seleção do(s) algoritmo(s) dentre os apresentados a serem utilizados e processamento efetivo do modelo. Alguns algoritmos precistam dos dados em formatos específicos, o que acaba causando diversos retornos à fase de preparação dos dados.

2.4.6.5 Avaliação do Modelo

Ao final da fase de modelagem, diversos modelos devem ter sido avaliados sob a perspectiva do analista responsável. Então, o objetivo passa a ser avaliar os modelos com a visão do problema, certificando-se que não existem falhas ou contradições com relação às regras do problema.

2.4.6.6 Divulgação ou Publicação do Modelo

A criação e a validação do modelo permitem o avanço de mais um passo, no sentido de tornar a informação gerada acessível. Isto pode ser feito de várias formas, desde a criação de um *software* específico para tal, até a publicação de um relatório para uso interno.

2.4.7 As Vantagens do *Data Mining*

O uso de *Data Mining* para construção de um modelo traz as seguintes vantagens:

- **Modelos são de fácil compreensão:** pessoas sem conhecimento estatístico (por exemplo, analistas financeiros ou pessoas que trabalham com *data base marketing*) podem interpretar o modelo e compará-lo com suas próprias idéias. O usuário ganha mais conhecimento sobre o comportamento do cliente e pode usar esta informação para otimizar os processos dos negócios.
- **Grandes bases de dados podem ser analisadas:** grandes conjunto de dados, de até vários *gigabytes* de informação podem ser analisados com *Data Mining*.
- ***Data Mining* descobre informações não esperadas:** como muitos modelos diferentes são validados, alguns resultados inesperados podem surgir. Em diversos estudos, descobriu-se que combinações de fatores particulares apresentaram resultados inesperados.
- **Variáveis não necessitam de recodificação:** *Data Mining* lida tanto com variáveis numéricas (quantitativas) quanto categóricas (qualitativas). Estas variáveis aparecem no modelo exatamente da mesma forma em que aparecem na base de dados.
- **Modelos são precisos:** os modelos obtidos por *Data Mining* são validados por técnicas de estatística. Desta forma, as predições feitas por modelos são precisas.

3 Aplicação Prática

Os dados, utilizados nesta aplicação prática, foram coletados da base de dados de compra de TI da empresa já apresentada no subtítulo Cenário do Capítulo 1 deste trabalho que, para facilitar, chamaremos de "Empresa Alfa".
Inicialmente, descreveremos o Contexto desta aplicação prática com uma pequena consideração sobre Gestão de Compras e uma breve exposição sobre o Modelo de Gestão de Compras adotada na "Empresa Alfa" para, em seguida, apresentar o estudo realizado, desde o modelo utilizado até a análise dos resultados obtidos após a aplicação do *Data Mining*.

3.1 Contexto

3.1.1 Gestão de Compras

Sabe-se que o processo de compra no mercado dos negócios é diferente do processo de compra de um consumidor comum.
Tanto Kotler e Armostrong (1993) como Semenik e Bamossy (1996) explicam que:

- **A estrutura do mercado e demanda -** o mercado organizacional é constituído por um número menor de compradores, porém em mais larga escala do que no mercado consumidor. É mais concentrado do ponto de vista geográfico. A demanda desse mercado, em sua grande maioria, denominada derivada, pois geralmente as empresas demandam bens de consumo ou insumos. Porém, certos mercados organizacionais têm demanda rígida ou inelástica, assim entendida como aquela que não é afetada por alterações de preços. A demanda nesse mercado também pode flutuar em função das condições de mercado e do tipo de categoria do bem envolvido (investimentos em bens de capital são sensíveis às condições econômicas; especulações em estoques diante de expectativas otimistas ou pessimistas).
- **A natureza de compra -** nas compras organizacionais, geralmente, há mais profissionalismo, pois são executadas por pessoas treinadas na tarefa de comprar. Dependendo do grau de complexidade da compra, a operação pode envolver muitas pessoas de diferentes níveis de responsabilidade;
- **Tipos de decisões e o processo de decisão -** os compradores organizacionais normalmente se deparam com situações de compra complexas, que envolvem grandes somas de dinheiro, detalhes técnicos e econômicos, especificações detalhadas, bem como necessitam de grande interação entre as pessoas e os níveis hierárquicos.

Churchill Jr. E Peter (2000) consolidaram esses pontos conforme a tabela 3 apresentada a seguir:

CARACTERÍSTICAS	CONSUMIDORES	COMPRADORES ORGANIZACIONAIS
Número de compradores	Muitos	Poucos
Tamanho das compras (quantidade e valor unitário)	Pequeno	Grande
Interdependência entre comprador e vendedor	Fraca	Forte
Critérios de Decisão	Racionais e Emocionais	Racionais
Número de pessoas envolvidas nas decisões de compra	Poucas	Muitas

Tabela 3: Diferenças entre consumidores e Compradores organizacionais

Quanto ao Planejamento das Compras, Parente (2000) afirma que existem três diferentes práticas a adotar:

- **De cima para baixo:** quando a gerência estabelece um montante em unidade monetária para as compras de toda a empresa. Nesse caso, os gestores de compras distribuem esse montante entre as várias categorias;
- **De baixo para cima:** quando os gestores de compra fazem as estimativas no nível do produto, passando pela categoria, pelos departamentos, até consolidar no nível de toda a empresa;
- **Interativa:** quando são fixados orçamentos financeiros de compras e os gestores de compras seguem a abordagem "de cima para baixo", com revisões periódicas feitas pelas gerências, com a finalidade de assegurar qua as metas financeiras, de estratégia de marketing e de abastecimento são cumpridas.

Várias outras faces da Gestão de compra poderiam ser abordadas, mas compreendemos ser suficiente para o entendimento desse estudo os aspectos abordados.

3.1.2 Modelos de Gestão de Compras da "Empresa Alfa"

Os processos para aquisição de pequenos bens, na "Empresa Alfa", seguem a seguinte classificação:

- Aquisição de bens considerados como Bem Patrimonial∗2 e material de consumo até o limite de R$ 50.000,00
- Bens com valores abaixo de R$ 326,62, independentemente das suas características, são considerados material de consumo;
- Aquisições de bens considerados Bem Patrimonial entre os valores de R$ 326,62 a R$ 1.306,48 são considerados pela "Empresa Alfa" como bens semi-permanentes recebendo identificação de Bem Patrimonial Único;

As demais Aquisições de Bens serão realizadas por meio dos Órgãos Centralizadores de Compras de acordo com os parâmetros definidos a seguir:

- Aquisição de Bem de Investimento
  - Valores menores do que R$ 500.000,00 (quinhentos mil reais) deve-se encaminhar a solicitação de compra para o Órgão Centralizado de Compra "A".
  - Valores maiores ou iguais a R$ 500.000,00 deve-se encaminhar a solicitação de compra para o Órgão Descentralizado de Compra de TI (C), que conduzirá o processo de aquisição junto ao Órgão Centralizado de Compra "B".
- Aquisição de Bem de Não Investimento
  - Valores menores ou iguais a R$3.000.000,00 (três milhões de reais) deve-se encaminhar a solicitação de compra para o Órgão Centralizado de Compra "A".
  - Valores maiores do que R$ 3.000.000,00 (três milhões de reais) deve-se encaminhar a solicitação de compra para Órgão Descentralizado de Compra de TI (C), que conduzirá o processo de aquisição junto ao Órgão Centralizado de Compra "B".

A Figura 11, abaixo, representa o fluxo físico do encaminhamento das solicitações das aquisições de bens, a realização das aquisições e respectivas áreas responsáveis.

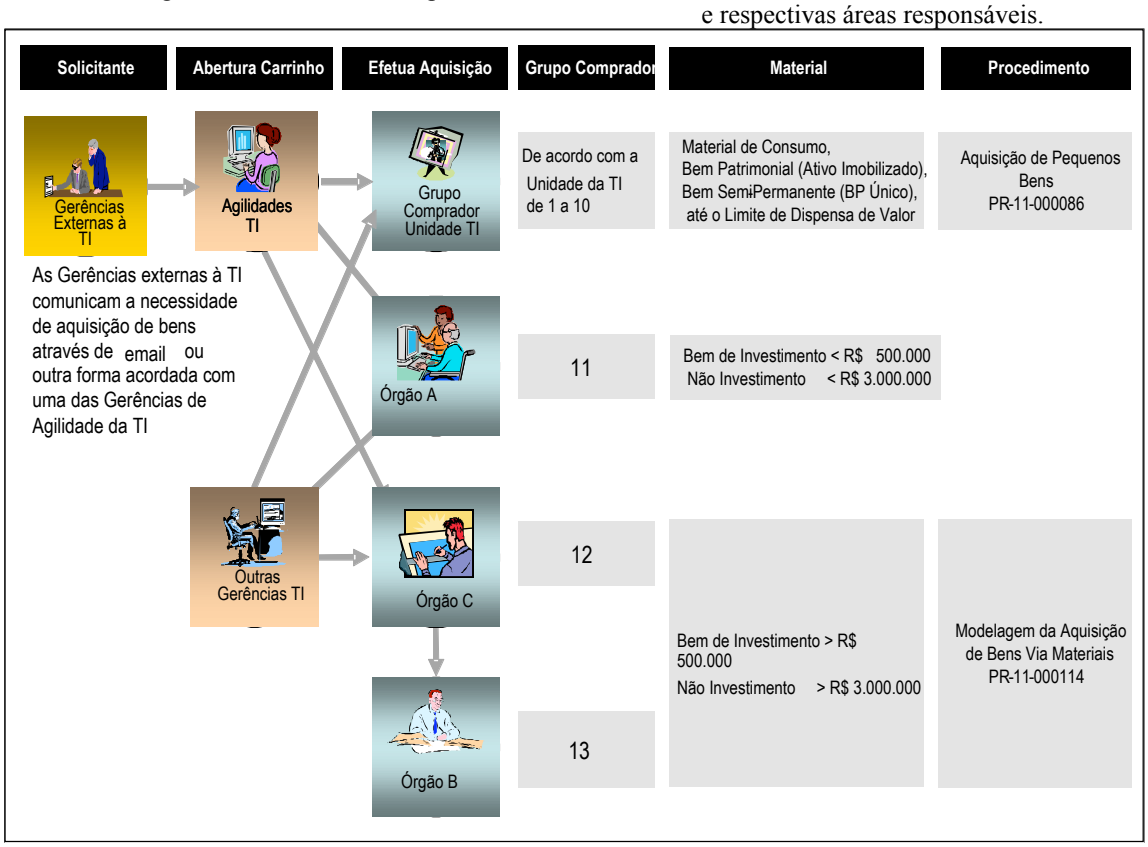

Figura 11: Encaminhamento das solicitações de compra

3.2 METODOLOGIA

A Figura 12, apresentada em seguida, resume o modelo que foi utilizado desde a fonte dos dados primários, passando pelas as etapas de preparação e mineração de dados, até a descoberta de conhecimento como produto final.

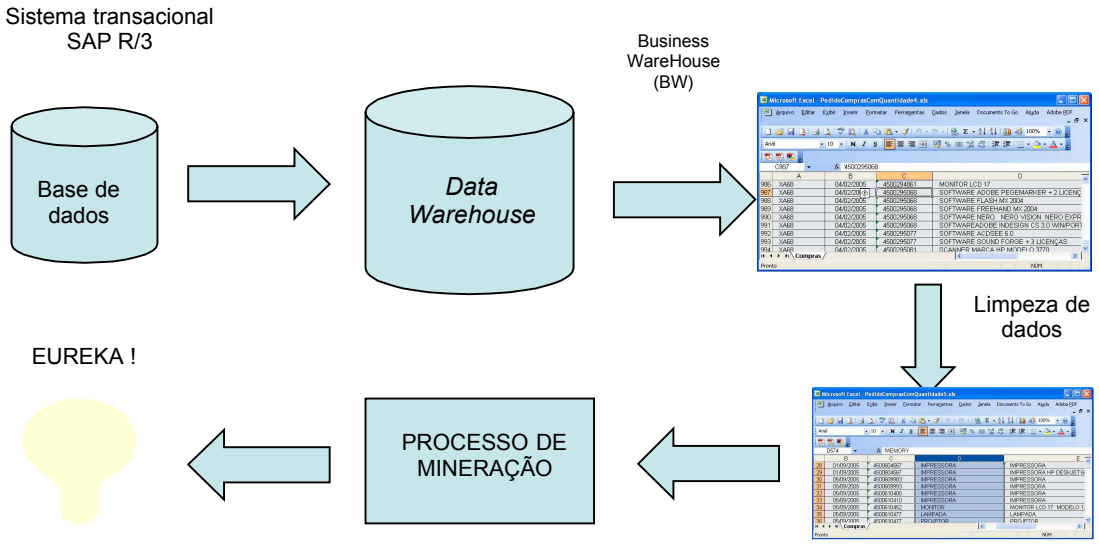

Figura 12: Diagrama do modelo utilizado para descoberta de conhecimento

Inicialmente os dados gerados diariamente pelo sistema transacional SAP R/3 são copiados, periodicamente, para outras bases de dados que contêm os registros históricos das compras efetuadas (*Data Warehouse*). Estes dados foram coletados utilizando-se a ferramenta da SAP denominada BW - *Business WareHouse*) – que armazena os dados de forma estruturada, facilitando a consulta e a análise, agregando valor para a tomada de decisões. O resultado desta etapa foi uma planilha do Microsoft Excel com todos e atributos e valores que estavam disponíveis na base de dados originais.

Posteriormente, foram feitas de forma manual, a limpeza, a codificação e o enriquecimento dos dados (detalhados no subtítulo seguinte), através da eliminação de linhas e colunas e da criação de novos atributos. Esta nova planilha obtida foi, então, convertida para o Microsoft Access de forma que possa ser lida pelo *software* de mineração de dados *WizRule*.

Finalmente, os dados foram processados pelo *software* e as regras de associação foram geradas em formatos de relatórios, propiciando assim a descoberta de conhecimento.

3.2.1 A Preparação dos dados

Esta etapa foi de fundamental importância para que o processo de mineração de dados pudesse gerar as regras de associação. Mais de 50% do tempo dedicado à pesquisa foram gastos nesta etapa preparatória.

- **Eliminação de itens, campos (colunas) e instâncias (linhas) desnecessárias para a análise:** Nessa etapa foram eliminados alguns dados, da planilha gerada pela *query* do BW, a fim de reduzir a quantidade e melhorar a qualidade de processamento. Foram excluídos todos os itens que não contemplavam compras de produtos de TI, as colunas que continham dados considerados não necessários a qualquer tipo de análise e as linhas que apresentavam inconsistências .
- **Preenchimento de campos em branco:** Alguns campos, ao migrarem do *Data Warehouse*, apareceram na planilha com valores zerados fazendo-se necessária a análise caso a caso com a respectiva ação adequada. Ou eliminou-se o item, ou atribuíram-se valores idênticos aos demais itens idênticos.
- **Codificação e padronização da descrição dos itens:** Para garantir a imparcialidade nas análises das associações apresentadas após a aplicação das regras de associação, foram elaboradas duas tabelas codificando os compradores e os fornecedores substituindo-se, em seguida, na planilha os nomes dos compradores e das empresas por suas respectivas codificações.
- **Enriquecimento dos dados:** Na primeira fase de enriquecimento de dados foi percebida a diversidade de formas de descrições existentes na base de dados para um mesmo produto, o que poderia não garantir resultados significativos após a aplicação das regras pelo *WizRule*. Para a solução dessa questão foi feita a inserção de uma coluna antes da coluna de descrição dos produtos, com a denominação "Tipo do Produto". Essa coluna foi posteriormente alimentada, manualmente, com a primeira palavra da Descrição do Produto (ex. *Notebook*) mantendo-se a descrição detalhada apenas na coluna Descrição de Produtos (ex. *NOTEBOOK HP COMPAQ NX5000* - TECLADO ABNT).

3.2.2 *Software* utilizado

O software utilizado para análise da base dados foi a versão 4.05 demo do *WizRule*. Este software é de fácil utilização e configuração sendo exigido muito pouco tempo de aprendizagem por parte do usuário para poder utilizá-lo de forma efetiva.

Possui alguns parâmetros de configuração relacionados ao grau de confiabilidade das regras permitindo fazer os ajustes necessários de acordo com cada caso analisado.

A interface padrão *"for Windows"* do produto e a capacidade de leitura arquivos de entrada gerados por outras ferramentas, em vários formatos possíveis (arquivos do *dbase/foxbase, access* e texto), tornam o produto ainda mais versátil.

Como já citado no Capítulo 1 deste trabalho, a versão demo deste *software* possui uma limitação de uso relacionada ao número máximo de linhas (1 000) que podem ser analisadas.

No entanto, esta versão foi escolhida por não haver necessidade do uso da versão *full*, já que após a etapa de pré-processamento, a base de dados final continha menos de 1 000 linhas.

O resultado final da análise é apresentado em um relatório que mostra, em tela, as regras geradas e os possíveis desvios (exceções às regras) dos dados.

As regras geradas são numeradas e se apresentam como segue:

1) *If* **Fornecedor** *is* **F45**
Then
Tipo do produto *is* ***SOFTWARE***
Rule's probability: **1,000**
The rule exists in **30** *records.*
Significance Level: Error probability is almost 0

Esta regra indica que quando o fornecedor é o F45 (código), o produto vendido é um *SOFTWARE*. A regra tem probabilidade igual a 1, existe em 30 registros e é baixa a probabilidade de haver exceções a esta regra.

A figura 13, exibida a seguir, mostra a tela principal do *WizRule Demo 4.05* já com a base de dados de compras, objeto deste estudo, carregada.

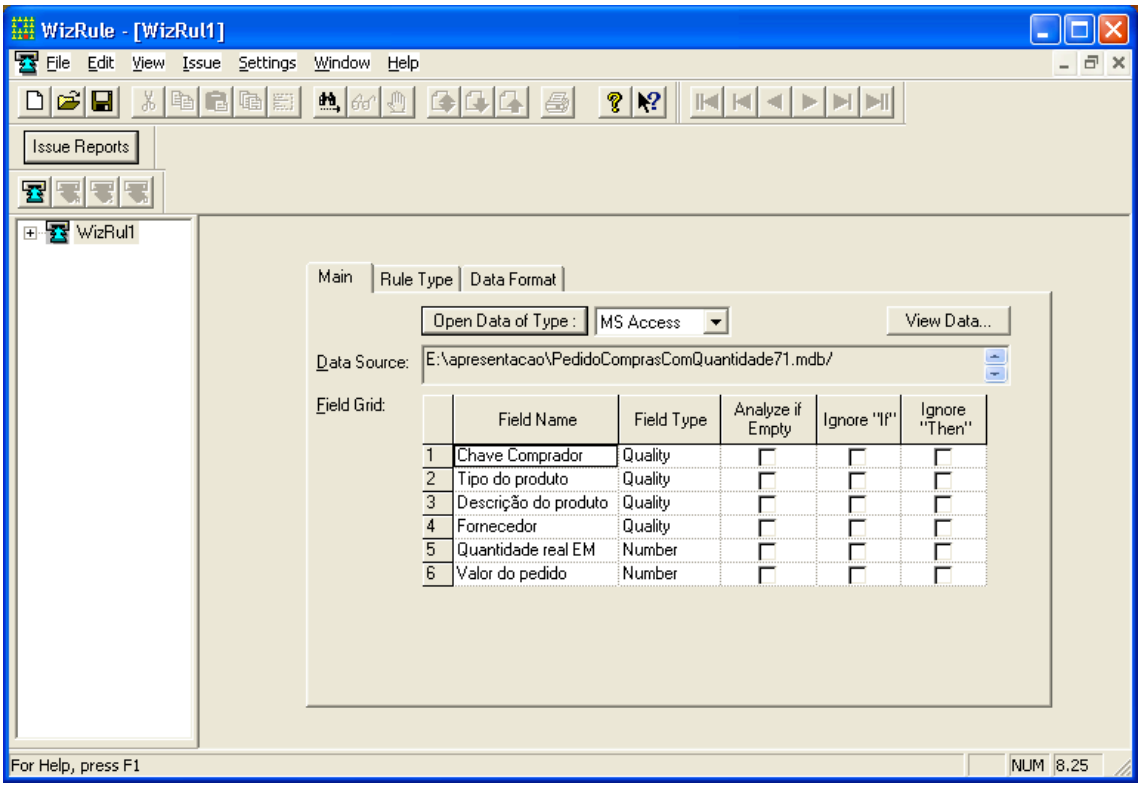

Figura 13: Tela inicial do *WizRule* com a base de dados importada do *Microsoft Access*

3.3 ANÁLISE DOS RESULTADOS

Foram analisados 935 registros de compras pelo *software* e geradas 52 regras de associação sendo configuradas de forma que somente regras com 90% de probabilidade e com pelo menos 20 ocorrências fossem consideradas.

A Tabela 4, apresentada na página seguinte, exibe a consolidação das primeiras 15 regras geradas.

REGRA	SE	E	ENTÃO	REGISTROS	DESVIOS
1	F03		C1	83	5
2	F45		SOFTWARE	30	0
3	F30		0-324 (VP)	50	0
4	F06		C1	46	2
5	F04		C4	85	4
6	C4	F09	0-329	49	2
7	F09		0-329	53	4
8	F11	1-5 (Q)	0-88 (VP)	46	2
9	F01		C1	26	0
10	MONITOR	F03	C1	25	0
11	F09	0-329	C4	49	3
12	MEMORIA	1-5 (Q)	C4	39	1
13	IMPRESSORA		1-5 (Q)	245	9
14	F45		C4	30	0
15	MEMORIA	0-329	C4	32	2

Tabela 4: Consolidação das 15 primeiras regras geradas pelo *WizRule*

Após a análise da tabela com todas as regras geradas e com o cruzamento de informações entre as regras, chegou-se as seguintes conclusões:

- Apesar do grande número de fornecedores que já forneceram pelo menos 1 item, a maior parte das compras (em registros) está concentrada em poucos;
- Esta concentração também ocorre na relação comprador X fornecedor e na relação fornecedor X item;
- Alguns fornecedores fornecem apenas um tipo de produto;
- Produtos tais como impressoras e scanners (entre outros) são comprados, na maioria dos casos, por unidade;

Ainda como resultado da análise foi possível estabelecer uma faixa de valores para uma grande quantidade dos itens de TI que são comprados. Estas faixas poderão ser utilizadas como parâmetros na análise de futuras compras.

4 CONCLUSÃO

A utilização de técnicas de mineração de dados mostrou-se útil para descoberta de conhecimento que estava oculto nas bases de dados de compra analisadas. Embora algumas das assertivas pudessem ser descobertas por outros meios (estatísticas dos atributos, por exemplo), a consolidação das regras e o cruzamento das mesmas permite uma melhor especificidade do caso avaliado.

O objetivo da pesquisa foi cumprido e o software utilizado mostrou-se adequado para a base de dados que foi utilizada. Este software apresentou uma boa performance e facilidade de uso, fundamentais para que esta pesquisa fosse realizada dentro do prazo planejado.

Quanto ao objetivo geral deste trabalho, de buscar contribuir com uma solução de otimização de tarefas capazes de auxiliar a tomada de decisão na gestão de compras em uma empresa de grande porte, aplicando a técnica de *Data Mining,* também foi atingido, apesar da utilização de uma pequena parte da base de dados de compra.

Após a aplicação da ferramenta de análise *Data Mining*, foram obtidas informações muito valiosas e que, certamente, demonstram que a utilização do *Data Mining* em base de dados de compra não só é possível como recomendada para auxílio na tomada de decisão de futuras compras.

Ficaram evidentes os benefícios oriundos das informações geradas. Esses resultados poderão ser aproveitados, na gestão das compras, nas ações gerenciais que visem à concentração das compras de alguns itens alcançando-se ganhos de escala, a implantação de medidas para minimizar os vícios identificados, a ampliação da escolha dos fornecedores evitando-se monopólios prejudiciais e as demais ações que contribuam para o alcance das metas do planejamento estratégico do negócio.

Estudos futuros poderão contemplar outras tarefas típicas do processo de mineração de dados tais como a classificação e a análise de agrupamentos. Estas tarefas poderão completar este estudo sobremaneira.

Através da classificação, poderia se buscar a estratificação dos itens comprados em uma graduação baseada no valor de compra. A decisão de comprar em lotes pode ser ajustada de acordo, também, com os valores praticados pelo mercado e pelos compradores.

Por meio da análise de agrupamentos, seria possível obter conhecimento quanto aos grupos de itens, identificando itens e fornecedores que poderiam ser objeto de contratos de fornecimento ao invés de recorrentes compras pontuais.

Recomendamos, no entanto, para os trabalhos futuros, a aplicação do *Data Mining* em toda a base de compras e da versão *full* do *software*, para que os resultados gerados sejam os mais completos possíveis e enriqueçam o processo decisório de toda a área de compras da empresa.

REFERÊNCIAS BIBLIOGRÁFICAS

[1].ADRIAANS, P.; ZANTINGE, D. **Data Mining**. England: Addison Wesley Longman, 1996.

[2].ARCHER, Earnest R. **How to Make a Business Decision: An Analysis of Theory and Practice**, Management Review, AMACOM, vol. 69, no. 2, fev. 1980, p.54-61

[3].__ O mito da motivação. In: **Psicodinâmica da vida organizacional.** São Paulo, Atlas, 1997.

[4].BISPO, C.A.F. **Uma Análise da Nova Geração de Sistemas de Apoio à Decisão.** Dissertação (Mestrado em Engenharia de Produção) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 1998.

[5].BISPO, C.A.F.; CAZARINI, E.W. **A evolução do processo decisório.** (CD-ROM) In: ENCONTRO NACIONAL DA ENGENHARIA DA PRODUÇÃO, 18., / CONGRESSO INTERNACIONAL DE ENGENHARIA INDUSTRIA,4., Niterói, 1998. *Anais*. Niterói, TEP-UFF, artigo 94.doc.

[6].BRACHMAN, R. and ANAND, T., **The process of Knowledge Discovery in Databases: A Human-Centered Approach, Advances in Knowledge Discovery and Data Mining**, AAAI/MIT Press,1996.

[7].CAMPOS, V. F. **TQC: controle da qualidade total (no estilo japonês)**. Belo Horizonte, Fundação Christiano Ottoni / Escola de Engenharia da UFMG. 1992.

[8].CHEN, M. S.; HAN, J.; YU, P. S. **Data Mining: an overview from database perspective.** IEEE Trans. on Knowledge and Data Engineering, New York, v. 8, n.6, p. 866-883, 1996.

[9].CODD, E. F.; CODD, S. B.; SALLEY, C.T. **Providing OLAP (on-Line Analytical Processing) to users-analysts: An IT mandate.** http://dev.hyperion.com/resource_library/white_papers/providing_olap_to_user_analysts.pdf Acesso em junho/2006.

[10].COSTA, P.W.A. **Como surgiram os data warehouses?** Computerworld, 03 nov., p.16. 1997.

[11].DIN - Departamento de Informática - UEM - Universidade Estadual de Maringá. GSI - Grupo de Sistemas Inteligentes - **Mineração de Dados**, 1998. Disponível em: http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html Acesso em: julho/ 2006.

[12].DRUCKER, P. F.. **Administrando em Tempos de Grandes Mudanças.** São Paulo: Pioneira, 1995.

[13].FAYYAD, Usama M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R.. **Advances in knowledge Discovery & Data Mining.** AAAI/MIT, 1996.

[14].FISHER, L. M. **Along the infobahn: data warehouses.** Strategy & Business, Third Quarter, 1996. Disponível em: http://www.strategy-business.com/press/article/17747?pg=0. Acesso em agosto/2006.

[15].FREITAS, A. A. Uma Introdução a Data Mining**. Informática Brasileira em Análise.** CESAR - Centro de Estudos e Sistemas Avançados do Recife. Ano II, n. 32, mai./jun. 2000.

[16].GATES, B. **A Estrada do Futuro**. São Paulo: Cia. Das Letras, 1995.

[17].GIL, A.C. **Como elaborar projetos de pesquisa**. 3.ed. São Paulo: Atlas, 1996. 159p.

[18]._____ **Métodos e técnicas de pesquisa social.** 5. ed. São Paulo: Atlas, 1999. 206p.

[19].HACKNEY, Douglas. **Data Warehouse Delivery: Who are You?** Part I. DM Review Magazine, v.8, n. 2, 1998.

[20].HALL, R. **Organizações: Estrutura, Processos e Resultados**. Rio de Janeiro. Ed. Prentice-Hall, 8ª edição. 2004.

[21].HAMMER, M. **Reengenharia: Revolucionando a Empresa em função dos Clientes, da Concorrência e das Grandes Mudanças da Gerência**. Rio de Janeiro. Campos, 1994.

[22].HUBER, J. et al. **Integrating Data Warehouse versus Build a Federated Data Warehouse: A Comparision**. In Proceedings of the 3rd Internacional Conf. on Data Warehousing and Knowledge Discovery, Munich, Germany, 2001. Disponível em: http://www.scch.at/index.jsp?menu=publications&link=/publications/publication.jsp&id=120 Acesso em: junho/2006.

[23].HYPERION. **The role of the OLAP server in a data warehousing solution.** http://dev.hyperion.com/resource_library/white_papers/olap_in_a_data_warehousing_solution.pdf Acesso em: junho/2006.

[24].INMON, W.H., **Building the Data Warehouse**, Four Edition, Indianapolis, Wiley Publishing, Inc, 2005.

[25].KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit - Guia Completo para Modelagem Dimensional**. Tradução da 2. ed. Rio de Janeiro: Campus, 2002

[26].KIMBALL, R., REEVES, L., ROSS, M. and THORNTHWAITE, W., **The data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouse**. John Wiley & Sons, New York, 1998.

[27].MANNILA, H. **Data mining: machine learning, statistics, and databases.** In: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANANGEMENT, Stockholm, 1996.

[28].NIMER, F.; SPANDRI, L.C. **Data Mining.** Revista Developers. Fev./1998, p.32.

[29].PEARSON, J. M., SHIM, J. P. **Na empirical investigation into DSS strutures and environments**. Decision Suport Systems, n. 13, p.141-158. 1995.

[30].PEREIRA, M.J.L.B.;FONSECA, J.G.M. **Faces da decisão: As mudanças de paradigmas e o poder da decisão.** São Paulo. Makron Books, 1997.

[31].POSSAS, B. A. V.; CARVALHO, M. L. B. de; REZENDE, R. S. F.; MEIRA JR., W. **Data Mining: Técnicas para Exploração de Dados.** Universidade Federal de Minas Gerais, 1998.

[32].POWER, D., **Decision Support Systems: Concepts and Resources of Managers** (text book binding). Quorum Books, 2002.

[33].RODRIGUES, A. M. **Escavando Dados no Varejo.** Centro de Estudos em Logística - COPPEAD - Universidade Federal do Rio de Janeiro, 2000.

[34].SILVA, E. M. **Avaliação do Estado da Arte e Produtos Data Mining.** UCB - Universidade Católica de Brasília, 2000.

[5].SIMON, Alan R,. **Strategic Database Technology: Management for the year 2000**. Morgan Kaufmann Publishers, Inc, 1995.

[36].SIMON, Herbert A. **Decision Making and problem solving.** National Academy Press. Washington, 1986.

[37]._____ **Comportamento administrativo: estudo dos processos decisórios nas organizações administrativas.** Fundação Getúlio Vargas. Rio de Janeiro, 1979.

[38].SINGH, Harry S. **Data Warehouse: Conceitos, Tecnologias, Implementação e Gerenciamento**. São Paulo: Makron Books, 2001.

[39].WEISS, S.M. and KULIKOWSKI, C.A., **Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and expert Systems** (Machine Learning Series), Hardcover, 1991.

[40].WELDON, J. L. **A career in data modeling**. Byte, jun. 1997. Disponível em: http://www.byte.com/art/9706/sec7/art3.htm Acesso em agosto/2006.

[41].WIZSOFT. **WizRule**. Disponível em: http://www.wizsoft.com/default.asp?win=8&winsub=8 Acesso em agosto/2006.

[42].THEARLING, K., BERSON, A., SMITH, S.. **Building Data Mining Appllications for CRM.** McGraw Hill, 2000.

---

* 1 Ambiente Transacional – ambiente onde são executadas, no dia-a dia, todas as transações primárias da empresa. (ex.: comprar, vender, baixar estoque, controlar manutenção, etc.).

* [2] Bem Patrimonial são bens contabilizados no ativo da "Empresa Alfa".