

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE CAMPINAS**

**Robinson Luz**

**MESTRADO EM CIÊNCIA DA INFORMAÇÃO**

**Integração da Lógica nebulosa à recuperação de informação em  
Data warehouse**

Campinas - S.P.  
2005

**Robinson Luz**

**Integração da Lógica nebulosa à recuperação de informação em  
Data warehouse**

Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciência da Informação da Pontifícia Universidade Católica de Campinas.

Área de concentração: Administração da Informação.

Linha de Pesquisa: Produção e disseminação da informação.

Orientadora: Profa. Dra. Nair Y. Kobashi

Co-orientador: Dr. Edberto Fereda

Campinas

2005

**ROBINSON LUZ**

**Integração da Lógica nebulosa à recuperação de informação em  
Data warehouse**

Dissertação apresentada como requisito parcial para  
obtenção do título de Mestre, ao Programa de Pós-  
Graduação em Ciência da Informação da Pontifícia  
Universidade Católica de Campinas.

Área de concentração: Administração da Informação.  
Linha de Pesquisa: Produção e disseminação da  
informação.

Campinas, 28 de junho de 2005.

Banca Examinadora

Profa. Dra. Nair Yumiko Kobashi (orientadora)  
PUC – Campinas

Prof. Dr. Paulo Celso da`Silva  
Universidade de Sorocaba

Prof. Dr. José Oscar Fontanini de Carvalho  
PUC- Campinas

**“Integração da Lógica Nebulosa à recuperação de Informação em Data warehouse”**

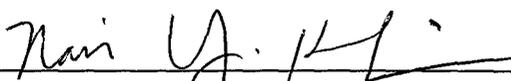
Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós Graduação em Ciência da Informação da Pontifícia Universidade Católica de Campinas.

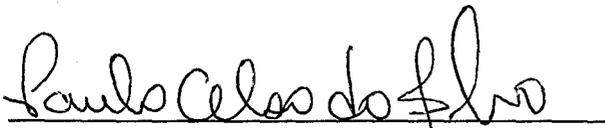
Área de Concentração: Administração da Informação.

Linha de Pesquisa: Produção e Disseminação da Informação.

Campinas, 28 de junho de 2005

**Banca Examinadora:**

  
\_\_\_\_\_  
**Profa. Dra. Nair Yumiko Kobashi (orientadora)**  
PUC - Campinas

  
\_\_\_\_\_  
**Prof. Dr. Paulo Celso da Silva**  
Universidade de Sorocaba

  
\_\_\_\_\_  
**Prof. Dr. José Oscar Fontanini de Carvalho**  
PUC - Campinas

## **Agradecimentos:**

À profa. Dra. Nair Y. KObashi por compartilhar comigo seu conhecimento, pela valiosa orientação do trabalho, segura, porém com suficiente liberdade para que eu pudesse, por mim mesmo, encontrar os caminhos que me aproximassem da realidade.

Ao Prof. Dr. Edberto Ferneda pela co-orientação e sugestões.

Meus agradecimentos à Roseli, minha esposa, pela paciência e pelo apoio dedicado ao longo desses 33 anos.

Aos Professores Maria de Fátima Tálamo, Paulo Jannuzzi, Raimundo Nonato Macedo dos Santos, Paulo de Martino Jannuzzi pelos ensinamentos recebidos, que foram fundamentais para essa jornada.

Aos Professores Antonio César Munari e Maria Angélica Cardieri pelo apoio na área de Banco de Dados.

Aos companheiros de curso Alexandre Camargo Castro, Amauri Pereira Ramos, Antonio Carlos da S. Junior e Daniel Pace cujas causas abraçávamos mutuamente.

Ao aluno Pedro Henrique Araujo pelo apoio na programação dos algoritmos referentes aos exemplos citados.

## **Integração da Lógica Nebulosa à recuperação de informação em Data warehouse.**

**Autor:** Robinson Luz

**Orientadora:** Profa. Dra. Nair Y. Kobashi

**Co-orientador:** Dr. Edberto Ferneda

### **Resumo**

A presente pesquisa tem como objeto de estudo a integração da “lógica nebulosa” às tecnologias de “data warehouse”. Objetiva, especificamente, propor, com base nas teorias e práticas da Ciência da informação, um modelo conceitual alternativo de organização e recuperação de informação. Para o desenvolvimento do modelo são descritos diversos tipos de bancos de dados e seu histórico, desde sua criação até os armazéns de dados chamados “data warehouse”. Quanto aos “data warehouse”, são expostos sua arquitetura, dimensões e volatilidade, para fundamentar e justificar a escolha deste tipo de banco de dados para realizar estudo comparativo de recuperação de dados utilizando-se a “lógica booleana” e a “lógica nebulosa”. São apresentados também conceitos básicos de “lógica nebulosa”, suas características, aplicações e sua importância para auxílio à tomada de decisão. São apresentadas, ainda, as vantagens do uso de métodos de tratamento nebuloso para lidar com sistemas que interagem com seres humanos por meio da simulação de raciocínios. São apresentados testes que exemplificam o grau de eficácia da recuperação de dados em “data warehouse” por meio da “lógica booleana” e por meio da “lógica nebulosa”. Como resultado do estudo, apresenta-se uma proposta de parâmetros que serão considerados na integração da “lógica difusa” aos “data warehouses”.

Palavras-chaves: Lógica Nebulosa, Data Warehouse e Recuperação de Informação.

# **Fuzzy Logic integration to information retrieval in Data Warehouse.**

**Author:** Robinson Luz

**Oriented by:** Profa. Dra. Nair Y. Kobashi

**Co-oriented by:** Dr. Edberto Ferneda

## **Abstract**

This research has as its objective to study the integration of fuzzy logic to “data warehouse” technologies, having its basis on Information Science theories and practices, it proposes an alternative conceptual model of organization and information retrieval. To develop the model it is described many types of data base in its history file, since its creation up to the data storage known as “data warehouse”. Regarding the “data warehouse”, it is showed its architecture, dimension and volatility to show its basis and justify the choice of this kind of database for rendering a comparative study of data recovery using “Boolean logic” and “fuzzy logic”. Basic concepts of “fuzzy logic”, its characteristics, uses and importance to support the decision taking are presented. It is still presented the advantages on using the methods of fuzzy treatment to deal with system that interact with human beings through simulation and thinking. It is presented tests which exemplify the effectiveness degree of data retrieval in “data warehouse” through Boolean logic and also through “fuzzy logic”. As a result of this study, it is shown a parameter proposal that will be taken in consideration when integrating the fuzzy logic to “data warehouses”.

Key-words: Fuzzy Logic, Data Warehouse and Information Retrieval

## SUMÁRIO

CAPÍTULO 1 – Introdução.....	1
1.1 Objetivos do trabalho .....	3
1.2 Hipótese de pesquisa.....	4
1.3 Metodologia.....	4
1.4 Organização da dissertação .....	4
CAPÍTULO 2 – A Ciência da Informação e a Recuperação de Informação .....	6
2.1 A Ciência da Informação.....	8
2.1.1 Perspectivas da Ciência da Informação .....	11
2.2 Gestão da Informação.....	13
2.3 Recuperação de Informação .....	15
2.3.1 Sistemas de recuperação de Informação.....	16
2.4 A era da informação .....	16
2.5 A informática e sistemas de informação .....	18
2.6 Sistemas de informação e sistemas de apoio à decisão.....	18
2.7 Dados operacionais e dados informacionais.....	19
2.8 Conclusões do capítulo.....	20
CAPÍTULO 3 – Banco de dados: dos arquivos até os “data warehouse” .....	22
3.1 Termos e Conceitos de organização de arquivos.....	24
3.2 Banco de dados e sistemas de gerenciamento de banco de dados .....	28
3.3 Modelos de Banco de Dados .....	30
3.3.1 Modelos Conceituais.....	30
3.3.2 Modelos Lógicos.....	31
3.3.2.1 Modelo em rede e hierárquico.....	31
3.3.2.2 Modelo Relacional.....	34
3.3.2.3 O modelo banco de dados “data warehouse”.....	39
3.4 Sistemas de gerenciamentos de bancos de dados SQL e segurança dos dados.....	39
3.4.1 Problemas suportados pelos gerenciadores de banco de dados .....	44
3.5 Conclusões do capítulo.....	46
CAPÍTULO 4 – Banco de dados “data warehouse” .....	47
4.1 Sistemas transacionais e o modelo dimensional .....	47
4.2 Abordagens para o desenvolvimento de um “data warehouse”.....	49
4.2.1 Conceitos básicos, definições e importância do “data warehouse” .....	50
4.2.2 Características do “data warehouse” .....	52
4.2.3 Arquitetura genérica do “data warehouse” .....	53
4.2.4 Metadados e sua importância .....	54
4.2.5 Granularidade.....	55
4.2.6 Dimensões.....	56
4.2.7 Tabela de fatos .....	57
4.2.8 Data Mining .....	57
4.3 Modelo de dados para, o “data warehouse” .....	58
4.3.1 Tratando as informações para o “data warehouse” .....	58
4.4 O nível de abrangências da base de dados no “data warehouse”.....	60
4.5 Ciclicidade dos dados.....	62
4.6 Escolha da duração do banco de dados .....	62
4.7 Busca ao “data warehouse” .....	63
4.8 Segurança das informações .....	63

4.9 Conclusões do capítulo.....	64
CAPÍTULO 5 – As lógicas: Booleana e Nebulosa como estratégia de recuperação de informação.....	66
5.1 Introdução.....	66
5.2 Histórico .....	69
5.3 Uma reflexão da teoria da “lógica booleana” e da teoria da “lógica nebulosa”.....	70
5.4 Lógica e Álgebra booleana .....	72
5.4.1 Modelo booleano.....	75
5.5 Conceitos da “lógica nebulosa” .....	76
5.6 Buscas usando conceitos e funções nebulosas.....	80
5.7 “Lógica nebulosa” no contexto da computação natural. ....	82
5.8 Conclusões do capítulo.....	84
CAPÍTULO 6 – Aplicação da “lógica nebulosa” à recuperação de informação em “data warehouse”.....	85
6.1 Porque a escolha do "data warehouse" para se utilizar a “lógica nebulosa” para tomada de decisão. ....	91
6.1.1 Processos que serão levados em consideração da carga dos dados no “data warehouse”. ....	93
6.2 Comparação de Resultados entre as lógicas booleana e nebulosa.....	96
6.3 Conclusões das análises entre “lógica nebulosa” e “lógica booleana” .....	99
CAPÍTULO 7 – Conclusões e considerações finais.....	100
7.1 Perspectivas .....	102
Referências: .....	105
Apêndice A.....	112
Apêndice B.....	115

## CAPÍTULO 1 – Introdução

A essência do processamento de recuperação de informação é “Dê-me o que eu digo que quero, então eu te direi o que eu realmente quero” Inmon (2004).

Os tomadores de decisão em geral percebem coisas, tais como mudanças no negócio, clientes, índices de vendas, procuram informações em bancos de dados através dos dados que explicam o que eles observam e se esforçam para que essas mudanças ou fenômenos sejam explicados e entendam a oportunidade quando esta é apresentada, e não se sentem satisfeitos até que as suas observações transformem-se em oportunidades de negócio. Inmon (2004).

A recuperação de informação em bancos de dados é utilizada em larga escala dentro de empresas; espera-se, no entanto, que os resultados obtidos através de buscas respondam efetivamente à necessidade de informação.

As buscas em banco de dados, quando realizadas utilizando-se lógica convencional, aqui definida como a “lógica booleana”, tendem a fornecer resultados que, muitas vezes, não satisfazem as necessidades do usuário. Isso acontece porque este sistema de busca se baseia em operações clássicas da Álgebra Booleana para decidir a respeito da pertinência de um elemento a um determinado conjunto. Como só existem duas possibilidades, pertence ou não, percebe-se que esta aproximação está sujeita a perdas de informação que podem vir a comprometer a confiabilidade do resultado final, visto que no mundo real, na maioria das vezes, elementos podem pertencer parcialmente a um conjunto e em nossa comunicação cotidiana usamos muitas palavras e sentenças com significado não preciso. Isto ocorre porque, tanto quem fala como quem ouve, não necessita somente de informações precisas e está acostumado a lidar com tipos de imprecisão, tais como os conceitos de: muito, pouco, pequeno, alto, bom, quente, frio, etc. Por exemplo, na área de vendas, alguém que nega uma venda por examinar o limite de crédito de um cliente por valores exatos, poderá deixar de vender uma soma significativa, por algum valor não significativo que extrapole a mais o seu limite de crédito.

Parece claro que, uma vez lançando mão do uso de “lógica booleana”, qualquer busca estará sujeita à perda de informação. Pode-se então, pensar que isto ocorre porque “subestimam-se” dados, isto é, desprezam-se informações que poderiam ser relevantes pelo fato de elas não obedecerem ao nosso padrão de classificação por uma diferença nem sempre significativa.

Com a evolução da Tecnologia de Informação e o crescimento do uso de computadores interconectados, praticamente todas as empresas utilizam sistemas informatizados para realizar seu controle administrativo, como controle de folha de pagamento, contas a pagar e receber, produção, etc o que, com o passar do tempo, gera uma enorme quantidade de dados armazenados em banco de dados relacionados aos negócios, mas não relacionados entre si.

Dados armazenados são um recurso, mas, de modo geral, em seu estado original raramente servem para tomada de decisão. Os sistemas convencionais de informática não são projetados para gerar e armazenar informações estratégicas, o que torna os dados vagos e sem valor para o apoio ao processo de tomada de decisão nas empresas. Um novo desafio é imposto então aos sistemas de informação; eles devem possuir uma interface amigável, que é definida por Pressman (2002) como: “Todo o projeto de interface com o usuário que crie um meio efetivo de comunicação entre o ser humano e o computador” e por Bispo (1995) como: “Navegar com extrema facilidade entre os diversos níveis de agregação dos dados e analisá-los, através de qualquer combinação possível entre os mesmos, possibilitando as mais variadas visões possíveis de negócio e apresentar aos usuários um sistema de apoio à decisão que possibilite recuperar, dados e informação, de forma fácil e eficiente, na hora certa e na forma correta para apoio ao processo decisório, de modo a melhorar a qualidade das decisões”.

Para Jannuzzi (2004) não obstante a importância atribuída à informação, afirmada por meio de grandes estoques e oferta de informações, uma das questões que hoje se levanta diz respeito à qualidade do fluxo da informação. Sob o ponto de vista do usuário, pois, apesar dos esforços de produção e disseminação, ainda assim, é comum haver por parte dos usuários, reclamações sobre a insuficiência da informação efetivamente utilizada. Perante esse quadro, as dúvidas recaem sobre o alcance dessas informações, ou seja, se a informação que é oferecida atende à demanda no formato e conteúdo apropriados ao usuário para o qual se destina.

Acoff (1967) faz considerações sobre os projetos da maioria dos sistemas de informação, que em geral apontam a deficiência crítica com que trabalham muitos tomadores de decisão. É a falta de uma boa quantidade de informações que deveriam possuir e a recusa em aceitar que esse seja o mais importante problema por eles enfrentado em termos de informação. Afirma ainda, que tomadores de decisão necessitam das informações que desejam, pois muitos dos indivíduos não projetam sistemas de informações baseando-se nas informações que tomadores de decisão gostariam de ter. Isso baseia-se na hipótese de que eles

sabem que informações exigem e efetivamente desejam, para isso teriam que estar cientes de cada tipo de decisão que deve se tomada, e na verdade, raramente essas condições são satisfeitas, pois muitos deles têm apenas algumas noções de que tipos de decisões lhes cabe tomar.

Ackoff ainda comenta que se o tomador de decisão receber a informação que ele deseja e exige, isso não significa que ele não terá problema algum para usá-la de forma eficaz, pois isso pode depender de rotinas de programações matemáticas ou caminho crítico. Se sua experiência e seu julgamento tiverem algum valor, ele talvez não se saia mal, mas raramente irá se sair muito bem.

A importância da informação para as organizações é universalmente aceita, e constitui, senão o mais importante, pelo menos um dos recursos cuja gestão e aproveitamento estão diretamente relacionados ao sucesso desejado. A informação também é considerada e utilizada em muitas organizações como um instrumento de gestão. Portanto, a gestão efetiva de uma organização requer a percepção objetiva da informação e ressalta o entendimento atual da informação como recurso valioso para a aquisição do conhecimento, constituído como fator de desenvolvimento social e econômico.

### 1.1 Objetivos do trabalho

Pretende-se neste trabalho estudar um modelo teórico e conceitual que explique, situe e abra caminhos para integração das tecnologias de “data warehouse” e “lógica nebulosa”, no conjunto de teorias e práticas de recuperação da informação, sob o ponto de vista da Ciência da Informação. Para isso será feito um estudo comparativo de recuperação de dados em “data warehouse” utilizando-se a “lógica booleana” e a “lógica nebulosa”; pesquisa de estratégias de integração da “lógica nebulosa” ao “data warehouse” e pesquisa das vantagens do uso de métodos de tratamento nebuloso para lidar com sistemas que interagem com seres humanos, utilizados para a recuperação de informações como extensão do seu próprio processo de raciocínio.

Justifica-se este trabalho por ser a informação elemento fundamental na tomada de decisão no ambiente empresarial. Nesta pesquisa descreveremos as possíveis interações entre o tomador de decisão e suas bases de dados, para melhorar a tomada de decisão.

De fato, a integração das tecnologias “data warehouse” e “lógica nebulosa” poderá ampliar as condições de busca e obtenção de informações por permitir expressar mais adequadamente o raciocínio humano.

Esta pesquisa poderá trazer contribuições teóricas específicas sobre as estratégias de busca de informações em banco de dados não relacional (multidimensional) para apoio à decisão ao comparar os valores dos resultados da “lógica booleana” e os resultados das pesquisas da “lógica nebulosa” utilizando a tecnologia “data warehouse”.

Esta pesquisa também poderá beneficiar pelo menos tres áreas: a área industrial, pela razão de disponibilizarmos recuperação de informação não só contidas na precisão da álgebra booleana, como também valores próximos a estes intervalos e que com critérios de determinação de aproximação poderão ser analisados e aceitos pelo tomador de decisão; a área da Ciência da Informação pela razão de oferecermos novos critérios de busca e recuperação de informação e a área das Tecnologias de Informação pelo fato de unirmos tecnologias de banco de dados não relacional com conceitos de “lógica nebulosa”.

## 1.2 Hipótese de pesquisa

A hipótese que se levanta neste trabalho é que no que se refere ao processo de recuperação de informação, a “lógica nebulosa”, ao contrário da “lógica booleana”, poderá modelar o modo aproximado de raciocínio humano, visando oferecer estratégias computacionais capazes de tomar decisões racionais em um ambiente de incerteza e imprecisão, fornecendo uma resposta aproximada para uma questão baseada em um conhecimento que é inexato, incompleto ou não totalmente confiável.

## 1.3 Metodologia

Para identificar idéias e dados para subsidiar a pesquisa serão, comparadas a “lógica nebulosa” e a “lógica booleana”, por meio de levantamento bibliográfico sobre “lógica nebulosa”, “data warehouse”, recuperação de informação e tomada de decisão para definir os conceitos e identificar os principais benefícios relativo ao tema da pesquisa.

Em seguida será construído um exemplo de aplicação de Recuperação de Informações para tomada de decisão em “data warehouse”, validando a pesquisa.

## 1.4 Organização da dissertação

A dissertação organiza-se em sete capítulos. No capítulo 1, teremos uma introdução onde serão apresentadas a contextualização do trabalho, os objetivos, hipótese de pesquisa e a metodologia.

No capítulo 2 serão abordadas as terminologias básicas de gestão de informação, os conceitos de gestão, recuperação de informação, gerenciamento e fontes de informação sob o ponto de vista de Ciência da Informação e suas perspectivas. Também serão analisados os conceitos básicos de gestão de informação e do conhecimento, sistemas de informação enfatizando o suporte à decisão e conceitos de dados “operacionais” e dados “informativos”.

No capítulo 3 serão tratados os banco de dados, sua definições, história e sua evolução até os “data warehouse” , conceitos, vantagens e desvantagens dos modelos em Rede, Hierárquico, Relacional. Apresentaremos também os conceitos da linguagem de recuperação e manutenção dos dados em Bancos de Dados chamada SQL e também conceitos sobre a segurança dos dados.

O capítulo 4 terá a apresentação das terminologias básicas de "data warehouse" como também conceitos, importância e características do "data warehouse", considerações sobre os metadados e sua importância, conceitos específicos de “data warehouse”, tais como: volatilidade, granularidade, fatos, dimensão, ciclicidade, mineração de dados e também os cuidados na transferências dos dados dos sistema tradicionais para os “data warehouse” com a finalidade de recuperar informação para tomada de decisão.

No capítulo 5 serão apresentados as terminologias básicas de “lógica nebulosa”, os conceitos e suas aplicações. Sua importância no processo de tomada de decisão e o posicionamento da “lógica nebulosa” no contexto da Computação natural. Serão feitas comparações das lógicas booleana e nebulosa na recuperação de informação para mostrar quanto melhorarão os resultados através da lógica nebulosa.

No capítulo 6, será feita uma comparação dos resultados de recuperação de dados através da integração de “data warehouse” com a “lógica nebulosa” e a “lógica booleana”. Características, vantagens e desvantagens de um sistema usando a tecnologia “data warehouse”. Também serão mostrado, os resultados das pesquisas à tomada de decisão.

O capítulo 7 se destina às conclusões deste trabalho, sugestões para trabalhos futuros e tendências das aplicações com "data warehouse" e a “lógica nebulosa”.

## CAPÍTULO 2 – A Ciência da Informação e a Recuperação de Informação

Neste capítulo serão abordadas as terminologias básicas de gestão de informação, os conceitos de gestão, recuperação, gerenciamento e fontes de informação sob o ponto de vista de Ciência da Informação e suas perspectivas. Também serão analisados os conceitos básicos de gestão de informação e do conhecimento, sistemas de informação enfatizando o suporte à decisão e conceitos de dados operacionais e dados informacionais.

Desde os anos 60 do século XX se menciona com diversas denominações a existência de uma prática profissional e acadêmica que lida com a informação, considerando sua produção, armazenamento, transferência e uso, como objeto específico de estudo. A este fazer se denominou prática ou atividade de informação e, ao campo do conhecimento, Ciência da Informação. No decorrer dos últimos 40 anos, esta nova prática profissional, como uma consequência lógica de seu próprio desenvolvimento, foi obrigada a buscar soluções práticas e teóricas. Em seu cenário fica mais clara e necessária a investigação, sendo que esta última é a atividade intelectual por excelência, que permite obter novos conhecimentos. O início da investigação na Ciência da Informação é provavelmente um dos melhores sinais de desenvolvimento deste novo espaço.

Mencionaremos aqui personagens e fatos que não podem ser esquecidos por suas contribuições para o desenvolvimento da Ciência da Informação.

O nascimento da Ciência da Informação pode ser visto como consequência de uma sucessão de técnicas relacionadas com o registro físico do conhecimento, principalmente a escrita. A escrita permitiu registrar, estocar e recuperar o conhecimento, gerando uma espiral cumulativa de textos cujo potencial foi amplificado quando Johann Gutenberg inventou o tipo móvel e apresentou a primeira prensa na Europa.

O sucesso do invento de Gutenberg só não foi mais imediato pelo fato de que naquela época poucas pessoas sabiam ler. Em uma sociedade basicamente agrária, os camponeses nada tinham a ganhar com a alfabetização, e em geral não aspiravam a ela, porém, a Revolução Industrial iniciada em meados do século XVIII provocou o êxodo das populações do campo para a cidade. Deu impulso à procura por mais informação e à qualificação necessária para sua interpretação e utilização. A construção de estradas e o surgimento das estradas de ferro facilitaram a expansão do comércio e a distribuição de livros e jornais. A velocidade das mensagens passou da velocidade do cavalo para a da locomotiva e desta para a eletricidade. McGarry, (1999, p. 90-93).

Paul Otlet publicou em 1934 uma obra cujo título, “Tratado de Documentação”, procurou dimensionar as novidades tecnológicas que permitiam condensar e organizar a informação de acordo com suas necessidades e objetivos. O “Tratado de Documentação”, consiste na exposição das, “noções relativas ao livro e ao documento, ao emprego separado dos elementos que constituem a Documentação” e “se propõe a servir de base inicial e geral para que, em cada caso, o documentalista, em seu centro de documentação, elabore seu próprio *Manual de Documentação*” Robredo (2003).

Paul Otlet e Henri la Fontaine, autores da Classificação Decimal Universal, fundaram também o *International Institute for Bibliography*, cujo objetivo era sintetizar toda a produção bibliográfica internacional em fichas padronizadas. Otlet e La Fontaine, de acordo com Lopez Yepes, são os precursores do ‘Movimento da Documentação’, e, suas idéias e realizações, ao introduzirem o novo conceito de ‘documentação’, introduziram, também, um novo paradigma.

De acordo com o exposto por Robredo (2003), “os repertórios bibliográficos são organizados de forma que uma ênfase especial é dada ao conteúdo dos documentos e à maneira como esse conteúdo é tratado e analisado, visando a armazenagem organizada da informação correspondente e, sobretudo, sua recuperação direcionada ao serviço do usuário”.

Trabalhos publicados por Vannevar Bush propunham a criação de sistemas de organização e recuperação das informações capazes de simular o caráter associativo do pensamento humano e também as principais idéias de um sistema hipertexto em sua obra “*As we may think*” (1945).

Em 1948, Claude Shannon e Weaver anunciaram a “Teoria matemática da comunicação”, além de contextualizar os elementos que integralizavam um sistema de engenharia da comunicação, possibilitou precisar um esquema da comunicação humana e portanto, determinar a informação que até aquele momento era conhecida intuitivamente, cuja teoria seria utilizada mais tarde para definir o objeto de estudo da Ciência da Informação.

Durante a II guerra, houve um desenvolvimento maciço da informação à serviço da indústria e da investigação bélica e militar. Com o término da guerra e o mundo dividido em duas concepções diretamente opostas: capitalismo e socialismo, teve início uma batalha que tinha como objetivo alcançar a hegemonia militar, econômica e política. Conseqüentemente, ocorreu o surgimento de uma quantidade enorme de informação, difícil de controlar e organizar com as ferramentas que havia até o momento.

A organização e uso da informação adquiriram um interesse primordial para os governos, principalmente a União Soviética e os Estados Unidos, que encaminharam recursos

para garantir melhor aproveitamento da recuperação de informação científica e tecnológica. O governo americano, preocupado em perder sua supremacia, cria um comitê especial (*Science Advisory Committee*) que resolveria o problema de acessar satisfatoriamente a literatura científica.

Outro fato, considerado como explosão da informação, ligado ao crescimento exponencial da produção e uso de documentos, em especial nas áreas científico-técnicas, que envolvia uma diversidade considerável de profissionais, em diferentes contextos. Aconteceu e decorreu, a partir do desmantelamento pelas potências aliadas, dos arquivos técnicos da indústria alemã, após o fim da Segunda Guerra Mundial. Com efeito, no anseio de conhecer a fundo o real nível dos conhecimentos científicos e tecnológicos da Alemanha, iniciou-se uma verdadeira maratona de análise do conteúdo de toda a documentação encontrada. Robredo (2003).

Tudo até aqui considerado, de algum modo, confirma que a Ciência da Informação surgiu de um grande volume de informação e da necessidade de utilizar métodos e ferramentas novas que permitam a busca e recuperação da informação.

## 2.1 A Ciência da Informação

Segundo Shera e Cleveland (1977), a década de 60 forneceu um clima favorável para o desenvolvimento da Ciência da Informação. Os problemas relacionados com o tratamento da informação começavam a ser abordados por parte da comunidade científica mundial, ao mesmo tempo em que se vivia um período de acelerado desenvolvimento tecnológico.

A primeira formulação do que seria a Ciência da Informação surgiu como resultado das conferências do Georgia Institute of Technology, realizadas entre 1961 e 1962: “Ciência da Informação é a ciência que investiga as propriedades e comportamento da informação, as forças que regem o fluxo da informação e os meios de processamento da informação para uma acessibilidade e usabilidade ótimas”, Shera e Cleveland, (1977). Os processos incluem a origem, disseminação, coleta, organização, recuperação, interpretação e uso da informação. O campo deriva ou relaciona-se com a matemática, a lógica, a lingüística, a psicologia, a tecnologia da computação, a pesquisa operacional, as artes gráficas, as comunicações, a biblioteconomia, a administração e alguns outros campos, Shera e Cleveland (1977).

Em 1968, Harold Borko formulou uma definição complementar, ressaltando suas características tanto de ciência pura como de ciência aplicada.

“Ciência da Informação é a disciplina que investiga as propriedades e o comportamento da informação, as forças que regem o fluxo da informação e os meios de processamento da informação para acessibilidade e usabilidade ótimas. Está relacionada com o corpo de conhecimento que abrange a origem, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação. Isto inclui a investigação das representações da informação nos sistemas naturais e artificiais, o uso de códigos para a transmissão eficiente de mensagem, e o estudo dos dispositivos e técnicas de processamento de informação tais como computadores e seus sistemas. É uma ciência interdisciplinar derivada e relacionada a vários campos tais como matemática, lógica, lingüística, psicologia, tecnologia da computação, pesquisa operacional, artes gráficas, comunicações, biblioteconomia, administração e outros campos similares. Possui um componente de ciência pura, que investiga o assunto sem considerar suas aplicações, e um componente de ciência aplicada, que desenvolve serviços e produtos”. Borko, (1968).

Outras definições:

Saracevic (1992) afirma, por sua vez:

“A Ciência da Informação é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais”.

Robredo (2003) define:

“Ciência da Informação é a que investiga as propriedades e comportamento da informação, as forças que regem o fluxo da informação e os meios de processamento da informação para um máximo de acessibilidade e uso. O processo inclui a origem, disseminação, coleta, organização, armazenamento, recuperação, interpretação e uso da informação. O campo deriva ou relaciona-se com a matemática, a lógica, a lingüística, a psicologia, a tecnologia computacional, as operações de pesquisa, as artes gráficas, as comunicações, a biblioteconomia, a gestão e alguns outros campos”.

Considerações da década de 90 aos dias atuais:

1989: Tendência crescente para associar Biblioteconomia à Ciência da Informação.

1991: Na Universidade de Tempere, Finlândia foi realizada a Primeira Conferência Internacional de Conceitos de Biblioteconomia e Ciência da Informação. Na mesma Conferência, Saracevic afirma sua opinião de que há uma diferença substancial entre biblioteconomia (aquela se ocupa das funções culturais e técnicas das bibliotecas) e Ciência da Informação sendo esta de natureza essencialmente tecnológica.

1994: É criada e realizada a primeira Conferência da Organização de Ciência da Informação (FIS) - *Foundations of Information Science*, na Universidade de Zaragosa,

Espanha, com o seguinte subtítulo: “Dos computadores e da física quântica às células, sistemas nervosos, e sociedades”.

1994: LE COADIC, caracterizou a Ciência da Informação enunciando procedimentos e atividades sobre a informação: “Tem por objeto o estudo das propriedades gerais da informação (natureza, gênese, efeitos) ou seja, mais precisamente:

i) a análise dos processos de construção, comunicação e uso da informação;

ii) a concepção dos produtos e sistemas que permitem sua construção, armazenamento e uso, e que a Ciência da Informação é ciência, produção consciente da espécie humana com origens bem precisas, um objeto e um conteúdo bem definidos e especialistas facilmente identificáveis.”

1995: SARACEVIC aborda o problema da interdisciplinaridade da Ciência da Informação, lembrando, ainda, a sua origem e antecedentes sociais, que evoluíram para a recuperação da informação constituindo “*a atividade principal da Ciência da Informação.*”

1996: Na Universidade de Tecnologia de Viena, foi realizada a Segunda Conferência da Organização de Ciência da Informação - *Second Conference on the Foundations of Information Science* (FIS 96), abordando os seguintes temas: abordagens para unificação; conceitos de informação; sistemas auto-organizados; vida e consciência, e sociedade e tecnologia.

1996: Em Copenhague, Dinamarca, foi realizada a Segunda Conferência Internacional de Conceitos em Biblioteconomia e Ciência da Informação - *International Conference on Conceptions of Library and Information Science*, tratando dos seguintes tópicos: o conceito de informação, busca de informação e comportamento, gestão da informação, métodos de investigação, organização do conhecimento e comunicação, interação na recuperação da informação; infometria e cientometria; redes e mídia.

1999: BATES, em artigo publicado no mesmo ano, dá continuidade a suas pesquisas sobre diversos problemas da indexação (representação de conteúdo, representação temática) em relação às implicações no desenho e sistemas de informação e à qualidade da recuperação em grandes bases de dados na Internet, e atribui à Ciência da Informação o papel de uma metaciência, ou seja, a ciência que paira para lá ou acima de todas as outras porque a todas diz respeito.

1999: É realizada a Terceira Conferência Internacional de Conceitos em Biblioteconomia e Ciência da Informação - *International Conference on Conceptions of Library and Information Science*, com ênfase nas bibliotecas digitais, conceitos interdisciplinares, desafios e oportunidades, com destaque dos artigos de BUCKLAND e

BRUCE, intitulados respectivamente “Vocabulário como um conceito central na Ciência da Informação e da biblioteca” e “As bibliotecas digitais e o fim dos sistemas de informação tradicionais”.

2002: É realizada uma nova Conferência Eletrônica sobre os Fundamentos da Ciência da Informação, sobre os auspícios da Organização de Ciência da Informação com o tema central: “Natureza da informação: conceitos, falsos conceitos e paradoxos.”, sendo também realizada a Quarta Conferência Internacional de Conceitos de Biblioteconomia e Ciência da Informação. *International Conference on Conceptions of Library and Information Science*, com o tema central: “Métodos e estruturas de trabalho emergentes”. Foram abordados temas como: informática social, tecnologia da informação.

Portanto, o conceito de informação, o processo de surgimento da Ciência da Informação e a evolução de seu conceito dentro de um movimento unificador da teoria da informação, expande o escopo da própria Ciência da Informação e uma nova visão do processo de comunicação associado a um enfoque sociológico da transmissão da informação e da geração do conhecimento, enfim, situam-se às atividades relacionadas com a documentação, os arquivos, a armazenagem, a difusão e a recuperação da informação, dentro desse conceito expandido da Ciência da Informação.

### 2.1.1 Perspectivas da Ciência da Informação

Como descrito nos itens anteriores, a Ciência da Informação teve seu aparecimento e expansão no pós guerra, principalmente a partir de 1950, quando pesquisas e documentos mantidos fora do conhecimento coletivo de informação foram então liberados. A grande crise da época foi então, como lidar com o enorme volume de informação disponibilizada, utilizando os mecanismos e tecnologias acessíveis naquela época. Era necessário gerenciar e controlar o grande volume de informação, estocar e caracterizar seu conteúdo, priorizar o seu uso de acordo com as diferentes comunidades informacionais e promover uma divulgação seletiva e retrospectiva para evitar a duplicação do esforço de pesquisa e permitir que a sociedade conhecesse os avanços que haviam sido efetivados.

Grande parte das pesquisas, realizadas nas duas décadas subsequentes a 1950 foram para tentar resolver problemas de divulgação, principalmente de pesquisas que não se tornavam conhecidas. Contudo, de uma maneira geral, a interação entre o receptor e os estoques disponíveis de informação era sempre mediada por um profissional da informação e o tempo de retorno da informação solicitada estava na dependência das características internas

de eficácia das unidades de informação, que hospedavam os estoques. O fluxo da informação era uni-direcionado, o receptor tinha acesso a um estoque de cada vez e avaliava a relevância de sua busca neste estoque, orientado sempre pelo mediador.

A crise inicial da Ciência da Informação, senão resolvida, foi bastante minorada pelo computador. A atual crise que se acerca da Ciência da Informação é mais profunda. Nestes quase 50 anos que se passaram, a área não acompanhou a mudança radical que aconteceu e continua em mudança nos modelos relacionados a tecnologia da informação, modificações estas advindas das transformações acontecidas na microeletrônica, na telecomunicação e suas técnicas assessórias.

As mudanças na tecnologia da informação ocorridas durante os últimos anos, reorganizaram todas as atividades associadas à Ciência da Informação. A sociedade sempre foi mais afetada pelas transformações, pela natureza da tecnologia do que pelo seu conteúdo, pelo menos a curto prazo. Contudo, a chegada da sociedade eletrônica de informação modificou novamente a delimitação de tempo e espaço da informação. A importância do instrumental da tecnologia da informação forneceu a infra-estrutura para modificações sem retorno, das relações da informação com seus usuários.

Barreto (1996) afirma, que tão importante como o instrumental tecnológico foram as transformações associadas a interatividade e interconectividade no relacionamento dos receptores com a informação e define:

a) interatividade - representa a possibilidade de acesso em tempo real pelo usuário à diferentes estoques de informação; às múltiplas formas de interação entre o usuário e às estruturas de informação contidas nestes estoques. A interatividade modifica a relação usuário, tempo e informação. A interatividade reposiciona os acervos de informação, o acesso à informação e a sua distribuição, e o próprio documento de informação ao liberar o receptor dos diversos intermediários que executavam estas funções em linha e em tempo linear passando para um acesso on-line e com linguagens interativas;

b) interconectividade - referência à possibilidade que tem o usuário de informação em deslocar-se, no momento de sua vontade, de um espaço de informação para outro espaço de informação. De um estoque de informação para um outro estoque de informação. O usuário passa a ser o seu próprio mediador na escolha de informação, o determinador de suas necessidades. Passa a ser o julgador de relevância do documento e do estoque que o contém em tempo real, como se estivesse colocado virtualmente dentro do sistema de armazenamento e recuperação da informação. A interconectividade reposiciona a relação usuário, espaço e informação.

Estas mudanças operadas no aspecto tecnológico das atividades de armazenamento e transmissão da informação vêm trazendo mudanças contínuas, também na relação da informação, com seus usuários, com seus intermediários, com a pesquisa e com o ensino em Ciência da Informação.

Destacam-se como instabilidades mais notáveis, as mudanças na estrutura de informação (a interação em tempo real com a estrutura da informação tem questionado o caráter alfabético e linear do documento texto); mudanças no fluxo de informação (o fluxo da informação entre os estoques ou espaços de informação e os usuários permeiam dois critérios: o da tecnologia da informação que almeja possibilitar o maior e melhor acesso a informação disponível e o critério da Ciência da Informação, que intervém para, também, qualificar este acesso em termos das competências para assimilação da informação) e o homem de informação e o trabalho em Ciência da Informação (o profissional desta área se encontra, nesta atualidade, como se em um ponto no presente entre o passado e o futuro; convive com tarefas e técnicas tradicionais de sua profissão mas precisa atravessar para uma outra realidade, onde estão indo seus clientes, e aprender a conviver com o novo e o inusitado, numa constante renovação da novidade).

## 2.2 Gestão da Informação

A Informação assume, hoje em dia, uma importância crescente. Ela se torna fundamental para a empresa na descoberta e introdução de novas tecnologias, exploração das oportunidades de investimento e ainda na planificação de toda a atividade industrial. É a importância que a informação assumiu na atualidade pós-industrial recoloca para o pensamento questões sobre a sua natureza, seu conceito e os benefícios que pode trazer ao indivíduo e no seu relacionamento com o mundo em que vive.

Com a crescente necessidade de administrar a informação, considerando-se os aspectos humanos e da tecnologia da informação relacionados, formou-se uma área profissional, a "Gestão da Informação", que se configurou como uma área de estudos já consolidada nos Estados Unidos e na Europa, cujos conteúdos teóricos e operacionais têm se transformado em ferramenta imprescindível para qualquer organização que necessite produzir, localizar, coletar, tratar, armazenar, distribuir e estimular o uso da informação. A relação transdisciplinar da Gestão da Informação com a Ciência da Informação, Administração e a Informática, resulta em um conjunto de habilidades e conhecimentos teórico-práticos que possibilitam a estruturação de sistemas de informação, assim como o oferecimento de serviços, produtos e atividades de informação.

A gestão de informação e a sua inserção na estratégia empresarial são um fator chave na criação de valor agregado e das vantagens competitivas para a empresa. Se, por um lado, ajudam a detectar novas oportunidades e criar vantagens competitivas, por outro, ajudam a defendê-la de ameaças provenientes da concorrência. A informação tornou-se uma necessidade crescente para qualquer setor da atividade humana, e é considerada como o ingrediente básico do qual dependem os processos de decisão.

De acordo com Jannuzzi (2004), a informação é fundamental para o bom desempenho de um sistema. Para as empresas, ela se apresenta como o recurso estratégico para a competitividade. Assim, a preocupação de se melhorar cada vez mais o uso de informações nas empresas tem sido objeto de muitos estudos e propiciado inúmeras publicações sobre o assunto, com ênfase, principalmente, em seu ambiente interno, implantação e desenvolvimento de sistemas de informação, alguns sobre as informações externas. Grande parte das informações que as empresas necessitam para o seu processo decisório existem e estão sendo de alguma forma ofertadas, mas a dificuldade de obtê-las provavelmente decorre do fato de que seu formato esteja distante das condições de percepção e assimilação dos usuários nas empresas.

### 2.2.1 Gestão do Conhecimento e o capital intelectual.

A gestão do conhecimento é o reconhecimento da importância estratégica da administração do conhecimento das empresas e configura-se como a mais recente fase da evolução da gestão de informação.

Para alguns autores informação como conhecimento, pode ser definida como:

Segundo Toffler (1995) à medida que os sistemas de informação passam a oferecer particularidades mais complexas, tais como: coleta, integração, avaliação dos dados, e passagem da informação por modelos sofisticados, seu valor potencial atinge o auge, pois hoje se cria e interligam-se novas redes de conhecimento<sup>2</sup>, ligando os conceitos uns com os outros de maneira impressionante, armando notáveis hierarquias de inferência, desovando novas teorias, hipóteses e imagens baseadas em inusitados pressupostos: novas linguagens, códigos e lógica. Empresas coletam e armazenam dados mais puros do que qualquer outra geração na história, dando-lhes contexto e, assim, transformando-os em informações.

---

<sup>2</sup> Conhecimento, para Toffler (95) é informação que foi submetida a um processo de refinamento, para se tornar afirmação mais geral. Este termo incluirá: informações, dados e imagens, atitudes, valores.

O conhecimento tornou-se o recurso essencial das organizações; os tomadores de decisão, ao usá-lo, não se limitam a transferir dados, pois também combinam dados históricos, analisam os dados estatísticos e os organizam, às vezes, criando um novo "*feedback*" e assim, o que se obtém como resultado é diferente do dado originário.

Na gestão do conhecimento não podemos deixar de considerar o valor do capital intelectual, na importância do ser humano como um dos principais personagens da economia, pois é ele quem detém o principal recurso competitivo das organizações: o conhecimento. O capital humano começa a crescer em competência e conhecimento, quando a empresa identifica e adquire as informações que faltam e compartilha esses conhecimentos entre si. Considerando então o capital intelectual como um dos principais geradores de riqueza das empresas, as empresas deverão gerenciar em um esforço multidisciplinar o capital intelectual e a gestão da informação.

### 2.3 Recuperação de Informação

Quando Calvin Mooers, no início de 1950 cunhou o termo "*information retrieval*", ele também definiu os problemas que as atividades endereçam: "Information retrieval abrange o aspecto intelectual da descrição da informação e sua especificação para pesquisa e também qualquer sistema, técnicas ou máquinas que são empregadas para controlar operações" . Mooers (1951).

Ainda segundo, Mooers, (1951), no contexto da Ciência da Informação, o termo "recuperação de informação" significa a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário, e também consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes. O termo pode ainda ser empregado para designar a operação que fornece uma resposta mais ou menos elaborada a uma demanda, e a resposta é convertida num produto cujo formato é acordado com o usuário.

Neste trabalho optou-se por uma abordagem que enfatiza os processos de busca de informação, (encontrar a informação desejada e apresentá-la para o usuário). Excluiu-se o tratamento documental que, embora complementar, mobiliza uma outra bibliografia, preocupando-se com a integração da "lógica nebulosa" à recuperação de informação em "data warehouse", que serão tratados nos capítulos a seguir.

### 2.3.1 Sistemas de recuperação de Informação

A Recuperação de Informação se firmou como uma área de pesquisa autônoma no seio da Ciência da Informação, com um acelerado desenvolvimento. Para Saracevic (1999), a Recuperação de Informação pode ser considerada a vertente tecnológica da Ciência da Informação.

O termo “recuperação de informação” atribuído a sistemas computacionais é ainda hoje bastante questionado, sendo que muitos autores preferem o termo “recuperação de documento” (“*document retrieval*”) ou “recuperação de textos” (“*text retrieval*”). De fato, os sistemas não recuperam “informação”, mas sim documentos ou referências cujo conteúdo poderá ser relevante para a necessidade de informação do usuário. Neste trabalho será utilizada a designação original “recuperação de informação”, ficando subentendido que se trata de uma “informação” potencial, uma probabilidade de informação contida nas bases de dados.

Os primeiros sistemas de recuperação de informação baseavam-se na contagem de frequência das palavras do texto e na eliminação de palavras reconhecidamente de pouca relevância. Nos trabalhos de Luhn e Salton observa-se inicialmente uma crença de que os métodos puramente estatísticos seriam suficientes para tratar os problemas relacionados à recuperação de informação. Porém, no transcorrer de suas pesquisas, percebe-se uma busca por métodos de análise semântica mais sofisticada. Desde os seus primeiros trabalhos, Salton se mostra interessado pela utilização de processos de tratamento da linguagem natural na recuperação de informação. Em livro de 1983, Salton e McGill apresentam em um capítulo intitulado “*Future directions in Information Retrieval*” a aplicação do processamento da linguagem natural e da “lógica nebulosa” na recuperação de informação.

Neste trabalho, trataremos da utilização de técnicas da “lógica nebulosa”, cujas pesquisas, utilizarão técnicas de SQL nebuloso para buscas em base de dados multidimensionais chamadas “data warehouse”.

A aproximação da Ciência da Informação com a “lógica nebulosa” e “data warehouse” para recuperação de informação, será descrita no Capítulo 6.

## 2.4 A era da informação

Para (Mcgowan (1997) apud Rossini, (2003) p. 68), nenhuma empresa escapa dos efeitos da informação. As drásticas reduções nos custos de obtenção, processamento e transmissão de informações alteram a maneira de se fazer negócio. Afeta todo o processo

através do qual elas criam os seus produtos e, além disso, reformulam o próprio produto (todo o pacote de bens físicos, serviços e informação que as empresas provêm para gerar valor aos seus compradores).

Vê-se a informação em todo e qualquer ambiente, desde o lazer através da multimídia em televisão, até a internet. A informação, agora, não tem mais barreiras. Para o mesmo autor, a revolução da informação afeta a competição de três formas vitais: muda a estrutura do setor e, ao fazê-lo, altera as regras da competição; cria vantagens competitivas, proporcionando às empresas novas formas de superar seus rivais e origina negócios completamente novos, começando, freqüentemente, dentro das operações já existentes na empresa.

Mcgowan afirma ainda que os administradores devem seguir alguns passos para aproveitarem as oportunidades que a revolução da informação criou, como: avaliar a intensidade das informações existentes, potenciais dos produtos e processos, e identificar as prioridades nas unidades de negócios para investimentos em tecnologia da informação; determinar o papel da tecnologia da informação na estrutura do setor; analisar como a tecnologia da informação poderá afetar as forças competitivas (não somente cada força terá a probabilidade de causar mudanças, mas também os limites do setor poderão mudar); investigar como a tecnologia da informação poderá gerar novos negócios (os administradores devem considerar oportunidades para criar novos negócios e diversificar os já existentes) e desenvolver um plano para tirar vantagem da tecnologia da informação.

Os primeiros quatro passos deveriam levar um plano de ação para capitalizar a revolução da informação. Esse plano de ação deve ser em ordem de importância.

Mcgowan diz que no passado os diretores simplesmente delegavam decisões tecnológicas aos sábios dos computadores da empresa e cuidavam de outros assuntos. Hoje, os diretores não podem mais evitar as decisões sobre o processo de tecnologia. Assim como os diretores não dominavam a linguagem tecnológica, os especialistas técnicos também não tinham compreensão suficiente dos objetivos e das metas dos negócios. Esse desconhecimento tendia a acarretar um adiamento nas decisões que trazia conseqüências graves, como: perda de oportunidades competitivas importantes, desperdícios de capital em tecnologias não produtivas e muitas despesas para consertar tudo. Tornava-se necessário, então, a adoção de uma nova abordagem com relação ao processo de tomada de decisão em tecnologia de informação: um processo que unisse o conhecimento técnico dos especialistas de computadores com a visão da alta gerência.

Pois bem, para que os empreendedores e administradores atuais possam tirar o melhor proveito da tecnologia é necessário que os mesmos conheçam a terminologia da linguagem da

informação, assim como os propósitos e aplicabilidade correta das novas ferramentas que surgem a cada instante, é necessário que o novo administrador tenha conhecimento de todo o processo da empresa e também da tecnologia da informação.

## 2.5 A informática e sistemas de informação

Consideremos a informática como a ciência que tem como objetivo viabilizar, garantir e suportar o tratamento e a comunicação das informações da organização. A informática considerada como ciência racional e automática da informação engloba vários componentes para o tratamento da informação; outros termos passaram a definir genericamente o tratamento da informação e os sistemas de informação como um campo multidisciplinar que situa-se nas áreas da computação, pesquisa operacional, sociologia, política, psicologia e gerenciamento, que contribuem para solucionar problemas na organização.

Nos conceitos da administração Albertin, (1996), os sistemas de informação têm grande influência na administração porque, com o uso estratégico de informação, pode-se afetar a produção, tanto física como intelectual, criar qualquer produto ou serviço que a organização vende, como também a coordenação da organização, por seu poder de encurtar distâncias e diminuir tempo, e, principalmente, as memórias das organizações, através de seus bancos de dados. Conclui-se, então, que os sistemas de informação têm permitido mudanças fundamentais na forma como o trabalho é feito; têm permitido também a integração de funções de negócios em todos os níveis internos e entre organizações; e causam mudança no clima de competitividade em muitas áreas, tais como indústrias, hospitais, universidades, etc. Como uso de sistemas de informação tem-se a oportunidade de apresentar novas estratégias para as organizações que estabelecem suas missões de aplicações bem-sucedidas e irão provocar transformações necessárias nas organizações, nas próximas décadas, para que possam prosperar no ambiente de competitividade.

## 2.6 Sistemas de informação e sistemas de apoio à decisão

Muitas empresas, através dos anos, têm construído uma grande e desordenada coleção de aplicações operacionais, que atendem ao processamento cotidiano das informações, porém, tais aplicações são frágeis e não integradas.

O que acontece é que essas informações estão desorganizadas, ou seja, existe uma grande dificuldade para que as informações estratégicas cheguem realmente até quem deve.

Ter as informações certas, na hora certa, é fundamental para que uma empresa se mantenha viva no mercado.

A disponibilidade de informações para uma melhor perspectiva da tomadas de decisão é mais que um simples sistema entrada-processamento-saída de informações; é gerenciamento de soluções, de problemas criados nos negócios, baseado na tecnologia de informação para um propósito específico e devem destinar-se a atender às necessidades informacionais dos tomadores de decisão de forma rápida. Para as tomadas de decisões críticas, devem-se gerenciar o uso intensivo de dados do seu macro ambiente tais como clientes, mercados, governo, concorrentes; devem proporcionar acesso a dados disponíveis em internet e intranet, a taxas e índices do mercado financeiro, proporcionar acesso às informações detalhadas; às sumarizações, e, principalmente fornecer recursos fáceis de busca e seleções de variações e exceções por um processo de perguntas e respostas.

Logo, as aplicações sobre o negócio são as aplicações que analisam o negócio ajudando a interpretar o que ocorreu e a decidir sobre estratégias futuras para a empresa, chamados de sistemas de suporte à decisão. Para construir uma arquitetura de dados adequada, deve haver a preocupação em dar suporte aos bancos de dados operacionais, para dar suporte às operações dos negócios; e os bancos de dados de suporte a decisão para dar apoio às aplicações sobre o negócio.

Há, então, necessidade de se separar dados que dão suporte aos sistemas de caráter operacional da empresa, daqueles que dão suporte aos processos gerenciais e de suporte à decisão.

## 2.7 Dados operacionais e dados informacionais

Na administração de dados, o ponto principal é o enfoque do compartilhamento de bases de dados único pelos sistemas de informação.

Os sistemas de bancos de dados internos e externos à corporação devem estar preparados para desencadear novos tipos de perguntas que surgem com os problemas que preocupam o nível estratégico da empresa, onde se deve olhar não só para o momento, como também para o futuro, moldar esse futuro e gerar seguidamente o sucesso nas décadas que virão. Hamel (1995).

Sistemas de Apoio à Decisão(SAD) e Sistemas de Informações Executivos (SIE) possuem funcionalidade e desempenho diferentes dos Sistemas de Produção (também chamados transacionais) da empresa.

Os Sistemas de Produção são suportados por bancos de dados de sistemas transacionais, cujos dados, armazenados e coletados, não estão disponíveis aos usuários finais de forma adequada para análises posteriores. Esses sistemas são muito eficientes para automatizar a operação de uma corporação, mas não são eficientes para análise do processo de negócio como um todo.

Os conceitos de SIE e SAD são, muitas vezes, confundidos. Apesar de estarem relacionados, eles tratam de problemas diferenciados e, tipicamente, atendem a públicos-alvo diferentes. Um SIE é específico para uso dos tomadores de decisão, sendo que a maioria das aplicações disponíveis é do tipo "*display only*", podendo-se consultar e imprimir, sem permitir a manipulação de dados. Por outro lado, um SAD é tipicamente projetado para o nível intermediário de gerência, nível tático para simulações e cálculos e análises "*what-if*". SAD e SIE normalmente lidam com poucos usuários por vez, e os requisitos em termos de tempo de resposta, não são críticos. No entanto, usualmente lidam com buscas complexas, não antecipadas ou previstas; envolvem grande quantidade de registros básicos referentes aos processos operacionais da empresa, e usualmente, lidam com tendências e não com um único instante de tempo: cada elemento de dado é acompanhado pelo correspondente período de tempo a que se refere.

A importância de separar os dados que dão suportes aos sistemas de caráter operacional da empresa daqueles que dão suporte aos processos gerenciais e de suporte à decisão, está no fato de que, cada tipo de aplicação pode se concentrar naquilo que faz melhor e oferecer melhor funcionalidade e desempenho para seu caso específico, conforme Campos e Rocha (1998).

Conclui-se que a análise, a partir de sistemas transacionais, é um processo custoso e que não traz resultado eficiente para tomada de decisões.

## 2.8 Conclusões do capítulo

Concluimos que no contexto da gestão da informação nas empresas, de maneira geral, a informação destina-se a subsidiar o processo de decisão e controle nos diversos níveis hierárquicos da organização. Sob esse prisma, a informação adquire a condição de mensagem e intencionalidade, que determinam a probabilidade de serem assimiladas pelos usuários da empresa, quesitos estes considerados como necessários para que a informação possa ter um poder de ação. Assim, quando se concebe a informação como recurso estratégico para a competitividade das empresas, está se falando, em outras palavras, do poder de ação da

informação, principalmente, nas decisões referentes às ações estratégicas da organização. Conclui-se também que um sistema de informação na empresa visa ao compartilhamento das informações necessárias ao desempenho das funções.

Vimos que a informação é fundamental para o bom desempenho de um sistema e para as empresas, ela se apresenta como o recurso estratégico para a competitividade.

Quanto a evolução das tecnologias de informação tem refletido de maneira excepcional nos sistemas de informação das empresas, por meio de recursos como o banco de dados, que, possibilita o aumento na capacidade de armazenamentos dos dados, permitindo desta forma, que arquivos, cadastros etc. possam ser agrupados conjuntamente; ou por recursos como os sistemas integrados de gestão, que, permitem à empresa que os dados do sistema possuam uma entrada e após o seu armazenamento, sejam disponibilizados, de forma integrada, para todos na empresa, segundo os critérios de segurança no acesso e buscas a esses dados, que trataremos no próximo capítulo sobre os bancos de dados, seus métodos e linguagens que permitam a busca e recuperação da informação.

### CAPÍTULO 3 – Banco de dados: dos arquivos até os “data warehouse”

Neste capítulo serão tratados os banco de dados, sua definições, história e sua evolução, conceitos, modelos em Rede, Hierárquico e o Relacional, uma breve descrição de cada modelo, suas vantagens, desvantagens. Apresentaremos também a linguagem de busca SQL, segurança dos dados e uma breve descrição de “data warehouse”.

Em determinado momento o cérebro humano deixou de ser o único meio de processamento de informações. Em uma evolução que partiu do ábaco, utilizado por comerciantes há mais de dois mil anos, passou pelo matemático e filósofo Blaise Pascal que construiu em 1642 um engenho mecânico capaz de somar e subtrair números de oito algarismos. Charles Babage, em 1834 idealizou um mecanismo capaz de executar seqüências matemáticas pré-definidas (os primeiros programas), por fim chegamos ao engenho de Hollerith, que com sua tecnologia, deu cabo da extenuante tarefa do censo norte-americano.

Em 1890, o norte-americano Herman Hollerith concebeu uma máquina eletromecânica que lia uma série de dados perfurados em cartões e que fez com que o censo de seu país fosse realizado naquele ano em um terço do tempo do censo anterior. Hollerith mais tarde fundaria a *Tabulating Machine Company*, hoje conhecida como IBM. O primeiro passo fora dado para o processamento mecânico de informações.

Posteriormente, na primeira metade do século XX, houve um acelerado desenvolvimento de diversos instrumentos mecânicos e eletromecânicos com o propósito de facilitar cálculos numéricos. Em 1946, foi ligado pela primeira vez o ENIAC (*Electronic Numerical Analyzer and Computer*), o primeiro computador eletrônico, resultado temporão do esforço de guerra norte-americano, que operava sob lógica digital, capaz de efetuar 3.500 multiplicações por segundo, o que era assombroso para a época.

O ENIAC é o primeiro na linha evolutiva dos computadores digitais. Essa máquina eletrônica cuja capacidade de processar e armazenar informações crescia a cada dia, manipulava dados em níveis sequer sonhados pelos seus criadores e dava à informação processada importância fundamental na dinâmica da sociedade moderna.

A partir da Segunda Guerra, a evolução dos computadores e de sua capacidade de processar informações dependia da ciência que, neste ramo, progredia a passos largos com o impulso da Guerra Fria e da demanda por novos sistemas de telecomunicações. O homem, pela primeira vez, deparava-se com um rival na faculdade de manipular informações. Seu relacionamento com o computador, que desde o início apresentou um elemento de

desconfiança, afinal era uma máquina capaz de realizar tarefas lógicas de grande volume e em velocidade extraordinária se comparada aos humanos, além de contar com o pressuposto da infalibilidade. Os primeiros receios humanos foram de que o computador iria tomar seus lugares na realização de diversas tarefas, relegando o homem a uma situação de inferioridade frente ao então alcunhado "cérebro eletrônico". A atual substituição de diversos postos de trabalho pelo trabalho automatizado demonstra que o receio não era de todo infundado.

Informações organizadas em fichas e pastas suspensas em armários de grandes dimensões passaram a ser dominadas por computadores com resultados excelentes, impossíveis até então de serem obtidos pelo trabalho humano, o que representava uma verdadeira revolução. Tornou-se possível armazenar um maior volume de informações e processá-las mais rapidamente, agregá-las e combiná-las dos mais diversos modos, obter aquelas necessárias em tempo irrisório e muito mais. Assim nascia o banco de dados informatizado.

Igualmente a muitas tecnologias na computação industrial, os fundamentos de bancos de dados relacionais surgiram na empresa IBM, nas décadas de 1960 e 1970, através de pesquisas de funções de automação de escritório. Foi durante um período da história na qual empresas descobriram que era muito custoso empregar um número grande de pessoas para fazer trabalhos como armazenar e indexar (organizar) arquivos. Por este motivo, valia a pena os esforços e investimentos em pesquisar um meio mais barato e ter uma solução mecânica eficiente. Muitas pesquisas foram conduzidas durante este período, cujos modelos hierárquicos, de rede e relacionais e outros modelos foram descobertos, bem como muita tecnologia utilizada hoje em dia.

Os sistemas de informação, abstraindo as rotinas e procedimentos manuais de interface, têm evoluído nos últimos anos através de estágios bem definidos. Em cada um destes estágios podemos observar a predominância de um aspecto tecnológico que influencia a arquitetura dos sistemas desenvolvidos e as próprias abordagens utilizadas no processo de desenvolvimento. Nos anos 60, por exemplo, os programas eram o foco central dos sistemas de informação, com os aspectos relativos aos dados e a apresentação dos dados aos usuários ocupando um plano secundário. A abrangência dos sistemas era a empresa como um todo, sistemas corporativos e a plataforma tecnológica utilizada era a dos equipamentos de grande porte (mainframes) e dos minicomputadores.

### 3.1 Termos e Conceitos de organização de arquivos

Com o surgimento dos discos magnéticos, os dados passaram a ser armazenados de forma distinta do código executável dos programas. Nesta nova abordagem, os dados eram organizados em coleções de itens logicamente relacionados, compondo *arquivos*. Embora cada programa fosse isolado dos dados, continuava possuindo cópias das rotinas necessárias para permitir a gerência dos dados (localização, recuperação e armazenamento de dados). Nota-se que os procedimentos relacionados com a gerência de dados eram embutidos no código-fonte do programa, juntamente com as partes relacionadas com a interação com o usuário e com a própria lógica da função de negócio.

Um sistema de computador organiza dados segundo uma hierarquia que começa com bits e bytes e prossegue até campos, registros, arquivos e bancos de dados. Um bit representa a menor unidade de dados que um computador pode tratar (0 ou 1). Um grupo de bits, designado bytes, representa um único caracter, que pode ser uma letra ou outro símbolo. Um agrupamento de caracteres em uma palavra, um grupo de palavras ou um número complexo como (nome ou a idade de uma pessoa) é chamado de **campo**. Um grupo de campos relacionados, tal como o nome do aluno, o curso que está fazendo, a data e a nota compreendem um **registro**; ou seja, registros são considerados linhas formadas por um conjunto de colunas. Os registros não precisam necessariamente conter dados em todas as colunas, os seus valores podem ser nulos e um grupo de registros do mesmo tipo é denominado de arquivo. O arquivo de alunos de um curso pode ser agrupado com arquivos que contém os históricos pessoais e financeiros dos alunos para criar um banco de dados de alunos. O registro descreve uma **entidade**, que pode ser uma pessoa, um lugar, uma coisa ou um evento sobre os quais se têm informações. Um “pedido” é uma entidade típica de um arquivo de “pedidos de venda”, que mantêm informações do arquivo de “pedidos de venda” da empresa.

Cada característica ou qualidade que descreve determinada entidade é um atributo. Por exemplo, número do pedido, data, quantidade, número e quantidade do item, cada um seria um atributo do item de pedido. Os valores específicos que esses atributos podem ter são encontrados nos campos do registro que descrevem a entidade, “pedido”, logo os campos são as colunas de um registro, também chamadas de campos de dados. Os campos possuem características que definem o dado que será armazenado. Por exemplo, em um campo do tipo numérico, serão somente armazenados números. Os sistemas de banco de dados possuem regras para consistir os dados que são armazenados.

Cada registro de um arquivo deve conter no mínimo um campo que o identifique exclusivamente, de modo que possa ser recuperado, atualizado ou ordenado. Esse campo de identificação é denominado de **campo-chave**. Um exemplo de **campo-chave** é o “número do pedido”, do registro “pedido”, portanto, um campo chave é um conjunto de um ou mais campos que determinam a unicidade de cada registro. Por exemplo, se um banco (base) de dados tem como chave Código do Produto, sempre que acontecer uma inserção de dados, o sistema, irá fazer uma consulta para identificar se o registro já não se encontra gravado no arquivo. Caso exista, um novo registro não será criado, sendo que apenas a alteração do registro existente será possível.

Campo: nome, endereço

Nome do Campo	Nome	Endereço
	Carla Patrícia	Rua 21 nº 12
	Ana Paula	Rua 33 nº 69

Registros: nome mais o campo

Dados Armazenados em forma de Registros	Nome	Endereço
	<b>Carla Patrícia</b>	<b>Rua 21 nº 12</b>
	<b>Ana Paula</b>	<b>Rua 33 nº 69</b>

Nos sistemas de arquivos tradicionais, os programas compartilhavam dados armazenados em disco, mas dependiam da organização dos arquivos e de seus métodos de acesso. Isto significa que qualquer alteração na estrutura dos arquivos implicava a alteração das funções de gerência de dados, que estão replicadas em cada programa.

Vejamos o exemplo do programa codificado em linguagem COBOL (iniciais de COmmon Business Oriented Language) referente à atualização do cadastro de funcionários. Haberkorn (1985).

Um programa escrito em COBOL era obrigatoriamente formado de quatro partes denominadas “divisões”, e que aparecem na seguinte ordem:

Identification Division (comandos de identificação do programa).

Program-Id. Atualiza-Cadastro.

Author. Robinson.

Essa é a identificação do programa e do programador.

Environment Division (comandos de definição dos equipamentos)

## Configuration Section

Input-Output Section.

File Control.

Select Entrada    Assign To Reader

Select impressora    Assign To Printer

Select Cadastro    Assign To Disk    Acces Is Random

Data Division. (comandos de definição e especificação dos arquivos de dados, registro de dados e áreas de trabalho)

### File Section

Fd    Entrada

Recording Mode Is Fixed

Record Contains 80 Characters

Label Record Is Omitted

Datarecord Is Area-Ent.

### 01 Area-Ent

02 Codigo-Reg                    Picture 99.

02 Matricula.                    Picture X(5).

02 Tipo                            Picture X.

02 Salario                        Picture 9(6) V 99.

02 Data                            Picture 9(6).

02 Nome                          Picture X(25).

02 Banco.

    03 Número                    Picture 99.

    03 Conta                      Picture 9(7).

02 Carteita-Profissional.

    03 Numero-Cp                Picture 9(6).

    03 Serie                        Picture 999.

02 Dir                              Picture 99.

02 Dsf                              Picture 99.

02 Seguro                         Picture 9(5) V99

Os endereços dos vários parágrafos são simbólicos e pode ter um nome qualquer, porém exclusivo. O ponto indica o fim de uma seqüência de comandos.

Procedure-Division (onde são escritas as instruções e comandos de execução do programa)

Início.

Open input entrada output impressora i-o cadastro.

Estão se abrindo o arquivo entrada, impressora e cadastro; um é de entrada, outro de saída e o terceiro de entrada/saída.

ler-registro.

Read Entrada At End Go To Fim.

Rot-Inclusao.

If Tipo Equalto 'H'perform Horista

Else If Tipo Equal To 'M' Perform Mensalista

Else Move Tipo Inválido To Texto Go To Imp-Erro.

Horista.

If Salario Greather Than Ortn

Move 'Salario Horista Invalido' To Texto

Goto Imp-Erro.

Exit.

Mensalista.

If Salario Less Than Ortn

Move 'Salario Mensalista Invalido' To Texto

Go To Imp-Erro.

Exit.

Close Entrada Impressora Cadastro.

Stop Run.

Após terminar os dados de entrada, o programa desviará para FIM, fechando todos os arquivos e encerrando o processamento.

Programar era escrever em uma linguagem que tem sua própria sintaxe, a tarefa definida pelo analista. Era escrever de forma exata e bem detalhada e prevendo-se todas as situações que possam ocorrer com os dados de entrada. A grande desvantagem desta dependência (programa e dados) é que qualquer alteração em um campo, por exemplo, o seu tamanho ou a inclusão de novos campos na estrutura dos dados, teria que alterar todos os programas que acessassem estes dados.

### 3.2 Banco de dados e sistemas de gerenciamento de banco de dados

Os bancos de dados têm como principal objetivo facilitar o manuseio de dados armazenados em disco; dar maior flexibilidade na busca de registros ou grupos de registros; permitir a alteração de seus formatos, tornando-os independentes dos programas; reduzir a redundância de informações e obter melhor aproveitamento do espaço físico alocado. Deverá também prover um sistema de segurança na leitura, gravação e integridade dos dados. Finalmente, colocar à disposição do usuário um potente e simples conjunto de comandos que vão desde aqueles orientados para a leitura, atualização, pesquisa, acesso, criação, cancelamento, listagem, edição. Fará apresentação dos dados até comandos aritméticos e lógicos e de transferência que permitem o desenvolvimento de programas altamente sofisticados e que, por isso, são chamados linguagens de 4ª geração (*Querys languages*).

Exemplos de respostas que se podem obter em banco de dados com uso de comandos relativamente simples:

Quais são os funcionários que têm formação superior, são solteiros e trabalham há mais de três anos na empresa;

Quais pedidos de clientes da capital estão com o prazo de entrega vencido;

Quais os títulos cuja data de vencimento é hoje;

Qual o resultado das despesas e receitas lançadas neste mês;

Quais os produtos cujos saldos em estoque estão menor que os estoques mínimos;

Quais os títulos em aberto de determinado cliente.

Muller (2002) define um **banco de dados** como um conjunto de informações com uma estrutura regular. Um banco de dados é normalmente, mas não necessariamente, armazenado em algum formato de máquina lido pelo computador. Há uma grande variedade de bancos de dados, desde simples tabelas armazenadas em um único arquivo até gigantescos bancos de dados com muitos milhões de registros, armazenados em salas cheias de discos rígidos.

Laudon (2004) define que um sistema de gerenciamento de banco de dados SGBD consiste em uma coleção de dados inter-relacionados e em um conjunto de programas para acessá-los. Um conjunto de dados, normalmente referenciado como *banco de dados*, contém informações sobre uma empresa particular, por exemplo. O principal objetivo de um SGBD é prover um ambiente que seja adequado e eficiente para recuperar e armazenar informações de banco de dados.

Ainda conforme Laudon (2004), os sistemas de banco de dados são projetados para gerenciar grandes grupos de informações. O gerenciamento de dados envolve a definição de estruturas para armazenamento de informação e o fornecimento de mecanismos para manipulá-las. Além disso, o sistema de banco de dados precisa fornecer a segurança das informações armazenadas caso o sistema dê problema, ou contra tentativas de acesso não-autorizado. Se os dados devem ser divididos entre diversos usuários, o sistema precisa evitar possíveis resultados anômalos. A importância das informações na maioria das organizações e o conseqüente valor dos bancos de dados têm orientado o desenvolvimento de conceitos e técnicas para o gerenciamento eficiente dos dados.

O surgimento dos sistemas de banco de dados favoreceu fortemente a difusão do uso dos recursos computacionais, e este pôde ser visto como um passo adiante na evolução do sistema de arquivos, promovendo a independência dos programas em relação aos dados armazenados. Esta independência de dados é decorrente da introdução do sistema de gerência de banco de dados (SGBD), que reúne todas as funções necessárias a localização e a manipulação dos dados de interesse dos sistemas de informação. Dessa forma, os SGBDs surgiram como uma camada lógica entre a aplicação e os dados, fornecendo um mecanismo de abstração de detalhes, como os relacionados com os caminhos de acesso aos dados e as estruturas de armazenamento utilizadas em sistemas computacionais. Bases de dados são gerenciadas por um sistema de gestão de bancos de dados ou SGBD.

Os SGBDs não são caracterizados tão-somente pelos seus modelos de banco de dados, embora estes sejam a principal característica de um SGBD. Um SGBD é caracterizado também pelas suas funções, que implementam um conjunto básico de facilidades e serviços que definem a capacidade de banco de dados que um sistema deve possuir para ser considerado um sistema de banco de dados. Dessa forma, um SGBD deve disponibilizar funções voltadas para assegurar a persistência dos dados, a definição e a manipulação, como também, a carga e a descarga do banco de dados, a cópia, a restauração, a recuperação, a segurança e a integridade dos dados. O controle de concorrência de transações e o desempenho das aplicações que têm acesso aos dados contidos no banco de dados.

O esquema de segurança em um SGBD exige que o usuário ao acessá-lo digite sua senha. Essa senha pode abrir um arquivo inteiro, somente determinado registro ou bloquear apenas alguns campos de determinado tipo de registro. Exemplo: pode-se deixar que o grêmio da empresa vasculhe os registros de funcionários para saber quem é solteiro ou casado para organizar um jogo de futebol, mas o campo de salários lhes será vetado embora fisicamente possa estar ao lado do estado civil.

Dentro desse tema de segurança, os SGBD possuem ainda eficientes rotinas de controle de arquivos em ambientes multi usuários, principalmente quando dois comandos de atualização de um registro forem efetuados simultaneamente, e rotinas de “LOG” que visam evitar que o usuário fique em situação desagradável quando ocorrer a perda total dos arquivos (rotinas de “LOG” são aquelas que gravam em meios magnéticos todas as atualizações realizadas pelo usuário. Ao ocorrer um acidente, seja por erro de paridade, queda de energia, ou problema nas unidades, o próprio sistema faz todo o processo de recuperação partindo do *back-up* (cópias de segurança) anterior e refazendo todas as movimentações).

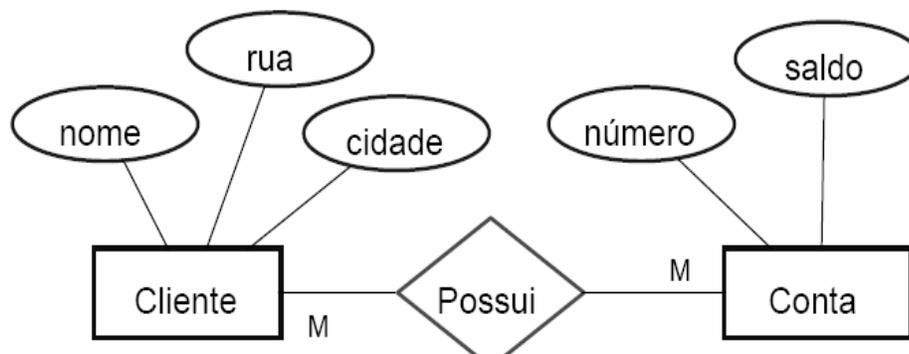
### 3.3 Modelos de Banco de Dados

A seguir são considerados modelos de banco de dados, de acordo com Melo (1997) e Laudon (2004).

#### 3.3.1 Modelos Conceituais

Os modelos conceituais, por incorporarem construções que possibilitam a apreensão do significado dos dados na realidade observada, são também chamados de *modelos semânticos*. No caso de bancos de dados, os modelos conceituais são usados para descrever objetos de dados. O Modelo Entidades-Relacionamentos (E-R) é o modelo mais utilizado no processo de projeto de banco de dados. No modelo E-R, uma *entidade* representa um objeto ou conceito existente no mundo real, como uma pessoa ou uma conta bancária.

Exemplo de um modelo Entidade-Relacionamento, onde um cliente pode ter muitas contas e uma conta pode pertencer a muitos clientes.



### 3.3.2 Modelos Lógicos

A maneira mais prática de classificar bancos de dados é de acordo com o modelo de programação associado ao banco de dados. Diversos modelos foram utilizados ao longo da história, por determinados períodos. Historicamente, o modelo de bancos de dados hierárquico foi implementado primeiro, depois apareceu o modelo de bancos de dados em redes; e então o modelo de bancos de dados relacional, que ganhou destaque.

Esses três modelos sem teoria própria surgiram basicamente a partir de estruturas e modelos de programação, não modelos de dados.

Modelos lógicos tradicionais descrevem dados por meio de estruturas de registros e são classificados de acordo com o tipo de estrutura e as operações em que se baseiam, (já descritos neste capítulo).

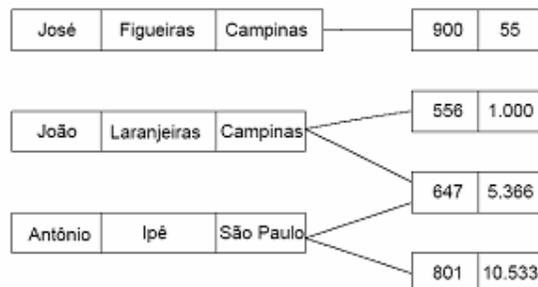
#### 3.3.2.1 Modelo em rede e hierárquico

O modelo em “rede” teve sua origem na linguagem de programação COBOL (*COmmon Busines Oriented Language*). Estendida com comandos de definição e manipulação de banco de dados, que deu ao usuário uma visão em rede do banco de dados. Consistia em conjuntos de entidades representadas por tipos de registro e relacionamentos onde neste tipo de estruturação lógica, uma determinada entidade pode estar subordinada a um número qualquer de entidades e, por sua vez, pode ter qualquer número de entidades a ela subordinado, caracterizando um relacionamento um-para-muitos entre os tipos de registro. Por exemplo, numa aplicação bancária, um tipo de registro PESSOA pode ter um relacionamento um-para-muitos com o uso de registro CONTA, indicando que cada pessoa pode possuir muitas contas. O modelo em rede permite que um tipo de registro esteja envolvido, como membro, em mais de um relacionamento, possibilitando a representação natural de relacionamentos muitos-para-muitos como tipos de registro. No exemplo, se uma conta puder pertencer a muitas pessoas, o relacionamento passa a ser um tipo de registro, membro de relacionamentos um-para-muitos com PESSOA e CONTA. Esta forma de estruturação lógica tende a gerar estruturas complexas devido ao número de relacionamentos que podem ser estabelecidos; no entanto, uma vantagem desse modelo é a flexibilidade, que permite às situações do mundo real serem representadas de uma forma direta. Uma desvantagem desse tipo de estrutura lógica é que sua complexidade torna difícil a construção de programas de aplicação. Devido ao fato de existirem diversos tipos de caminhos de acesso,

não existe simetria na formulação das recuperações o que torna a linguagem de manipulação de dados complexa.

## O Modelo de Rede

- Os dados são representados por coleções de registros e os relacionamentos por elos.



O modelo do banco de dados hierárquico apresenta dados aos usuários segundo uma estrutura semelhante a uma árvore, que não permite que um tipo de registro não esteja envolvido, como filho, em mais de um relacionamento. Dentro de registro, os elementos de dados são organizados em pedaços de registros chamados *segmentos*. Para o usuário, cada registro parece um organograma empresarial, com um segmento no nível mais alto denominado *raiz*. Um segmento de nível superior é ligado logicamente a um segmento de nível inferior em uma relação semelhante à relação pai-filho. Um segmento pai pode ter mais de um filho, mas um filho pode ter somente um pai. Este modelo pode ser considerado como uma restrição do modelo de redes.

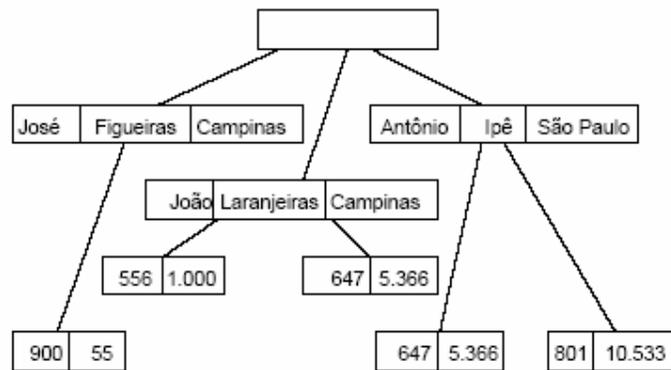
Uma estrutura hierárquica que poderia ser usada para um banco de dados de recursos humanos, o segmento-raiz é 'funcionário', contém informações básicas, como nome, endereço e número de identificação. Imediatamente abaixo dele há três segmentos-filhos: remuneração (contendo salário e dados de promoção), descrição do cargo, (contendo dados sobre cargos e departamentos) e benefícios (contendo dados sobre beneficiários) e opções de benefício. O segmento remuneração tem dois filhos abaixo dele: avaliação de desempenho (contendo dados sobre as avaliações de desempenho do funcionário) e histórico salarial (contendo dados históricos sobre seus salários anteriores). Abaixo do seguimento benefícios estão os segmentos-filhos pensão, seguro de vida e seguro saúde, que contêm dados sobre esses planos de benefícios.

Enquanto estruturas hierárquicas representam relacionamentos um-para-muitos, os bancos de dados em rede representam dados segundo uma lógica de relacionamentos muitos-

para-muitos. Em outras palavras, pais podem ter múltiplos filhos e um filho pode ter mais de um pai. Um típico relacionamento muitos-para-muitos para um banco de dados em rede é o relacionamento aluno-curso. Há muitos cursos em uma universidade, e um curso tem muitos alunos e um aluno faz muitos cursos.

## O Modelo Hierárquico

- Os dados e relacionamentos são representados por registros e ligações, respectivamente.
- Os registros são organizados como coleções arbitrárias de árvores.



Uma forma de referência posicional entre os diversos registros é feita utilizando-se o conceito de nível hierárquico, onde o segmento raiz é considerado como sendo o nível 1 e os níveis abaixo como sendo nível 2, nível 3.. nível n. Esta forma de estruturação, em comparação ao modelo de redes, tem a vantagem de simplificar os problemas de programação, em situações do mundo real que sejam naturalmente representadas de forma hierárquica, pois uma hierarquia estabelece um caminho de acesso pré-definido para recuperação dos registros no banco de dados. O preço desta simplificação é que este modelo apresenta a desvantagem de dificultar a modelagem de relacionamentos do mundo real, pois obriga todos os relacionamentos a serem expressos em termos de hierarquias, o que, obviamente nem sempre é a forma mais adequada. Um problema que surge é a necessidade de duplicação de dados para permitir que certas situações do mundo real sejam modeladas.

Consideremos a relação currículo entre as entidades curso e disciplinas. Esta relação por do tipo muitos-para-muitos, não pode ser representada de forma natural, quando utilizamos o modelo hierárquico. Uma alternativa para representar essa relação consiste em considerar as disciplinas como sendo subordinadas aos cursos, esta representação apresenta o

inconveniente de ocorrência de redundância de dados, pois, como o relacionamento é muitos-para-muitos, uma determinada disciplina poderá estar subordinada a diversos cursos. Portanto, seus dados serão duplicados tantas vezes quanto forem os cursos aos quais ela está subordinada, acontecerá que no caso das disciplinas que formam o núcleo comum de diversos cursos, esta redundância será elevada. Outro aspecto interessante a ser observado é que esse modelo não é flexível em relação a alterações, ou seja, é difícil adicionar novas entidades ou novos relacionamentos, o que, praticamente, obriga a estrutura de dados a ser estática. Esse fato representa uma desvantagem desse modelo em relação ao de redes, no qual é relativamente mais simples alterar a estrutura lógica para acompanhar a evolução das necessidades da organização.

### 3.3.2.2 Modelo Relacional

Em 1970, um pesquisador da IBM, E. F. Codd, publicou o primeiro artigo sobre bancos de dados relacionais. Este artigo tratava do uso de cálculo e álgebra relacional para permitir que usuários não técnicos armazenassem e recuperassem grande quantidade de informações. Codd visionava um sistema em que o usuário seria capaz de acessar as informações através de comandos e visava também prover mais flexibilidade na organização de grandes bancos de dados e aliviar alguns dos problemas dos modelos anteriores. Uma das principais razões que justificam o interesse por esse modelo é que, oferece uma representação simples e natural do banco de dados e seus usuários. A estrutura lógica oferecida para representar o mundo real consiste em tabelas denominadas relações, devido a esse modelo estarem fortemente baseada na teoria matemática sobre relações. Esse modelo oferece uma estruturação lógica, clara e flexível. O banco de dados pode evoluir, novas relações podem ser criadas sem causar impacto às estruturas existentes, devido ao fato de não existirem relacionamentos dependentes de subordinações, elos e outras construções que tornam os outros modelos pouco flexíveis.

Surgiu então o termo tabela, todos os dados de um banco de dados relacional são armazenados em tabelas. Uma tabela é uma simples estrutura de linhas e colunas. Cada linha contém um mesmo conjunto de colunas, mas as linhas não seguem qualquer tipo de ordem. Em um banco/base de dados, podem existir uma ou centenas de tabelas. O limite é imposto unicamente pela ferramenta de software utilizada.

Temos dois tipos de chaves: Chave Primária: é a chave que identifica cada registro dando-lhe unicidade. A chave primária nunca se repetirá. Chave Estrangeira: é uma chave

formada pela chave primária de outra tabela e a chave de um campo da tabela que recebe o relacionamento. Define um relacionamento entre as tabelas e pode ocorrer repetidas vezes.

Índices: Os dados são armazenados nas tabelas de forma não organizada. Sendo assim, para a recuperação dos dados é necessária a existência de mecanismos que facilitem a consulta e proporcionem uma performance aceitável para a mesma. Para isso, os sistemas de bancos de dados relacionais criam os índices das tabelas, sendo que esses índices são atualizados constantemente. Caso o índice se corrompa por algum motivo, é possível que pesquisas possam retornar resultados não desejados ou que inserções de chaves duplicadas aconteçam. Nesse caso, o banco de dados será corrompido também. Os sistemas de bancos (bases) de dados possuem mecanismos para evitar que esses eventos ocorram como também possibilitam a recuperação dos índices e consistência da tabela, caso eles aconteçam.

Diferentemente dos bancos de dados em rede, os relacionamentos entre as tabelas não são codificados explicitamente na sua definição. Em vez disso, se fazem implicitamente pela presença de atributos de mesmo tipo. Como resultado, bancos de dados relacionais podem ser reorganizados e utilizados de maneira flexível e de formas não previstas pelos projetistas originais. Por causa dessa flexibilidade, muitos bancos de dados são baseados no modelo relacional.

O modelo Relacional é simples, elegante e baseado em conceitos matematicamente fundamentados da Álgebra Relacional e do Calculo de Predicados (podendo utilizar expressões e fórmulas e operações and, or, etc.). A linguagem de consulta relacional como a SQL (*Structured Query Language*) é mais declarativa do que a linguagem navegacional, permitindo ao usuário especificar o que buscar no banco de dados. Desta maneira, as ligações inter-tabelas ocorrem logicamente, diminuindo sensivelmente a dependência física do modelo.

Abaixo são mostrados:

a) o esquema relacional de PESSOA e CONTA, em que o relacionamento um-para-muitos é logicamente definido pelo campo Limite - crédito em CONTA que referencia CPF em PESSOA;

b) o esquema relacional de PESSOA, POSSUI e CONTA, em que o relacionamento muitos-para-muitos é definido pelos campos Limite - crédito e Número-conta em POSSUI, que referenciam, respectivamente, CPF em PESSOA e Número em CONTA..

Exemplos das informações em um banco de dados

nome	rua	cidade	conta	saldo
José	Figueiras	Campinas	900	55
João	Laranjeiras	Campinas	556	1.000
João	Laranjeiras	Campinas	647	5.366
Antônio	Ipê	São Paulo	647	5.366
Antônio	Ipê	São Paulo	801	10.533

## O Modelo Relacional

Tabela Cliente (dados)

cód-cliente	nome	rua	cidade
015	José	Figueiras	Campinas
021	João	Laranjeiras	Campinas
037	Antônio	Ipê	São Paulo

Tabela Conta (dados)

nro-conta	saldo
900	55
556	1.000
647	5.366
801	10.533

Tabela Cliente-Conta  
(relacionamento)

cód-cliente	nro-conta
015	900
021	556
021	647
037	647
037	801

Figura 3.01: Modelo Relacional

Um exemplo de relacionamento entre as tabelas Funcionários, Produtos e Fornecedores.



Figura: 3.02: Representação dos relacionamentos Funcionários, Produtos e Fornecedores.

Outro exemplo, considerando as entidades “Departamento e Professor” e a relação “Emprega” entre essas entidades. Date (1988).

#### DEPARTAMENTO

CÓDIGO	NOME
FIS	Física
QUI	Química
INF	Informática

#### PROFESSOR

MATRÍCULA	NOME	IDADE	FORMAÇÃO
P1	Paulo	25	Engenheiro
P5	Carlos	31	Matemático
P9	José	27	Físico
P4	Robinson	55	Matemático

## EMPREGA

DEPARTAMENTO	PROFESSOR
INF	P5
FIS	P9
INF	P4
QUI	P1

Figura 3.03: Representação das entidades departamento e professor e da relação emprega.

Uma forma de representação dessas relações, sem apresentar os valores assumidos pelos atributos é:

DEPARTAMENTO (CÓDIGO, NOME)

PROFESSOR (MATRÍCULA, NOME, IDADE, FORMAÇÃO)

EMPREGA (DEPARTAMENTO, PROFESSOR)

Algumas diferenças entre o modelo relacional e o de redes é no que diz respeito ao número de operações necessárias nas linguagens de manipulação de dados: para atuar sobre uma estrutura em rede a linguagem necessita de mais operadores que para atuar sobre uma estrutura relacional. O modelo Relacional oferece uma estruturação lógica, clara e flexível. Outra diferença é que o modelo Relacional separa completamente as características estruturais do modelo das características de armazenamento físico, atingindo com isso um alto grau de independência física. Já no modelo hierarquico tanto as características de estruturas lógicas quanto as físicas são definidas em conjunto, o que aumenta a complexidade e reduz a independência física.

Sobre as estruturas lógicas, é interessante observar que, embora exista muita controvérsia sobre qual dos modelos é melhor para os usuários, não se pode dizer que exista um modelo melhor em todos os sentidos para todas as classes de usuários, devido ao fato de a utilização de banco de dados estar tornando-se abrangente. A arquitetura de um Sistema Gerencial de Banco de Dados deve oferecer recursos para a construção de diversos tipos de interfaces, cada uma voltada para as necessidades de uma classe de usuários. É importante que algumas interfaces possam basear-se num modelo e outras em modelos diferentes, podendo, assim, existir uma coexistência entre esses modelos.

A partir de meados da década de 1980, Sistema Gerencial de Banco de Dados relacionais passaram a ganhar também a plataforma de microcomputadores pessoais, com produtos como as diversas variações de DBase com SQL, Borland Paradox, Lotus Approach e Microsoft Access, dentre outros, além de versões *desktop* dos bancos de dados corporativos.

Atualmente, mais de 90% do mercado de banco de dados corporativo são dominados por produtos relacionais. Grandes fornecedores de Sistema Gerencial de Banco de Dados relacionais, a seguir em ordem alfabética: Borland, Computer Associates, Digital, Gupta, IBM, Informix, Microsoft, Oracle, Progress, Sybase, Tandem e Unisys. Melo (1997).

### 3.3.2.3 O modelo banco de dados “data warehouse”.

Um “data warehouse” ou (armazém de dados) é um sistema de computação utilizado para armazenar informação relativa as atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios e análise de grandes volumes de dados e obtenção de informações estratégicas que podem facilitar a tomada de decisão. O assunto “data warehouse” será tratado em detalhes no próximo capítulo.

## 3.4 Sistemas de gerenciamentos de bancos de dados SQL e segurança dos dados

A IBM montou um grupo de pesquisa conhecido como *System R*, cujo objetivo era desenvolver um projeto para criar um sistema de banco de dados relacional o qual eventualmente se tornaria um produto. Os primeiros protótipos foram utilizados por muitas organizações, tais como MIT Sloan School of Management (uma escola renomada de negócios norte-americana). Novas versões foram testadas com empresas de aviação para rastreamento do manufaturamento de estoque.

Eventualmente, o Sistema R evoluiu para SQL/DS, cuja função é dar suporte à definição, manipulação e controle dos dados em banco de dados relacional, o qual posteriormente tornou-se o DB2. A linguagem criada pelo grupo do Sistema R foi a Structured Query Language (SQL), Linguagem de Consulta Estruturada. Esta linguagem tornou-se um padrão na indústria para bancos de dados relacionais e hoje em dia é um padrão ISO (*International Organization for Standardization*).

Um sistema de gerenciamento de banco de dados (DBMS) é simplesmente o software que permite a uma organização centralizar seus dados e gerenciá-los com eficiência e proporciona acesso a programas, aplicativos e aos dados armazenados. O DBMS serve de interface entre os programas aplicativos e os arquivos físicos de dados. Quando o programa

aplicativo requisita um item de dado, tal como 'remuneração bruta', o DBMS o encontra no banco de dados e o passa ao programa aplicativo. Utilizando arquivos de dados tradicionais, os programadores teriam de especificar o tamanho e o formato de cada elemento de dado utilizado no programa e então informar ao computador onde estaria localizado. Um DBMS elimina a maioria dos comandos de definição de dados encontrados em programas tradicionais.

O DBMS livra o programador ou usuário final da tarefa de entender onde e como os dados estão realmente armazenados, separando as visões lógica e física dos dados. A visão lógica apresenta os dados tais como seriam vistas por usuários ou especialistas da empresa, ao passo que a visão física mostra-os como estão realmente organizados e estruturados nos meios de armazenamento físico. Existe apenas uma visão física dos dados, mas pode haver muitas visões lógicas diferentes. O software de gerenciamento de banco de dados disponibiliza o banco físico para as diferentes visões lógicas apresentadas por vários programas aplicativos. Por exemplo, o programa de benefícios de aposentadoria de um funcionário poderia utilizar uma visão lógica do banco de dados de recursos humanos, que requer apenas o nome do funcionário, seu endereço, número da Previdência Social, plano de pensão e dados de benefícios de aposentadoria.

Um sistema de gerenciamento de banco de dados possui três componentes: uma linguagem de definição de dados; uma linguagem de manipulação de dados e um dicionário de dados.

A **linguagem de definição de dados** é a linguagem formal que os programadores utilizam para especificar o conteúdo e a estrutura do banco de dados. Essa linguagem define cada elemento de dado como aparece no banco antes de ser traduzido para os formatos exigidos pelos programas aplicativos.

A maioria dos DBMS tem uma linguagem especializada, denominada **linguagem de manipulação de dados**, utilizada em conjunto com alguma linguagem de programação para manipular os dados no banco. Essa linguagem contém comandos que permitem aos usuários finais e especialistas em programação buscar dados do banco para atender a requisições de informações e desenvolver aplicações. A linguagem de manipulação de dados mais usada hoje é a **linguagem estruturada de consulta** ou **SQL**. Usuários finais e especialistas em sistemas de informação podem utilizar a SQL como linguagem de consulta interativa para acessar dados dos bancos; os comandos SQL podem ser gravados nos programas aplicativos escritos em linguagens de programação, convencionais.

O terceiro elemento de um DBMS é um **dicionário de dados**. Esse dicionário é um arquivo automatizado ou manual que armazena definições de elementos e características de dados, como utilização, representação física, propriedade, (quem, dentro da organização, é responsável pela manutenção do dado), autorização e segurança. Muitos dicionários podem produzir listagens e relatórios de utilização de dados, grupamentos, localização de programas e assim por diante, portanto o dicionário de dados é a função de apoio ao administrador que completa os requisitos de um Sistema Gerenciador de Banco de Dados. No sentido de permitir o controle centralizado dos dados e como os usuários podem ser informados, sobre os recursos de dados existentes, visando o aproveitamento potencial da integração de dados.

Um DBMS pode reduzir a dependência programa/dados, além dos custos de desenvolvimento e manutenção do programa. Acesso e disponibilidade da informação podem ser aumentados porque usuários e programadores podem executar consultas *ad hoc* de dados no banco de dados. O DBMS permite que a organização centralize o gerenciamento dos dados, sua utilização e segurança.

A IBM sendo a companhia que inventou o conceito original e o padrão SQL, não foi a responsável por produzir o primeiro sistema comercial de banco de dados. O feito foi realizado pela Honeywell Information Systems Inc., cujo sistema foi lançado em junho de 1976. O sistema era baseado em muitos princípios do sistema que a IBM concebeu, mas foi modelado e implementado fora da IBM. Os primeiros sistemas de banco de dados construídos baseado nos padrões SQL começaram a aparecer no início dos anos 80 com a empresa Oracle e depois com a IBM através do SQL/DS, servindo como sistema e repositório de informações de outras empresas.

Estes sistemas somente nasceram a partir da insistência de um jornal técnico em utilizar BNF para SQL e este jornal publicou tal artigo. BNF é o conjunto de sintaxes de linguagem de computador que explica exatamente como cada comando interage com os outros comandos e o que pode ou não ser realizado, como os comandos são formados e assim por diante. Por causa da publicação deste artigo, empresas puderam utilizá-lo para modelar seus próprios sistemas, os quais seriam 100% compatíveis com o sistema da IBM.

O padrão SQL passou da IBM para a ANSI (*American National Standards Institute*) - Instituto Nacional Americano para Padrões, e a ISO, os quais formaram um grupo de trabalho para continuar o desenvolvimento. Este desenvolvimento ainda acontece com outras novas versões dos padrões definidos.

O padrão SQL sempre usa o termo “linha” e “coluna”, nunca usando “registro” e “campo”, de modo que os comandos de “manipulação de dados”, ou seja comandos do SQL

que executam funções de recuperação ou atualização dos dados, podem ser chamados ou interativamente ou dentro de um programa de aplicação. Em geral a chamada interativa significa que o comando em questão é executado de um terminal (video) interativo e no caso de recuperação de dados o resultado é apresentado no terminal. A chamada de dentro de um programa de aplicação significa que o comando é executado como parte de um processo em execução onde o programa e o resultado, no caso de recuperação de dados, são trazidos para uma área de entrada dentro desse programa.

Já verificamos que em bancos de dados relacionais as informações são guardadas em tabelas. Para recuperar uma informação necessária ao usuário, deve-se buscá-la em várias tabelas diferentes, estabelecendo-se um *relacionamento* entre elas. Tabelas são na verdade conjuntos de dados. Por exemplo, quando em um sistema existe uma tabela de vendas, esta tabela corresponde ao conjunto de todas as vendas feitas por uma organização. A tabela de vendedores corresponde ao conjunto de vendedores que trabalham em uma organização. Cada “linha ou registro” da tabela corresponde a um elemento do conjunto.

Uso de índices: quando fazemos consultas em uma tabela estamos selecionando registros com determinadas propriedades. Dentro do conceito de álgebra relacional, estamos fazendo uma simples operação de determinar um subconjunto de um conjunto. A forma trivial de realizar esta operação é avaliar cada um dos elementos do conjunto para determinar se ele possui ou não as propriedades desejadas. Ou seja, avaliar, um a um, todos os seus registros.

Em tabelas grandes, a operação descrita acima pode ser muito custosa. Imaginemos que se deseje relacionar todas as apólices vendidas para um determinado cliente, para saber seu histórico. Se for necessário varrer toda a tabela de apólices para responder esta questão o processo certamente levará muito tempo.

A forma de resolver este problema é o uso de índices. Índices possibilitam ao banco de dados o acesso direto às informações desejadas.

Fisicamente, a tabela não está organizada em nenhuma ordem. Os registros são colocados na tabela na ordem cronológica de inserção. Eliminar linhas nos Bancos de Dados ainda causam mudanças nesta ordem. Um índice é uma estrutura onde todos os elementos de uma tabela estão organizados, em uma estrutura de dados eficiente, ordenados segundo algum critério. Um registro no índice é composto pelo conjunto de valores dos campos que compõem o índice e pelo endereço físico do registro na tabela. Ao escrever uma consulta SQL, não informamos especificamente qual índice será usado pela consulta. Esta decisão é

tomada pelo banco de dados. Cabe a nós escrever a consulta de forma que o uso do índice seja possível.

Alguns exemplos do estilo sintático de comandos para se chamar o SQL de um programa de aplicação:

a) Para definição de dados

O comando **CREATE** é usado na criação de tabelas

Vamos ao exemplo de como criar uma tabela, observe os detalhes em relação aos parênteses, nome das colunas e tipo de dado.

```
Create Table Dept (  
  Deptno      Number(2) Not Null,  
  Dname       Char(14),  
  Loc         Char(13));
```

Construíram-se a tabela Departamento com as colunas departamento, nome departamento e localização.

b) Para manipulação de dados

Após criarmos a tabela vamos inserir dados na mesma, com um comando **INSERT**

```
Insert Into Dept Values (10,'Compras','São Paulo');
```

```
Insert Into Dept Values (20,'Vendas','Sorocaba');
```

Na primeira linha do comando, armazenou-se o valor 10 em código do departamento, “Compras” em departamento e “São Paulo” na localização.

Para eliminarmos linhas da tabela, usamos o comando **DELETE**

```
Delete from Dept where Deptno = 10, eliminando a linha cujo código de departamento é = 10.
```

c) Para consultar os dados

Por exemplo, a consulta:

```
select coluna1  
from tabela  
where coluna1 = 3  
      and  coluna2 > 4  
      and  coluna4 between 10 and 20.
```

Disponibiliza para uso as colunas: coluna1, coluna2 e coluna4. Nos casos das colunas: coluna2 e coluna4, o acesso será para buscar registros que possuam valor numa determinada faixa.

Poderíamos mostrar outros exemplos verificando o uso da função **AVG** que tem a finalidade de trazer a média de uma determinada soma. poderíamos também usar a função **MAX** e **MIN** que tem como função trazer o máximo e o mínimo valor. Também temos a função **SUM** que faz a soma de valores de colunas.

### 3.4.1 Problemas suportados pelos gerenciadores de banco de dados

Imaginemos parte de um banco que mantenha informações sobre todos os clientes e contas de poupança em sistemas de arquivos permanentes no banco. O sistema tem ainda uma quantidade de programas aplicativos que permitem ao usuário manipular os arquivos, incluindo: um programa para debitar ou creditar em uma nova conta; um programa para adicionar uma nova conta; um programa para calcular o saldo de uma conta e um programa para gerar posicionamentos mensais.

Esses programas aplicativos foram desenvolvidos por programadores em resposta às necessidades da organização bancária. Novos programas de aplicação são adicionados ao sistema à medida que as necessidades aparecem. Logo, mais arquivos e mais programas são acrescentados ao sistema. O típico sistema de processamento de arquivos acima é suportado por um sistema operacional convencional. Registros permanentes são guardados em diversos arquivos, e uma série de diferentes programas aplicativos é escrito para extrair e adicionar registros aos arquivos apropriados.

Este esquema tem uma série de desvantagens:

a) Redundância e inconsistência de dados: muitos programadores diferentes e programas implementados em linguagens diferentes podem gerar arquivos de formatos diferentes. Informações podem estar duplicadas em diversos lugares. Gera inconsistência pois estas cópias podem estar com valores diferentes.

b) Dificuldade no acesso aos dados: um diretor deseja a lista de todos os clientes que moram na cidade de CEP 34863. Ou ele extrai manualmente esta informação de uma lista de clientes ou pede a um programador escrever um programa aplicativo. Suponha mais tarde que o mesmo diretor deseje uma lista com os clientes com mais de \$10000. Tal lista não existe e novamente o diretor tem as duas opções.

c) Isolamento dos dados: como os dados estão espalhados, em arquivos separados e com formatos diferentes, é difícil escrever novos programas aplicativos para recuperar os dados adequados.

d) Anomalias de acesso concorrente: para aperfeiçoar o desempenho geral do sistema e obter tempo de resposta mais rápido, deixamos que múltiplos clientes acessem e atualizem os dados simultaneamente. Isso gera dados inconsistentes. Exemplo: dois clientes sacarem dinheiro de uma mesma conta corrente.

e) Problemas de segurança: nem todo usuário do sistema de BD deve ter acesso a todos os dados. Por exemplo: o RH pode ter acesso às informações cadastrais dos clientes, mas não aos valores de conta corrente. Se novos programas aplicativos forem adicionados, é difícil assegurar tais restrições de segurança.

f) Problemas de integridade: os valores dos dados armazenados necessitam satisfazer certas restrições. Por exemplo, o saldo nunca estar abaixo de R\$ 1000,00. Estas restrições podem estar contidas nos programas aplicativos, mas quando novas restrições forem adicionadas, é difícil de alterar estes programas.

Por estes motivos fez-se necessário uma abordagem relacionada à aplicação de um SGBD, a fim de evitar os problemas descritos.

a) Segurança: um SGBD deve, também, prover mecanismos de segurança de acesso para consulta ou atualização dos objetos persistentes. Em geral, estes mecanismos são implementados por meio de comandos de concessão/revogação (*grant/revoke*) de privilégios de acesso a usuários individuais ou grupos de usuários. Dependendo da aplicação, o nível de segurança pode ser de coleções de dados (tabelas e registros ou linhas), atributos isolados (campos ou colunas) e até mesmo de conteúdo (valores de dados).

b) Integridade: um banco de dados deve estar sempre num estado consistente, satisfazendo permanentemente algumas condições de consistência, chamado de restrições de integridade. O SGBD tem a incumbência de garantir a integridade do banco de dados na passagem de um estado para outro, que ocorre ao final de cada transação. Controle de *concorrência de transações*: uma transação de banco de dados é uma seqüência atômica de operações regida pela regra do tudo ou nada; isto é, uma transação é executada inteiramente ou, caso não termine, nenhuma das suas operações é confirmada. Num ambiente típico de execução de um SGBD, as transações são executadas concorrentemente. Para garantir a consistência do banco de dados e das transações o SGBD usa estratégias de controle de concorrência.

c) Desempenho: além de executar eficazmente as funções anteriormente descritas, um SGBD deve executá-las eficientemente. Muitas estruturas de dados, métodos de acesso e técnicas

de otimização que melhoram o desempenho de SGBDs têm sido analisadas extensivamente, tanto em protótipos de pesquisa como em produtos comerciais. Com o advento de novas aplicações, por exemplo, multimídia, novos requisitos de armazenamento e acesso têm surgido. Cabe ao SGBD atender a estes novos requisitos para garantir aos usuários índices satisfatórios de desempenho, como tempo de resposta em consultas e volume de transações processadas.

### 3.5 Conclusões do capítulo

Neste capítulo, verificamos que o avanço das tecnologias de informação é de fundamental importância para as empresas, e pelo volume de informações geradas no ambiente interno e externo das organizações, mostramos a importância do armazenamento de dados e historicamente a evolução da independência dos programas em relação aos dados armazenados.

Verificamos a utilização dos sistemas de gerência dos bancos de dados, que reúne todas as funções necessárias à localização e manipulação dos dados e de interesse dos sistemas de informação, tratando da segurança e integridade dos dados.

Quanto aos processos de recuperação de informação necessárias ao usuário, estudamos sobre uma linguagem de busca chamada SQL e que no capítulo 5, será estendida para buscas com conceitos da “lógica nebulosa”.

## CAPÍTULO 4 – Banco de dados “data warehouse”.

Neste capítulo faremos a apresentação das terminologias básicas de “data warehouse” como também conceitos, importância e características do “data warehouse”, diferenças entre os sistemas transacionais e o modelo multidimensional; considerações sobre os metadados e sua importância, conceitos específicos de “data warehouse”, tais como: volatilidade, granularidade, fatos, dimensão, ciclicidade, “data mining” e também os cuidados na transferências dos dados dos sistemas tradicionais para os “data warehouse”.

Os “data warehouse” surgiram como conceito acadêmico na década de 80. Com o amadurecimento dos sistemas de informação empresariais, as necessidades de análise dos dados cresceram paralelamente. Os sistemas OLTP (*Online Transaction Processing*), usados para armazenar as operações de negócios, não conseguiam cumprir a tarefa de análise com a simples geração de relatórios. Nesse contexto a implementação do “data warehouse” (possibilita a análise de grandes volumes de dados), passou a se tornar realidade nas grandes corporações. O mercado de ferramentas de “data warehouse”, que faz parte do mercado de *Business Intelligence*, cresceu então, e ferramentas melhores e mais sofisticadas foram desenvolvidas para apoiar a estrutura do “data warehouse” e sua utilização. Pela sua capacidade de sumarizar grandes volumes de dados e de possibilitar análises, os “data warehouse”s são atualmente o núcleo dos sistemas de informações gerenciais e apoio à decisão das principais soluções de *business intelligence* do mercado.

O “data warehouse” é comumente acessado através de “data marts”, que são pontos específicos de acesso a sub-conjuntos do “data warehouse”. Os “data marts” são construídos para responder a prováveis perguntas de um tipo específico de usuário. Por exemplo: um Data Mart de Vendas, poderia armazenar informações consolidadas dia-a-dia para um usuário gerencial e em periodicidades maiores (semana, mês, ano) para um usuário no nível da diretoria.

### 4.1 Sistemas transacionais e o modelo dimensional

Sistemas transacionais em uma corporação são projetados para atender a duas características básicas: eficiência na aquisição de informações (alta performance para armazenar as transações executadas, não comprometendo o ritmo operacional da organização) e manutenção da integridade e consistência das informações adquiridas. Dentro destas

necessidades, o modelo de entidade relacionamento, bem como as técnicas tradicionais de modelagem estão solidificadas como ferramentas de suma importância.

Os bancos de dados operacionais armazenam informações necessárias para as operações diárias da organização, e são utilizados por todos os funcionários para registrar e executar operações pré-definidas; por isso seus dados podem sofrer constantes mudanças, conforme as necessidades atuais da organização.

Já o usuário em busca de informações analíticas tem um perfil diferente. Sua atuação não é planejada. Novas necessidades de negócio levam a novas necessidades de análise. Tentar antecipá-las, todas, é inviável. Portanto, nesse ambiente, é muito difícil ter um sistema intermediando o contato entre o usuário e a base de dados, o usuário não precisa ter nenhum conhecimento da base de dados e nem precisa entender como os dados são armazenados. Ele deve sim, analisá-los, saber usar o sistema e também saber de recuperação de informação.

O que é realmente precisamos são sistemas, com uma interface bem-simples e intuitiva, acessíveis ao usuário, e que permitam a navegação e recuperação pela base de dados, permitindo comparação e análise de dados.

Kimball (1997) explica que, para esse fim, o modelo relacional dentro das técnicas tradicionais de modelagem, se mostra inadequado. Ele é excessivamente complexo, confunde o usuário e dificulta a extração de informações úteis. Para esse tipo de aplicação, um modelo conceitual dimensional é bastante mais adequado. Nesse tipo de modelo, trabalha-se com algumas entidades centrais, chamada de tabela de fatos. Esta entidade capta a ocorrência de um evento que diz respeito à atuação da corporação e esse modelo é muito mais compreensível para o usuário, que é capaz de reconhecer a natureza de sua operação no modelo dimensional.

Um modelo dimensional armazena dados analíticos, destinados às necessidades dos tomadores de decisão das entidades no processo de tomada de decisões. Pode-se envolver buscas complexas que necessitam acessar um grande número de registros; por isso, é importante a existência de muitos índices criados para acessar as informações da maneira mais rápida possível. Um modelo dimensional armazena informações históricas de muitos anos e, por isso, deve ter uma grande capacidade de processamento e armazenamento dos dados.

OLAP (On-line Analytical Processing) representa um conjunto de tecnologias projetadas para suportar análise e buscas esporádicas. Sistemas OLAP ajudam analistas e gerentes a sintetizar informações sobre a empresa, através de comparações, visões personalizadas, análise histórica e projeção dos dados em vários cenários de “e se ...”e

oferecem respostas rápidas e consistentes às buscas iterativas executadas pelos usuários, independentemente da complexidade da busca.

A característica principal dos sistemas OLAP é permitir uma visão conceitual e multidimensional dos dados de uma empresa. A modelagem dimensional é a técnica utilizada para se ter uma visão multidimensional dos dados. Nessa técnica, os dados são modelados em uma estrutura dimensional conhecida por cubo. As dimensões do cubo representam os componentes dos negócios da empresa, tais como “cliente”, “produto”, “fornecedor” e “tempo”. A célula resultante da interação das dimensões é chamada de métrica, e geralmente representa dados numéricos como “unidades vendidas”, “lucro” e “total de venda”. Tempo é uma dimensão quase permanente em todas as análises de negócio e se caracteriza por ser seqüencial: um servidor OLAP entende perfeitamente essas seqüências do tempo. Expressões como este mês, último mês, cinco primeiros meses deste ano etc., utilizadas normalmente pelos usuários, são perfeitamente entendidos e interpretados, o que facilita sensivelmente a preparação das aplicações.

Além dos componentes dimensão e medida, outro importante aspecto do modelo multidimensional é a consolidação dos dados, uma vez que, para a tarefa de análise é mais útil e significativa a agregação (ou sumário) aos valores indicativos dos negócios.

#### 4.2 Abordagens para o desenvolvimento de um “data warehouse”

Praticamente todas as organizações utilizam sistemas informatizados para realizar seus processos mais importantes, o que com o passar do tempo, gera uma enorme quantidade de dados relacionados aos negócios, mas não relacionados entre si. Esses dados são um recurso, mas de modo geral, raramente servem para o uso estratégico no seu estado original. Os sistemas convencionais de informática não são projetados para gerar e armazenar informações estratégicas, o que os torna sem valor para o apoio ao processo de tomada de decisões.

Analisar informações para tomada de decisão não é uma atividade recente, mas sim a tecnologia de suporte a esse processo. “data warehouse” é uma evolução dos sistemas de suporte a decisão que surgiram a partir das facilidades providas pelas tecnologias de banco de dados.

“data warehouse” é construído para que tais dados possam ser armazenados e acessados, de forma que não sejam limitados por tabelas e linhas estritamente relacionais. A função do “data warehouse” é tornar as informações corporativas acessíveis para o seu entendimento, gerenciamento e uso. O “data warehouse” deverá estar separado dos bancos de

dados operacionais; as buscas efetuadas pelos usuários não interferem nesses sistemas, que ficam resguardados de alterações indevidas ou perdas dos dados.

Em termos simples, um “data warehouse” é um armazém de dados, e pode ser definido como um banco de dados especializado, o qual se integra e gerencia o fluxo de informações, a partir dos bancos de dados corporativos e fontes de dados externos às instituições.

Um “data warehouse” oferece os fundamentos e os recursos necessários para um sistema de apoio a decisão eficiente, disponibilizando dados integrados e dados históricos que servem desde o corpo diretivo da universidade, que necessita de informações mais resumidas, até os coordenadores de cursos, quando os dados detalhados ajudam a observar aspectos táticos da universidade. Nele o nível estratégico pode obter de modo imediato, respostas para perguntas que normalmente não possuem retorno em seus sistemas convencionais e, com isso, tomar decisões com base em fatos, não em mera intuição ou especulação.

#### 4.2.1 Conceitos básicos, definições e importância do “data warehouse”

Nos dias atuais, o acesso aos dados tem-se tornado um dos desafios mais críticos no mundo dos negócios. Informações estratégicas nas mãos daqueles que são responsáveis por tomadas de decisão permitem a criação e manutenção de uma maior vantagem competitiva por parte das organizações.

Com o surgimento do “data warehouse” são necessários novos métodos de estruturação de dados e nova tecnologia, tanto para armazenamento como para recuperação de informações. A necessidade desses novos métodos e tecnologia surgiu a partir de duas constatações: 1) existe uma necessidade de informação não atendida pelos aplicativos comerciais convencionais, que atuam em nível operacional do negócio; 2) a tecnologia de armazenamento de dados utilizada nesses aplicativos não atende às necessidades. Graças aos avanços tecnológicos e às novas estruturas dos bancos de dados, a tecnologia da informação pode permitir que organizações possam elaborar um “data warehouse” próprio.

A seguir algumas definições pesquisadas sobre “data warehouse”:

a) Segundo Inmon (1997), “data warehouse” é uma coleção de dados orientados por assuntos integrados, variáveis com o tempo e não voláteis, para dar suporte a um processo gerencial de tomada de decisão.

b) Conforme Harjinder e Rao (1996), “data warehouse” é um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para

atender à necessidade de buscas estruturadas e esporádicas, relatórios analíticos e de suporte a decisão.

c) Segundo Barquini, citado por Dal'Alba (1998), “data warehouse” é uma coleção de técnicas e tecnologias que juntas, disponibilizam um enfoque pragmático e sistemático para tratar do problema com o usuário final que irá acessar informações que estão distribuídas em vários sistemas da organização.

O “data warehouse” é um banco de dados que contém dados extraídos do ambiente de produção da empresa, que foram selecionados e depurados, tendo sido otimizados para processamento de busca e não para processamento de transações. Em geral, um “data warehouse” requer a consolidação de outros recursos de dados, além dos armazenados em banco de dados relacionais, inclusive informações provenientes de planilhas eletrônicas, documentos textuais, etc.

O modelo de dados transacional, que tem como objetivo construir e integrar todos os dados das empresas surgiu no início da década de 80, mas foi abandonado por pensarem que ele apenas atrapalhava o desenvolvimento rápido de aplicações. Apesar disto, muitas outras empresas vislumbraram o valor desse tipo de modelo e notaram que poderia servir como uma arquitetura comum através da qual novas aplicações poderiam ser projetadas, integrando-se às antigas. Apesar da existência desse modelo, que muitas vezes facilita a integração dos dados, nota-se que não é o mais adequado para aplicações que tenham como objetivo suporte a decisão, devido às características específicas para o mundo transacional. Têm-se utilizado, então, os modelos multidimensionais, também chamados de modelo estrela, para suportar os sistemas de processamento analítico.

A tecnologia de “data warehouse”, portanto, surgiu como uma necessidade para superar as limitações dos sistemas OLTP, cujos dados armazenados e coletados não estão disponíveis aos usuários finais de forma a permitir posteriores análises.

A idéia do “data warehouse” é integrar os dados dos diversos sistemas OLTP em uma estrutura única, através de processos de extração, compatibilização das diferenças semânticas e consolidação dos dados, o que torna possível o acesso às informações estratégicas.

Como se vê, existem diferentes visões sobre o que seria um “data warehouse”: uma arquitetura, um conjunto de dados semanticamente consistente, com o objetivo de atender a diferentes necessidades de acesso a dados e extração de relatórios, ou ainda, um processo em constante evolução, que utiliza dados de diversas fontes heterogêneas, para dar suporte a buscas esporádicas, relatórios analíticos e à tomada de decisão.

É importante considerar, no entanto, que um “data warehouse” não contém apenas dados resumidos, mas pode conter também dados primitivos. É desejável prover ao usuário a capacidade de aprofundar-se num determinado tópico, investigando níveis de agregação menores ou mesmo o dado primitivo, permitindo também a geração de novas agregações ou correlações com outras variáveis.

Além do mais, é extremamente difícil prever todos os possíveis dados resumidos que serão necessários: limitar o conteúdo de um “data warehouse” apenas a dados resumidos significa limitar os usuários apenas às buscas e análises que eles puderem antecipar frente a seus requisitos atuais, não deixando qualquer flexibilidade para novas necessidades.

#### 4.2.2 Características do “data warehouse”

Um “data warehouse” pode armazenar grandes quantidades de informação, às vezes divididas em unidades lógicas menores que são chamadas de “data marts”. O esquema de dados mais utilizado é o “Star Schema”, também conhecido como Modelagem multidimensional. Geralmente o “data warehouse” não armazena informações sobre os processos correntes de uma única atividade de negócio e sim cruzamentos e consolidações de várias unidades de negócios de uma empresa. Ele pode ser:

a) Orientado por temas

Refere-se ao fato de o “data warehouse” armazenar informações sobre temas específicos e importantes para o negócio da empresa. Exemplos típicos de temas são: produtos, atividades, contas, clientes, etc. Em contrapartida, o ambiente operacional é organizado por aplicações funcionais.

b) Integrado

Refere-se à consistência de nomes, das unidades das variáveis, etc., no sentido de que os dados foram transformados até um estado uniforme. Por exemplo, considere-se sexo como um elemento de dado. Uma aplicação pode codificar sexo como M/F, outra como 1/0 e uma terceira como H/M. Conforme os dados são trazidos para o “data warehouse”, eles são convertidos para um estado uniforme, ou seja, sexo é codificado apenas de uma forma; ele será convertido em uma representação única, ao ser colocado no “data warehouse”.

c) Variante no tempo

Refere-se ao fato de o dado em um “data warehouse” referir-se a algum momento específico, significando que ele não é atualizável, enquanto o dado de produção é atualizado

de acordo com mudanças de estado do objeto em questão, refletindo, em geral, o estado do objeto no momento do acesso. Em um “data warehouse”, a cada ocorrência de uma mudança, uma nova entrada é criada, para marcar essa mudança.

d) Não volátil

Significa que o “data warehouse” permite apenas a carga inicial dos dados e sua busca, o chamado ambiente *“load-and-access”*. Após serem integrados e transformados em um estado padrão, os dados são carregados em bloco para o “data warehouse”, para que estejam disponíveis ao acesso dos usuários. No ambiente operacional, ao contrário, os dados são, em geral, atualizados registro a registro, em múltiplas transações.

#### 4.2.3 Arquitetura genérica do “data warehouse”

Uma arquitetura genérica e seus componentes proposta por Campos & Rocha (98) será descrita, a seguir. Note-se que essa arquitetura procura apenas sistematizar papéis no ambiente de “data warehouse” e permitir que as diferentes abordagens encontradas no mercado atualmente possam se enquadrar.

Nos últimos anos, o conceito de “data warehouse” evoluiu rapidamente de um considerável conjunto de idéias relacionadas para uma arquitetura voltada para a extração de informação especializada e derivada a partir dos dados operacionais da empresa. O estudo de uma arquitetura que descreva o ambiente de “data warehouse” permite compreender melhor a estrutura geral de armazenamento, integração, comunicação, processamento e apresentação dos dados que servirão para subsidiar o processo de tomada de decisão nas empresas.

Camada de bancos de dados operacionais e fontes externas correspondem aos dados das bases de dados operacionais da organização, junto a dados provenientes de outras fontes externas, que serão tratados e integrados para compor o “data warehouse”.

Camada de Metadados é a informação sobre os dados mantidos pela empresa, informação em um diagrama Entidade-Relacionamento, (dados em um dicionário de dados são exemplos de metadados). Para poder manter a funcionalidade de um ambiente de “data warehouse” é necessário ter disponível uma grande variedade de metadados, desde dados sobre as visões dos usuários até dados sobre os bancos de dados operacionais. Idealmente, o usuário deve poder ter acesso aos dados de um “data warehouse” sem que tenha que saber onde residem estes dados ou a forma como estão armazenados. Características dos metadados associados a um “data warehouse” serão discutidas com mais detalhes no próximo item.

A Camada de gerenciamento de processos está envolvida com o controle das diversas tarefas a serem realizadas para construir e manter as informações do dicionário de dados e do “data warehouse”. Ela é responsável pelo gerenciamento dos processos que contribuem para manter o “data warehouse” atualizado e consistente.

Camada de transporte: controla o transporte de informações pelo ambiente de redes. É usada para isolar aplicações, operacionais ou informacionais, do formato real dos dados nas duas extremidades. Também inclui a coleta de mensagens e transações e se encarrega de entregá-las em locais e tempos determinados.

Camada do “data warehouse” ou o “data warehouse” propriamente dito: corresponde aos dados usados para fins "informacionais". Em alguns casos, “data warehouse” é simplesmente uma visão lógica ou virtual dos dados, podendo de fato não envolver o armazenamento desses dados. Em um “data warehouse” que exista fisicamente, cópias dos dados operacionais e externos são de fato armazenadas, de modo a permitir um fácil acesso e alta flexibilidade de manipulação.

Camada de gerenciamento de replicação: inclui todos os processos necessários para selecionar, editar, resumir, combinar e carregar o “data warehouse” e as correspondentes informações de acesso, a partir das bases operacionais e fontes externas. Normalmente isso envolve programação complexa, mas cada vez mais são disponibilizadas ferramentas para facilitar tais processos. Essa camada pode também envolver programas de análise da qualidade dos dados e filtros que identificam padrões nos dados operacionais.

#### 4.2.4 Metadados e sua importância

Diferentes aplicações desenvolvidas em tempos diferentes no âmbito operacional da empresa, geralmente, contêm dados que são inconsistentes ou redundantes. A menos que seja guiado pelos princípios de uma administração de dados efetiva, um “data warehouse” não atingirá seu objetivo de integração dos dados. Os metadados constituem-se no principal recurso para a administração de dados e assumem maior importância ainda no ambiente de “data warehouse”.

Metadados para Inmon (1997) são normalmente, definidos como "dados sobre os dados". Ou seja, metadado é uma abstração dos dados, ou ainda, dados de mais alto nível que descrevem dados de um nível inferior. Sem metadados, os dados não têm significado. São exemplos de metadados as descrições de registros em um programa de aplicação ou o

esquema de um banco de dados descrito em seu catálogo, ou ainda, as informações contidas em um dicionário de dados.

Em um ambiente operacional, os metadados são especialmente valiosos para os desenvolvedores de aplicação e os administradores do banco de dados. Os bancos de dados operacionais são usualmente utilizados via aplicações, que já contêm as definições de dados embutidas. Seus usuários simplesmente interagem com as telas do sistema, sem precisar conhecer como os dados são mantidos pelo banco de dados.

O ambiente de suporte da decisão, por sua vez, é bastante distinto. Nele analistas de dados e tomadores de decisão procuram por fatos não usuais e correlações que serão reconhecidas, quando encontradas. Aplicações rotineiras e pré-definidas não fazem sentido nesse ambiente.

#### 4.2.5 Granularidade

Granularidade diz respeito ao nível de detalhe ou de resumo contido nas unidades de dados existentes no “data warehouse”. Quanto maior o nível de detalhes, menor os níveis de granularidade. O nível de granularidade afeta diretamente o volume de dados armazenados e, ao mesmo tempo, o tipo de busca que pode ser respondida.

Exemplo de níveis de granularidade:

Baixa				Alta			
produto	data	quantidade	valor	mês/ano	produto	quantidade	valor
A1	28/05/04	10	100,00	05/04	A1	120	1.200,00
B1	29/05/04	45	200,00	05/04	B1	45	200,00
A1	29/05/04	110	1.100,00				

Figura 4.01: Granularidade

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

#### 4.2.6 Dimensões

Os bancos de dados multidimensionais simulam um cubo com "n" dimensões. A análise multidimensional representa os dados como dimensões, em vez de tabelas. Combinando-se essas dimensões, o usuário tem uma ótima visão da empresa e pode efetuar ações comuns como *"slice and dice"*, que consiste na mudança das dimensões a serem visualizadas e *"drill-down/up"* onde é a navegação entre os diversos níveis do detalhamento, onde "drill-down" em um "data warehouse" significa adicionar cabeçalho de linha de tabelas de dimensão, e "drill-up" subtrair cabeçalhos de linha.

Dimensão é uma reunião de membros<sup>1</sup>, todos eles de um tipo semelhante; cidades, regiões e países formariam a dimensão (localização geográfica). As dimensões representam as possíveis formas de visualizar os dados; são as quebras de níveis comparadas às base de dados operacionais, como também as variáveis e os valores que são sumarizados.

As dimensões para Kimball são a guia da base de dados; as plataformas para se percorrer os valores-limite permitidos e a aplicação dessas limitações, e são a fonte dos cabeçalhos de linhas de relatórios finais dos usuários; e elas contêm o vocabulário da empresa para os usuários.

O primeiro contato com as dimensões ocorre quando se avalia a granularidade das tabelas de fatos de um "data warehouse". Ao se estabelecer a granularidade, defini-se as informações pretendidas e a sua forma de apresentação.

Dimensões são possíveis formas de se visualizar os dados. São os "por" dos dados; exemplificando: "por mês", "por país", "por produto" e devem ser escolhidas, tendo-se em mente o "data warehouse" a longo prazo. Esta opção apresenta o momento inicial, quando o arquiteto da base de dados deve desviar sua atenção dos detalhes desta base e analisar os planos a longo prazo.

Um conjunto de dimensões bem arquitetado faz com que a base de dados seja compreensível e de fácil utilização.

Intuitivamente, o usuário pode identificar cruzamento de dados que lhe interessam. Por exemplo, é preciso visualizar as pessoas por sexo, por faixa etária e por tempo de estudo. Essa necessidade identifica três dimensões: sexo, faixa etária e tempo de estudo.

Um "data warehouse" sempre pressupõe o acompanhamento dos dados através do tempo. No caso da dimensão tempo, é útil manter alguns atributos para documentar feriados e diferentes marcos de períodos significativos, tais como trimestre, quadrimestre, semestre, etc.

---

<sup>1</sup> Membro é um dado dentro de uma dimensão; ex. janeiro de 99, primeiro trimestre de 98 (dimensão tempo); lojas, distribuidores, vendedores (dimensão canais de distribuição).

#### 4.2.7 Tabela de fatos

Tabela de fatos<sup>2</sup> é a tabela central, considerada a tabela que interliga as dimensões. Nela são armazenadas medições numéricas<sup>3</sup>, em que cada medida é obtida na interseção de todas as dimensões.

#### 4.2.8 Data Mining

Para Fernando Numer e Luiz Carlos Spandri (Dvelopers, fev. de 1998), a garimpagem de dados ou "Data Mining" é o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma organização, através da análise de grandes quantidades de dados armazenados em seus bancos de dados, usando técnicas de reconhecimento de padrões estatísticas e matemáticas.

Esse processo trabalha no modo de descoberta indutiva, ou seja, os dados são analisados através de um conjunto de algoritmos e critérios especificados, possibilitando o cruzamento de informações na base de dados, com respostas impensáveis, em que o usuário final formula uma hipótese sobre o seu negócio e, então, utiliza ferramentas de mineração para aprovar ou desaprovar essa hipótese.

A expressão "*data mining*", até certa época, era usada em textos relacionados com lógica, inteligência artificial e redes neurais (software de rede neural é um modelo matemático computacional que abstrai o modo como as células cerebrais, os neurônios operam, isto é, aprender através da experiência, desenvolver regras e reconhecer padrões), mas hoje é mencionada com o assunto banco de dados.

Algumas aplicações em potencial para a tecnologia de "data mining" em algumas áreas são: bancos (previsão de níveis de maus empréstimos e uso fraudulento de cartões de crédito); marketing (previsão de quais clientes responderão à mala direta ou vão comprar um produto em particular); manufatura, vendas no varejo (previsão de vendas, determinação de níveis confiáveis de estoque); manufatura, produção (previsão de falhas em máquinas); seguros (previsão de coberturas médicas) e hardware e software (previsão de falhas e necessidades de manutenção).

---

<sup>2</sup> Fato é uma medição, normalmente numérica e aditiva

<sup>3</sup> Medições numéricas ou fatos mensuráveis constituem: número de unidades vendidas, números de dólares (moeda) vendidos, custo estendido (custo unitário x quantidade).

### 4.3 Modelo de dados para, o “data warehouse”

Para (Oliveira, Djalma (1998)) poucas são as empresas que investem em soluções para municiar a informática na tarefa de construção e manutenção de ambientes das diversas bases de dados, a fim de suportar os processos de extração, limpeza, regras de negócio e sintetizar os dados operacionais, transformando-os em informação útil para o processo decisório.

Em princípio, um “data warehouse” provê um banco de dados especializado que gerencia o fluxo de informações dos bancos de dados corporativos e fontes de dados externas à empresa, para as aplicações do usuário final. A função do “data warehouse” é tornar as informações corporativas acessíveis para entendimento, gerência e uso. Os dados operacionais e de fontes externas armazenados no “data warehouse” devem ser tratados, a fim de se eliminarem inconsistências e serem integrados num banco de dados orientado a assuntos específicos.

Como todo “data warehouse” depende de dados disponíveis nas bases de dados operacionais da organização, uma das preocupações para a sua construção é o mapeamento dessas bases e a sua conciliação em um único ambiente, para que o administrador possa assegurar-se da qualidade dos dados utilizados e fazer o gerenciamento de dados do projeto de "warehousing".

Em relação às considerações iniciais, verifica-se que o componente mais importante de um sistema de “data warehouse” é o servidor de banco de dados relacional, usado para armazenar vasta quantidade de informações, rápidas e confiáveis, com a finalidade de responder a uma grande quantidade de questões de negócios.

#### 4.3.1 Tratando as informações para o “data warehouse”

Os dados introduzidos em um “data warehouse” geralmente passam por uma área conhecida como área de stage. O stage de dados ocorre quando existem processos periódicos de leitura de dados de fontes como sistemas OLTP. Os dados podem passar então por um processo de qualidade, desnormalização e gravação dos dados no “data warehouse”.

Na arquitetura do “data warehouse”, ao representar a estrutura de dados, deve-se analisar a camada externa da base de dados transacional, que é representada por todos os dados operacionais, que são os resultados das transações efetuadas diariamente pela organização.

Como essas bases de dados são muitas vezes antigas e projetadas para trabalhar em nível transacional, não suportam o acesso dos dados para fins de análise. O objetivo do “data warehouse” é libertar as informações que estão presas na base de dados transacionais e juntá-las às informações de outras fontes de dados da própria organização ou, mesmo, de fontes externas.

Os sistemas de “data warehouse” revitalizam os sistemas da empresa, porque:

permitem que sistemas mais antigos continuem em operação;

consolidam dados inconsistentes dos sistemas mais antigos em conjuntos coerentes;

c) extraem benefícios de novas informações oriundas das operações correntes;

d) provêem ambiente para o planejamento e arquitetura de novos sistemas de cunho operacional.

O tratamento de informações consiste em recuperar o que realmente interessa ao “data warehouse”, portanto, deve-se iniciar a extração das informações das bases de dados existentes, com a preocupação da transformação, limpeza, observando as padronizações, integração dos dados de diferentes aplicações.

Ao extrair dados operacionais e levá-los ao “data warehouse”, deve se levar em consideração alguns critérios, conforme Soares (1998) e Dal'Alba (1998):

Efetuar a condensação dos dados, pois o volume de dados contido no “data warehouse” poderá ficar excessivamente grande para ser controlado, dificultando o gerenciamento e degradando o seu desempenho;

A não extração de informações, quando da sua carga no “data warehouse” pode levar a análises incorretas; tal problema ocorre em função de as informações operativas estarem constantemente sendo atualizadas (alta volatilidade), o que dificulta a interpretação dos resultados de buscas;

Ao pesquisar dados que apresentem marcas de tempo, é necessário que as aplicações assinalem o momento da última alteração ou atualização em um registro para que, ao ser executada a varredura para o “data warehouse”, só sejam examinados os registros que tenham a data de atualização maior ou igual do que a data da última pesquisa;

Nesta fase deve-se avaliar a necessidade de adicionar atributos que sejam calculados a partir de outros, da própria dimensão;

O “data warehouse”, ao contrário de modelos operacionais, não deve ser normalizado<sup>4</sup>. Sempre que possível, as informações devem estar agregadas, para facilitar o entendimento do

---

<sup>4</sup> Normalizar: Date(86) processo usado para eliminar a redundância de dados.

usuário final, e a desnormalização reduz a quantidade de ligações necessárias para elaboração de buscas;

Para atingir um grau de segurança na confiabilidade dos dados, devemos catalogar o Metadados como plataforma, fontes de dados, tabelas, campos, índices, chaves primárias, chaves estrangeiras, parâmetros e programas.

Como metadados são utilizados normalmente como um dicionário de informações, devem incluir:

Origem de dados: pois todo elemento de dado precisa identificar, sua origem ou o processo que o gera. Esta informação deve ser única, ou seja, cada dado deve ter uma fonte de origem;

Formato de dados: todo elemento de dados deve ter identificado seu tamanho e tipo;

Definições de negócios: cada elemento de dados deve ser suportado por uma definição do mesmo no contexto da área de negócio;

Buscas: é importante que o usuário seja capaz de também compreender os dados, e para tal, informações gerais sobre as suas derivações precisam estar disponíveis;

Tabelas: nome e identificador de cada coluna, critérios de classificação.

#### 4.4 O nível de abrangências da base de dados no “data warehouse”

No modelo dimensional, a estruturação dos dados tem a característica de “cubo de dados” (nome dado a um banco de dados dimensional), pois leva-se em consideração uma tabela dominante central com múltiplas junções, relacionando-se às outras tabelas. Cada uma das tabelas secundárias possui apenas uma junção com a tabela central.

O modelo dimensional permite às pessoas pensar sobre o negócio em termos simples de dimensões como tempo, mercado e produto.

Pensar no negócio em termos de um cubo com dimensões formando a base do cubo, o ponto de interseção das três dimensões dentro do cubo equivale a um ponto de medição para o negócio.

Lembre-se que a "tabela de fatos" é a grande tabela central do projeto dimensional, que possui uma chave de múltiplas partes, onde cada um dos componentes é uma chave externa para uma tabela de dimensão individual.

Para definir o nível de abrangências do “data warehouse”, deve-se pensar na sua granularidade, que se refere ao nível de detalhe em que as unidades de dados são mantidas.

Quanto maior o nível de detalhes, menor o nível de granularidade. Esta é uma questão fundamental no projeto, porque afeta diretamente o volume de dados armazenados no “data warehouse”, e ao mesmo tempo, o tipo de busca que pode ser respondida.

Conclui-se que o “data warehouse” pode ser considerado como uma estrutura que cresce, e cresce em direções que não podem ser previstas. Se forem enfocadas exigências funcionais atuais, será uma solução altamente aperfeiçoada para os problemas de hoje. Porém, como os problemas mudam, a solução pode ficar cada vez menos ótima e, conseqüentemente, ficará inaceitável pelo alto volume de dados. Um caminho melhor é trocar otimização por flexibilidade. A solução restante poderá não ser a ideal, mas será mais adaptável e, conseqüentemente, terá um tempo de vida maior.

Quando há a análise de volumes de dados, tem-se a preocupação, no “data warehouse”, de satisfazer os requisitos às respostas pretendidas, como também quanto a sua eficiência.

Com relação ao volume de dados devemos levar em consideração a renovação do “data warehouse”, cujo primeiro passo é a leitura da base de dados herdados, e Inmon(97) sugere o método mais atraente para essa renovação, que consiste em capturar os dados no ambiente herdado à medida que eles são atualizados e ao capturar os dados durante a atualização, quando chegar o momento de renovar o “data warehouse”, não há necessidade de fazer as varreduras completas das tabelas..

Para (Oliveira, Adelize, (1998)) existem duas preocupações em função da escolha dos níveis de granularidade e do volume de dados são: os "backups" e sua recuperação e também a eliminação periódica de dados, pois dados encontrados nesse ambiente servem como históricos e apresentam, necessariamente, uma baixa probabilidade de acesso, ou chegam a zero; portanto, eles devem ser eliminados e este aspecto é importante, pois irá influir na velocidade de acesso, nos "backups" e recuperação (tempo menor para efetuá-los).

Portanto, a granularidade da tabela central, chamada também de tabela de fatos não deve chegar a níveis muito detalhados, a ponto de extrapolar a capacidade de armazenamento do “data warehouse”, o que sugere alguns cálculos. Parte-se de um valor bruto do somatório de um fato, em um dado período, e de um valor médio daquele mesmo fato na granularidade que se deseja, para obter o número de registros. Divide-se o primeiro pelo segundo, para obter o número de registros que constituirá a tabela de fatos. Multiplica-se este número pelo total de períodos que se pretende manter. Para saber o valor em bytes, é preciso estimar o tamanho de cada registro e multiplicá-lo pelo total de registros.

#### 4.5 Ciclicidade dos dados

Na Ciência da Informação, conforme Robredo (2003), a ciclicidade dos dados é representada pela informações difundidas, que depois de um certo tempo não estarão mais disponíveis. Já no modelo “data warehouse”, ela refere-se ao tempo necessário para uma alteração sobre dados do ambiente operacional.

Inmon (1997) lembra que dados que são alterados refletem-se no ambiente operacional, o mais breve possível. Uma vez que os dados tenham sido refletidos no ambiente operacional, as alterações precisam ser passadas para o “data warehouse” colocando-se como questão: em quanto tempo esse ajuste do “data warehouse” deveria ser feito ? Como regra, pelo menos 24 horas deveriam passar-se entre um momento em que a alteração é observada pelo ambiente operacional e sua repercussão no “data warehouse”. Ele afirma que não deverá haver pressa para tentar passar as alterações para o warehouse, o mais rápido possível. E explica várias razões para a necessidade de colocar uma "dobra de tempo" nos dados.

A primeira razão consiste no fato de que quanto mais rigidamente o ambiente operacional for emparelhado com “data warehouse”, mais dispendiosa e complexa será a tecnologia. Uma dobra de tempo de 24 horas pode facilmente ser alcançada com tecnologia convencional.

#### 4.6 Escolha da duração do banco de dados

A duração mede o alcance retroativo no tempo da tabela de fatos. Em muitas empresas há a necessidade natural de analisar o mesmo período um ano antes. Geralmente essa necessidade exige, pelo menos, cinco trimestres de dados. Próximo ao final do ano normal, significa dois anos de dados. Esses dados podem ser repetidos com menos intensidade para dados com dois anos atrás, três anos, e assim por diante.

As diversas organizações que precisam de relatórios financeiros históricos podem definir durações mais longas para as tabelas de fatos, muitas vezes, abrangendo sete ou mais anos anteriores.

Essas tabelas de fatos de longa duração apontam, pelo menos, dois problemas no projeto de “data warehouse”. Em primeiro lugar, é cada vez mais difícil obter dados retroativos. Quanto mais antigos os dados, maior a probabilidade de que haja problemas na leitura e na interpretação de arquivos ou fita. Em segundo, é obrigatória a utilização das versões antigas das dimensões mais importantes, e não as versões mais atuais. As descrições adequadas do produto antigo e do cliente antigo devem ser usadas junto com o histórico da

transação correspondente. Muitas vezes, o “data warehouse” deve atribuir uma chave genérica para essas dimensões importantes, a fim de distinguir diversas amostras de clientes e produtos em um período qualquer de tempo.

#### 4.7 Busca ao “data warehouse”

O sucesso do “data warehouse” depende em muito da ferramenta de busca disponível. Se os usuários não puderem obter respostas para as questões importantes do negócio, os dados armazenados se tornarão informações inúteis.

Mesmo sabendo que a informação se encontra de alguma forma armazenada no “data warehouse” da organização, ainda pode existir um longo caminho a ser percorrido até que essa informação esteja de fato disponível. Segundo Campos e Rocha (1998), a “pesquisa” eficaz, de modo a poder subsidiar decisões, depende da existência de ferramentas especializadas que permitam a captura de dados relevantes mais rapidamente e sua visualização através de várias dimensões.

Os sistemas não devem apenas permitir acesso aos dados, mas também permitir análises de dados, de maneira que transformem dados brutos em informação útil para a empresa. Não se pode esquecer de que o usuário em busca de informações analíticas tem um perfil diferente, sua atuação nem sempre é planejada e novas necessidades de negócio levam a novas necessidades de análise. Tentar antecipá-las é o maior objetivo.

#### 4.8 Segurança das informações

Fraudes eletrônicas, vingança de funcionários, acessos não autorizados, quebra de sigilo da informação, desastres e erros são ocorrências cada vez mais frequentes nas empresas.

Atualmente, a informação é um fator crítico para a efetivação do negócio de praticamente todas as organizações. Uma indisponibilidade ou perda de dados acarreta prejuízos e, dependendo do tipo de negócio, podem tirar a empresa do mercado, portanto é importante que esteja claro para todos o objetivo da segurança: garantir a disponibilidade, integridade, sigilo e legalidade da informação, ponto chave para o desenvolvimento da empresa.

A informação é um bem da organização e tem valor essencial. Ela tem valor para a organização, para seus concorrentes, para os funcionários insatisfeitos, para os ladrões eletrônicos. Portanto, ela deve ser alvo de um processo de segurança compatível com seu

porte. O recurso à informação precisa ter um processo contínuo que controle os acessos dos usuários, que descreva regras de utilização, que estabeleça responsáveis e que possa ter todos esses procedimentos auditados periodicamente.

Com o objetivo de contemplar todos os aspectos e facilitar a implementação desse processo, Edison Fontes (Revista Developers Magazine, fevereiro de 1998) definiu alguns aspectos de segurança da informação.

**Organização:** devem ser contempladas as normas, as políticas, os termos de compromisso e as auditorias, onde a organização explicitará o seu desejo e sua filosofia em relação ao recurso informação;

**Controle de acesso à informação:** aqui se define como cada usuário se identificará para ter acesso ao ambiente computacional e como ele será autenticado; também deverá ser definido quem será o gestor, que é o responsável pela liberação dessa informação para todos os usuários;

**Recurso humano:** esse aspecto é fundamental. De nada adiantará implementarmos todos os aspectos anteriores se o funcionário que trata com a informação não tiver conscientização sobre segurança da informação. A assinatura de um termo de responsabilidade e uma campanha de conscientização dos usuários são medidas práticas que devem ser adotadas. Ao analisar o processo de segurança através desses aspectos, está se equacionando de forma profícua o problema com mais condições de dimensionar a tarefa de implementação deste processo.

Não se deve esquecer dos fatores críticos de sucesso: continuidade, disponibilização de recursos e conscientização dos usuários. Qualquer que seja a empresa, ela deve possuir seu plano de segurança, que será compatível com seu porte organizacional, com o seu negócio e com a criticidade da informação para esse negócio.

Para que a segurança da informação aconteça, precisa-se que as pessoas que compõem a empresa deixem de estar apenas "envolvidas com a segurança" para estarem "comprometidas com a segurança".

#### 4.9 Conclusões do capítulo

Grande parte das informações que as empresas necessitam para o seu processo decisório existe e está sendo de alguma forma ofertada, mas a dificuldade de obtê-las provavelmente decorre do fato de que seu formato esteja distante das condições de percepção e assimilação dos usuários nas organizações. Estudamos neste capítulo que Sistemas

transacionais em uma organização são projetados para atender a duas características básicas: eficiência na aquisição de informações e manutenção da integridade e consistência das informações adquiridas. Estudamos também que os bancos de dados operacionais armazenam informações necessárias para as operações diárias da organização e seus dados sofrem constantes mudanças, dificultando as análises para tomadas de decisão. O tomador de decisão em busca de informações analíticas tem um perfil diferente, como sua atuação não é planejada, novas necessidades de negócios levam a novas necessidades de análise.

Ao estudarmos as características do “data warehouse” mostramos que esta técnica de armazenamento de dados permite a visualização multi-dimensional do universo da organização e para os usuários, esta é uma das maneiras mais natural, e segura de trabalhar com a recuperação de informações para tomada de decisão. O “data warehouse” é voltado para o processamento de consultas e análise de tendências, devido a sua não volatilidade e a possibilidade de agregar dados históricos.

O que é realmente precisamos são sistemas, com uma interface amigável e intuitiva, acessíveis ao usuário, e que permitam a navegação e recuperação pela base de dados, permitindo comparação e análise de dados, estudo esse que será complementado no capítulo 6 juntando-se conhecimentos adquiridos sobre “data warehouse” e de “lógica nebulosa” que será apresentada no capítulo 5.

## CAPÍTULO 5 – As lógicas: Booleana e Nebulosa como estratégia de recuperação de informação.

Neste capítulo serão apresentados as terminologias básicas de “lógica nebulosa”, os conceitos e suas aplicações. Características e sua importância para ajudar à tomada de decisão. Abordaremos o posicionamento da “lógica nebulosa” no contexto da Computação natural. Comparações das lógicas booleana e nebulosa e estratégias de recuperação em um banco de dados multidimensional com o conceito de “lógica nebulosa”.

### 5.1 Introdução

Se analisarmos um dicionário ou uma gramática, descobriremos que uma linguagem é definida como um sistema de símbolos que serve como meio de comunicação. Nota-se que isso não se restringe à comunicação entre humanos; hoje em dia existem dezenas de linguagens de programação que, poderíamos dizer, servem também para comunicar instruções de um humano a uma máquina. Estas seriam exemplos de linguagens artificiais, ao contrário do português, inglês, e assim por diante, que são chamadas linguagens naturais ou línguas. Sendo a linguagem um instrumento muito complexo, seu emprego, bem como a sua interpretação podem ser causas de certos erros. Uma palavra pode representar uma ampla classe de coisas ou acontecimentos físicos, levando dessa forma, a ter vários significados a depender do contexto.

Lidar com situações que envolvem fatores como ambigüidade, incerteza e informações vagas na resolução de problemas é característica do pensamento humano, que usa o seu conhecimento, experiências do cotidiano, intuição e até mesmo senso comum para lidar com esses fatores. Nos problemas de difícil solução, em que se faz necessário o auxílio matemático computacional, modelar tais fatores é extremamente complicado, principalmente em relação ao armazenamento e recuperação das informações.

Os bancos de dados sempre possuíram grande importância dentro da Computação, não só por sua utilização em larga escala dentro de empresas, como também por seu aproveitamento natural em pesquisas científicas, visto que permitem a recuperação de informações de maneira direcionada e organizada. É essencial que os resultados obtidos através de consultas possuam máxima coerência em relação ao mundo real. Bancos de dados convencionais foram criados para lidar com consultas exatas. Num tal sistema, um usuário especifica a consulta impondo condições restritas e o conjunto de resposta conterá apenas

elementos que respeitem, de forma exata, todas as condições formuladas. Assim, se um usuário consultar um banco de dados através da “lógica booleana”, por exemplo, valores que não estão compreendidos entre 3 e 6, a resposta será nula, mesmo que existam elementos “próximos do intervalo pesquisado” tais como 2.9 ou 2.99 ou 6.01. Por serem valores próximos, embora não pertencentes ao intervalo e que poderiam ser aproveitados, se retornasse com esses valores ao usuário a informação do grau de pertinência.

Este objetivo, contudo, nem sempre pode ser alcançado através dos conceitos rígidos que constituem a base da Álgebra Booleana, na qual, por sua vez, se baseia a lógica utilizada pelas linguagens de consulta a bancos de dados, esta Álgebra, baseia-se nos conceitos do “Verdadeiro/Falso” e só admite dois valores de verdade, sem meio-termo. Estes estados de verdade recebem sua respectiva “nomenclatura” de acordo com o contexto em que estão inseridos, mas inevitavelmente decorrem das variações de Liga/Desliga, Verdadeiro/Falso, 0/1, Sim/Não...

Para Lugo (2003), as buscas a banco de dados, quando realizadas utilizando “lógica booleana”, tendem a retornar resultados, muitas vezes, que não satisfazem as necessidades do usuário. Isso acontece porque este sistema se baseia em operações clássicas de união, interseção, complemento e negação da Álgebra Booleana para decidir a respeito da pertinência de um elemento a um determinado conjunto. Como só existem duas possibilidades, pertence ou não, percebe-se que esta aproximação está sujeita à perdas de informação que podem vir a comprometer a confiabilidade do resultado final, visto que no mundo real, na maioria das vezes, elementos podem pertencer “parcialmente” a um conjunto. Nota-se que, qualquer busca estará sujeita à perda de informação, isto ocorre porque estamos desprezando informações que poderiam ser relevantes pelo fato de as mesmas não obedecerem a um padrão de classificação por uma diferença nem sempre significativa.

Suponhamos, por exemplo, que dispomos de um banco de dados cujos registros possuem, entre outras informações, a altura e a idade de certo grupo de indivíduos. Digamos que se deseja listar as pessoas altas e jovens deste conjunto. Usando as técnicas convencionais, teríamos de, primeiramente, definir de forma precisa os termos: “alto” e “jovem”. O que é, afinal, uma pessoa alta? É possível usar nosso bom senso para estabelecer os limites dessa busca. Poderíamos, desta forma, pedir que o banco de dados nos respondesse quais pessoas possuem mais de 1,75m e cuja idade varia entre 15 e 30 anos. Se em um dos registros um indivíduo com 1,74m e 21 anos pertencesse a essa base de dados, ele seria “desprezado”, não sendo retornado pela busca. Tudo por causa de um centímetro. Mesmo que

mudássemos o limiar em um centímetro não resolveríamos o problema, uma vez que voltaríamos a ter a mesma situação para 1,73m e assim sucessivamente.

Precisamos, portanto, de uma forma de obter estes resultados desprezados pelas buscas através da “lógica booleana”. Mais do que isso, se pode ir além e lembrar que o conceito “alto” pode variar de país para país, de cultura para cultura, e essa tradução baseada no bom senso do usuário do banco de dados pode não ser muito confiável.

Poderíamos citar outro exemplo, na área financeira de um Banco Mercantil, onde um cliente com uma movimentação bancária expressiva poderia ter o seu cheque devolvido em função de um valor muito próximo acima do “estouro” de sua conta bancária. Citar também que, em nossa comunicação cotidiana usamos muitas palavras e sentenças com significado não preciso (ou vago). Isto acontece porque, tanto quem fala como quem ouve, não necessita de informações mais precisas e está acostumado a lidar com tais tipos de imprecisão. Por exemplo, alguém solicita uma sopa bem quente, está muito pouco preocupado com a real temperatura da sopa, o que ele deseja é que a temperatura da sopa esteja bastante acima do que ele considera como morna. Certamente, ninguém será capaz de determinar o ponto preciso em que a sopa passa de morna para quente. Seria bom se pudéssemos evitar essas situações impertinentes.

Para lidar com este tipo de afirmações precisamos recorrer a definições nebulosas (fuzzy), também chamadas difusas e dentro desta perspectiva, a sopa é mais ou menos quente, dependendo de sua temperatura.

Lógica nebulosa é Segundo Zadeh (1965), a lógica que se propõe a modelar o modo de raciocínio inexato, e pode também ser definida, como a lógica que suporta os modos de raciocínio que são aproximados, ao invés de exatos, como estamos acostumados a trabalhar.

Assim, os conjuntos nebulosos são na verdade uma “ponte” que permite ainda o emprego de quantitativas, como por exemplo “muito quente” e “muito frio”.

Ainda segundo Zadeh (1965), raciocínio aproximado à “lógica nebulosa” consiste em aproximar a decisão computacional da decisão humana. Isto é feito de forma que a decisão de uma máquina não se resume apenas a um "sim" ou um "não", mas também tenha decisões "abstratas", do tipo "um pouco mais", "talvez sim", “perto de”, “quase bom” e outras tantas variáveis que representem as decisões humanas.

## 5.2 Histórico

Aristóteles, filósofo grego (384 – 322 a.C.), foi o fundador da ciência lógica apresentando a primeira sistematização da lógica da qual se tem notícia. Já antes de Aristóteles, havia uma certa preocupação com a questão da validade dos argumentos, principalmente por parte dos sofistas e de Platão. Mas esses pensadores, embora tenham se ocupado um pouco com tais questões, de fato nunca desenvolveram uma teoria lógica, nunca se preocuparam em fazer um estudo sistemático dos tipos de argumentos válidos. Através da leitura dos diálogos de Platão, Aristóteles descobriu que existe uma lei que rege o pensamento para que se atinja o conhecimento de algo, a verdade, sem cair em contradição. Para Aristóteles, a lógica seria um instrumento para a ciência e a filosofia. A lógica aristotélica estava a serviço de uma explicação da realidade e baseava-se na distinção entre verdadeiro e falso.

Investigando os tipos de raciocínio, Aristóteles construiu uma teoria cujo núcleo é a caracterização e análise dos silogismos. Um exemplo típico de silogismo é:

Todo homem é mortal  
 Sócrates é homem  
 Logo, Sócrates é mortal.

Outro exemplo

Todos os homens bem sucedidos são homens profundamente interessados em seu trabalho.

Nenhum homem que esteja profundamente interessado em seu trabalho se distrai facilmente, quando está trabalhando.

Portanto, nenhum homem que se distrai facilmente quando está trabalhando é um homem bem sucedido.

Apesar das limitações para representar todos os tipos de inferências, o domínio da lógica silogística prevaleceu até o século XIX, quando George Boole (1815-1864), concebeu um sistema de símbolos e regras aplicável desde números até enunciados. Com esse sistema é possível codificar proposições em linguagem simbólica e manipulá-las quase da mesma maneira como se faz com os números.

Com a publicação do livro em 1849, “Investigação sobre as leis do Pensamento”, a Lógica afasta-se da Filosofia e aproxima-se da Matemática, e essa publicação deu início à simbolização, da lógica, o que se constituiu fazer uma linguagem simbólica, artificial, o que Aristóteles havia começado em grego, uma linguagem natural. Boole, na verdade apresentou

um cálculo lógico, hoje conhecido também como álgebra booleana contendo um número infinito de formas válidas de argumentos. Daí em diante a lógica alcançou um desenvolvimento extraordinário com implicações em praticamente todos os campos do saber humano.

No entanto, mesmo com tal desenvolvimento, os princípios básicos da lógica de tradição aristotélica não foram questionadas, dentre elas o chamado “Princípio da Não Contradição”, ou seja, uma declaração é falsa ou verdadeira, não podendo ser ao mesmo tempo parcialmente verdadeira ou parcialmente falsa.

Na década de 60, um afegão professor de Engenharia Elétrica e Ciência da Computação da Universidade de Berkeley, chamado de Lotfi Zadeh observou que muitas regras que as pessoas usavam para fazer inferências não eram conscientes, ou seja, não podia ser explicadas pelas pessoas que as usavam. Por exemplo, podemos olhar para uma pessoa e dizer que “ela parece ter em média 18 anos”, entretanto não conseguiremos explicar como chegamos a essa conclusão. Zadeh publicou em 1965 um artigo resumido sobre os conceitos dos conjuntos Fuzzy, revolucionando esse assunto e a primeira aplicação foi controlar uma máquina a vapor e depois dessa vieram, os controladores Fuzzy de plantas nucleares, refinarias, processos biológicos e químicos, tratamento de água e sistema de operação automática de trens e metrô.

Em função dos desenvolvimentos nas áreas citadas e pelas enormes possibilidades práticas de aplicações que se apresentaram, os estudos sobre Sistema Fuzzy avançaram rapidamente, culminando com a criação em 1984, da Sociedade Internacional de Sistemas Fuzzy, constituídas principalmente, por pesquisadores dos países mais avançados tecnologicamente.

O desenvolvimento de Sistemas Fuzzy, técnicas de Inteligência Artificial e Redes Neurais são consideradas algumas das mais promissoras técnicas para o século XXI.

### 5.3 Uma reflexão da teoria da “lógica booleana” e da teoria da “lógica nebulosa”.

A lógica de Aristóteles estabelece um conjunto de regras rígidas para que as conclusões sejam aceitas como logicamente válidas; seguindo uma linha de raciocínio baseada em premissas e conclusões. Por exemplo:

“Todo ser vivo é mortal” (premissa 1)

“Maria é um ser vivo” (premissa 2)

Como conclusão temos:

“Maria é mortal”.

Esta lógica formal é considerada binária, pois uma declaração é verdadeira ou falsa, não podendo ser ao mesmo tempo parcialmente verdadeira e parcialmente falsa. Obrigatoriamente, uma declaração deve fazer parte de apenas um conjunto, o das declarações verdadeiras ou o das declarações falsas. Esta suposição é o “Princípio da Não-contradição”, que coloca A e não A cobrindo todas as possibilidades, sendo considerada a base do pensamento lógico tradicional.

A “lógica nebulosa” viola estas suposições. O conceito de dualidade estabelece que algo pode e deve coexistir com seu oposto; o que fez a “lógica nebulosa” parecer natural e até mesmo inevitável. A “lógica booleana” trata com valores “verdades” das afirmações, classificando-as como verdadeiras ou falsas. Mas muitas experiências humanas não podem ser classificadas simplesmente como verdadeiras ou falsas, sim ou não, branco ou preto. Na verdade, entre a certeza de ser e a certeza de não ser existem infinitos graus de incerteza. Ao contrário da “lógica booleana”, a “lógica nebulosa” não impõe limites bruscos, mas proporciona graus de pertinência de elementos a uma determinada categoria. Para exemplificar, utilizaremos a palavra meia-idade. É comum classificarmos uma pessoa de meia idade ou não, apesar de não se saber exatamente quando começa e termina esse período. Vamos assumir que o período vai de 35 a 55 anos. Pela lógica tradicional, uma pessoa de 34 anos só pertenceria ao grupo de meia-idade depois do dia de seu aniversário de 35 anos. Da mesma maneira, uma pessoa de 56 anos não pertenceria mais a este grupo.

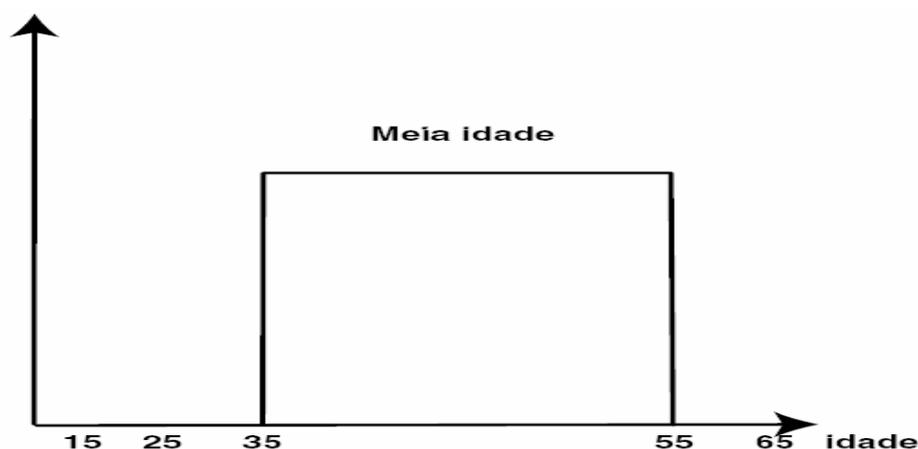


Figura 5.01: Gráfico Booleano

Mas não é desejada tamanha precisão em relação a este conceito, pois o limite de meia-idade não pode ser definido precisamente. Note que os limites do conjunto acima são

muito rígidos. O ideal é ter limites graduais, onde são associados graus de certeza que uma pessoa seria ou não de meia idade.

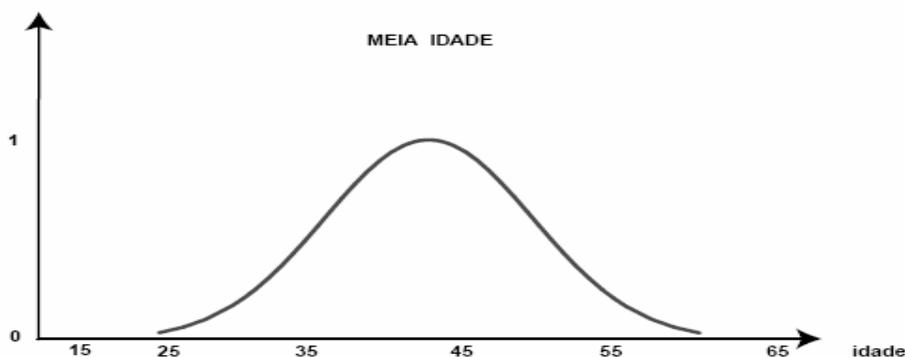


Figura 5.02: Gráfico nebuloso

#### 5.4 Lógica e Álgebra booleana

A Álgebra de Boole passou a ter utilização prática a partir de 1937, quando foi feita a primeira aplicação à análise de circuitos e reles por A Nakashima.

E em 1938 Claude E. Shannon mostrou, em sua tese de mestrado no Departamento de Engenharia Elétrica do MIT a aplicação da álgebra booleana na análise de circuitos de reles. Daghlion (1995).

A álgebra de Boole considera o mundo em que vivemos situações com dois estados apenas, que mutuamente se excluem, tais como: Verdadeiro, Falso; Ligado, Desligado; Sim, Não etc. não tratando de situações tais como: morno, alto, diferentes tonalidades de vermelho etc, que não se apresentam como estritamente dicotômicas, ou seja, com dois estados excludentes bem definidos.

Consideremos então alguns conceitos elementares de especial interesse que são: a noção de conjunto, elementos de um conjunto e a relação de pertinência.

Assim, dado um conjunto  $A = \{1,2,3\}$ , dizemos que 1, 2 e 3 são elementos do conjunto A, em conseqüência, pertencem ao conjunto A, caso tenhamos um elemento 4 que não aparece no conjunto A, diríamos que o 4 não pertence ao conjunto A.

Consideremos também as sentenças ou proposições<sup>5</sup> por ex. :

A lua é um satélite da Terra.

O sol gira em torno da Terra.

Representamos por V ou F o valor verdade de uma proposição, considerando V quando a proposição for verdadeira e F caso contrário. Nos exemplos apresentados a proposição (a) é verdadeira (V) e a proposição (b) é falsa (F).

As proposições se classificam em:

**Simples:** aquela que não contém nenhuma outra proposição como parte integrante de si mesma. Ex. “Carlos é solteiro”.

**Composta:** aquela formada pela combinação de duas ou mais proposições simples. Ex. “Carlos é solteiro e Pedro é estudante”.

Para as proposições compostas surgem então os conectivos, que são palavras usadas para formar novas proposições a partir de outras, tais como: “se .. então”; “e”, “ou”, “se e somente se” e “não”. Trataremos apenas dos conectivos “e” e o “ou” neste trabalho.

Chama-se conjunção de duas proposições “p” e “q” a proposição, representada por “p e q”, cujo valor-verdade é a verdade quando p e q são ambas verdadeiras e a falsidade nos outros três casos. A conjunção pode ser representada pela seguinte tabela

p	q	p e q
V	V	V
V	F	F
F	V	F
F	F	F

---

<sup>5</sup> Proposições, todo o conjunto de palavras que exprimem um pensamento de sentido completo. Alencar (1968).

Exemplos:

a) p: Roma é capital da Itália (V); q:  $2 + 3 = 5$  (V);

Roma é capital da Itália e  $2 + 3 = 5$  (V);

b) p: Roma é capital da Itália (V); q:  $2 + 3 = 4$  (F);

Roma é capital da Itália e  $2 + 3 = 4$  (F);

c) p: Roma é capital da França (F); q:  $2 + 3 = 5$  (V);

Roma é capital da França e  $2 + 3 = 5$  (F);

d) p: Roma é capital da França (F); q:  $2 + 3 = 4$  (F);

Roma é capital da França e  $2 + 3 = 4$  (F);

Chama-se disjunção de duas proposições “p” e “q” a proposição, representada por “p ou q”, cujo valor-verdade é a verdade quando ao menos uma das proposições “p” e “q” é verdadeira e falsa quando ambas são falsas. A disjunção pode ser representada pela seguinte tabela.

p	q	p ou q
V	V	V
V	F	V
F	V	V
F	F	F

Exemplos:

a) p: Roma é capital da Itália (V); q:  $2 + 3 = 5$  (V);

Roma é capital da Itália ou  $2 + 3 = 5$  (V);

b) p: Roma é capital da Itália (V); q:  $2 + 3 = 4$  (F);

Roma é capital da Itália ou  $2 + 3 = 4$  (V);

c) p: Roma é capital da França (F); q:  $2 + 3 = 5$  (V);

Roma é capital da França ou  $2 + 3 = 5$  (V);

d) p: Roma é capital da França (F); q:  $2 + 3 = 4$  (F);

Roma é capital da França ou  $2 + 3 = 4$  (F);

#### 5.4.1 Modelo booleano

No modelo booleano um documento é representado por um conjunto de termos de indexação que podem ser definidos de forma manual por profissionais especializados ou automaticamente, através da utilização de algum tipo de algoritmo computacional. As buscas são formuladas através de uma expressão booleana composta por termos ligados através dos operadores lógicos AND, OR e NOT (E, OU e NÃO), e apresentam como resultado os documentos cuja representação satisfaz às restrições lógicas da expressão de busca.

O modelo booleano, apesar de bem formalizado, possui limitações que diminui sua atratividade.

Algumas dessas limitações são:

a) Sem um treinamento apropriado, o usuário leigo será capaz de formular somente buscas simples. Para buscas que exijam expressões mais complexas é necessário um conhecimento sólido da “lógica booleana”.

b) Existe pouco controle sobre a quantidade de documentos resultante de uma busca. O usuário é incapaz de predizer quantos registros satisfarão à restrição lógica de uma determinada expressão booleana, sendo necessárias sucessivas reformulações antes que seja recuperado um volume aceitável de documentos;

c) O resultado de uma busca booleana se caracteriza por uma simples partição do conjunto de documentos em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual utilidade para o usuário;

d) Não existe uma forma de atribuir importância relativa aos diferentes termos da expressão booleana. Assume-se implicitamente que todos os termos têm o mesmo peso.

Apesar de suas limitações, o modelo booleano está presente em quase todos os sistemas de recuperação de informação, seja como a principal maneira de formular as expressões de busca, seja como um recurso alternativo. Uma razão para isso é que para usuários experientes este modelo oferece certo controle sobre o sistema. Se o conjunto de documentos resultante a sua quantidade é significativa ou não, e é fácil saber quais os operadores necessários para diminuir ou aumentar a quantidade de documentos até atingir um resultado satisfatório.

Uma das maiores desvantagens do modelo booleano é a sua inabilidade em ordenar os documentos resultantes de uma busca; no entanto, o ordenamento dos documentos é de extrema importância face ao volume de documentos que geralmente pode ser recuperado. Apesar disso, muitos desses sistemas se desenvolveram utilizando o modelo booleano como ponto de partida para a implementação de novos recursos de recuperação. Neste sentido o modelo booleano pode ser considerado o modelo mais utilizado nos sistemas de recuperação de informação, e nos sistemas de banco de dados, onde o seu poder se expressa através da linguagem SQL, (Structured Query Language).

Os operadores booleanos podem ser combinados com os operadores de proximidade a fim de formar expressões de busca mais restritivas ou mais genéricas e para isso faremos a utilização da “lógica nebulosa” e de operadores de proximidade com o SQL nebuloso.

### 5.5 Conceitos da “lógica nebulosa”

Os conceitos da “lógica nebulosa” utilizados compreendem: Zadeh (1965)

**Conjuntos nebulosos:** um conjunto  $A$  pertencente a um universo  $X$  é denominado conjunto nebuloso, se  $x$  pertence a  $A$ ,  $x$  é membro do conjunto  $A$  com um grau de pertinência  $u(x)$ ;

**Grau de pertinência:** o grau de pertinência de um elemento  $x$ , denotado por  $u(x)$ , em um conjunto nebuloso  $A$ , compreende um valor numérico que estabelece o “quanto” esse elemento pertence ou não a esse conjunto;

**Função de pertinência:** é a função que estabelece o grau de pertinência de um elemento  $x$  em um conjunto nebuloso  $A$ , e possui sua imagem no intervalo  $[0,1]$ ;

**Relação de similaridade:** compreende o cálculo da similaridade entre elementos de um mesmo domínio, e deve obedecer às seguintes propriedades: considere  $s$  uma relação de similaridade e  $D$  um domínio de elementos; para todo  $x, y$  e  $z$  pertencente a  $D$  tem-se:

(1) reflexiva:  $s(x,x) = 1$ ;

(2) simétrica:  $s(x,y) = s(y,x)$ ;

(3) transitiva:  $s(x,z) \geq \max$  (para todo  $y$  pertencente a  $D$ )  $\{ \min[s(x,y), s(y,z)] \}$

**Tabela de similaridade:** é uma tabela cujos valores apresentam a relação de similaridade entre todos os elementos de um mesmo domínio.

O objetivo da lógica fuzzy é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa.

A teoria dos conjuntos fuzzy é baseada no fato de que os conjuntos existentes no mundo real não possuem limites precisos. Um conjunto fuzzy é um agrupamento indefinido de elementos no qual a transição de cada elemento de não-membro para membro do conjunto é gradual. Esse grau de imprecisão de um elemento pode ser visto como uma “medida de possibilidade”, ou seja, a “possibilidade” de que um elemento seja membro do conjunto.

Em muitos problemas em física e matemática, nós não temos dificuldade em classificar elementos como pertencentes ou não a um dado conjunto clássico. Dessa forma, dado um conjunto  $A$  e um elemento  $x$  do conjunto universo  $U$  conseguimos muitas vezes dizer se  $x$  pertence a  $A$  ou  $x$  não pertence a  $A$ . Podemos, por exemplo, afirmar sem receio que o número 7 pertence a  $N$  e que  $-7$  não pertence a  $N$ . Este é um caso sobre o qual não se tem dúvida, sendo a “lógica booleana” perfeitamente aplicável. No entanto, podemos discordar quanto ao fato de 6,5 pertencer ou não ao conjunto dos números aproximadamente iguais a 7. Neste caso a resposta não é única e objetiva, pertencer ou não vai depender do tipo de problema que estamos analisando. Para ilustramos esta situação, suponhamos que um aluno obteve médias 6,5. O professor nessa situação poderia analisar o caso levando em consideração a capacidade do aluno, a sua dedicação durante o curso. E pode decidir pela aprovação do aluno mesmo não tendo atingido a média 7. Note que, neste caso, o número 6,5 está sendo considerado como pertencente ao conjunto dos aproximadamente iguais a 7. Existem inúmeras situações em que a relação de pertinência não está bem definida, e nestes casos, não sabemos dizer se o elemento pertence ou não a um determinado conjunto.

A intenção de Zadeh foi justamente trabalhar de forma a flexibilizar a pertinência de elementos aos conjuntos, criando uma idéia de grau de pertinência. Podendo, assim, um determinado elemento pertencer parcialmente a um conjunto.

Se desejarmos, por exemplo, construir o conjunto dos números aproximadamente iguais a 5, citado acima, como deveria procedermos? Será que os números 2 e 10 pertenceriam a este conjunto? Claramente esta resposta dependeria do contexto. O que Zadeh nos propõe é considerarmos uma função que nos forneça o quanto um determinado número pertence ao conjunto considerado. Sendo assim, chamando de  $K$  o conjunto dos números aproximadamente iguais a 5, no universo dos números naturais  $N$ , podemos propor por exemplo uma função de pertinência onde:

10 pertence a  $K$  com grau de pertinência 0,00 (corresponde a não pertinência clássica).

2 e 8 pertencem a  $K$  com grau de pertinência 0,25

3 e 7 pertencem a  $K$  com grau de pertinência 0,50

4 e 6 pertencem a K com grau de pertinência 0,75

5 pertence a K com grau de pertinência 1,00 (corresponde a pertinência total).

Esta extensão da função característica da lógica clássica para o intervalo  $[0,1]$ , originou os conjuntos fuzzy e possibilitou, entre outras coisas, a utilização de variáveis lingüísticas, permitindo a exploração do conhecimento humano no desenvolvimento de vários sistemas.

A figura abaixo, apresenta a função de pertinência de conjunto fuzzy que corresponde aos números aproximadamente iguais a 5. Note que nesse caso a função de pertinência é discreta, mas também pode ser contínua. A “lógica nebulosa”, também pode ser definida, como a lógica que suporta os modos de raciocínio que são aproximados, ao invés de exatos, como estamos naturalmente acostumados a trabalhar. Ela está baseada na teoria dos conjuntos nebulosos e difere dos sistemas lógicos tradicionais em suas características e detalhes. Nesta lógica, o raciocínio exato corresponde a um caso limite do raciocínio aproximado, sendo interpretado como um processo de composição nebulosa.

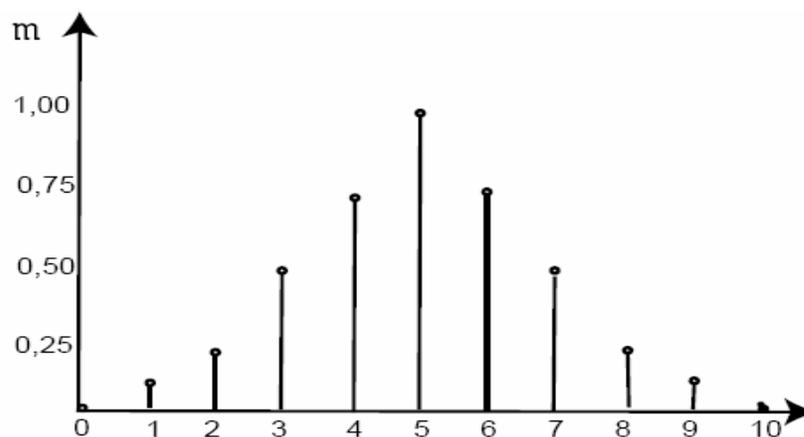


Figura 5.03: Gráfico nebuloso representando graus de pertinência.

A Teoria dos Conjuntos clássica fornece todas as regras para pertinências de conjuntos e a relação entre os mesmos e que são utilizadas em toda a Matemática. Essas relações são definidas através de uma função de avaliação que nos diz se um dado elemento pertence ou não a um determinado conjunto. Ou seja, ela recebe um elemento  $x$  e um conjunto  $X$  e retorna 0 (falso) se  $x$  pertence a  $X$  ou 1 (verdadeiro) caso  $x$  não pertence a  $X$ . Este tipo de definição nos expõe as deficiências claras, ainda mais quando tratamos de conjuntos em que a incerteza

é um fator importante. Mais do que isso, o limiar entre pertencer ou não a uma dada definição é muito rígido, o que pode nos trazer problemas de perda de informação.

Ao tratarmos os conjuntos com regras de relação e pertinência nebulosas, passamos a definir o retorno da função citada anteriormente como um número real no intervalo  $[0,1]$ . Ou seja, ao invés de termos 0 **ou** 1, temos qualquer valor **entre** 0 e 1. Esta função é denominada de “**função de inclusão**”. Dessa forma temos uma avaliação mais sutil e flexível da pertinência de um elemento a certo conjunto. Podemos, agora, fazer uma definição mais formal acerca dos conjuntos nebulosos e suas relações com os respectivos elementos. Assim sendo, seja uma coleção de objetos indicados genericamente por  $X$ . Então um **conjunto nebuloso**  $A$  em  $X$  é um conjunto de pares ordenados:  $A = \{(x, mA(x)) \mid x \text{ pertencente } X\}$  onde  $mA$  é a chamada função de inclusão e  $mA(x)$  é o grau de inclusão de  $x$  em  $X$ .

Uma **variável nebulosa** é definida através de quatro elementos: um nome simbólico, um conjunto de termos ou **rótulos**, um universo de discurso e uma regra semântica. O nome simbólico nada mais é do que uma identificação que usamos para distinguir a variável, como, por exemplo “temperatura”. O universo de discurso é o mesmo que domínio, no caso poderia ser entre  $-20^{\circ} \text{ C}$  e  $45^{\circ} \text{ C}$ . Este domínio pode ser subdividido em diversos subconjuntos, para os quais damos o nome de rótulos, tais como muito frio (entre  $-20^{\circ} \text{ C}$  e  $5^{\circ} \text{ C}$ ), frio (entre  $-5^{\circ} \text{ C}$  e  $17^{\circ} \text{ C}$ ), ameno (entre  $15^{\circ} \text{ C}$  e  $27^{\circ} \text{ C}$ ), quente (entre  $25^{\circ} \text{ C}$  e  $35^{\circ} \text{ C}$ ) e muito quente ( $31^{\circ} \text{ C}$  e  $45^{\circ} \text{ C}$ ). Vale notar que os rótulos não são, necessariamente, mutuamente exclusivos. Esta, aliás, é uma das principais características da “lógica nebulosa” e que representa uma grande vantagem em relação à tradicional. As regras semânticas são séries de funções de inclusão que associam elementos com os rótulos. Além disso, encontramos novas interpretações para as operações elementares da Teoria dos Conjuntos, tais como união, interseção, negação e complemento. Suponhamos três conjuntos  $A$ ,  $B$  e  $C$ . Usando a lógica tradicional, se fizermos  $C = A \text{ união } B$ ,  $C$  será o conjunto cujos elementos são todos aqueles que pertencem a  $A$  **ou** a  $B$ . Por outro lado,  $C = A \text{ interseção } B$ , faria com que  $C$  fosse formado por todos os elementos que pertencem a  $A$  **e** a  $B$  **ao mesmo tempo**. Finalmente,  $C = \text{não } A$  atribui a  $C$  todos os elementos do universo considerado que não pertencem a  $A$ .

Tabela de Características e Vantagens da “lógica nebulosa”.

Características	Vantagens	Desvantagens
A “lógica nebulosa” está baseada em palavras e não em números, ou seja, os valores verdades são expressos lingüisticamente. Por exemplo: quente, muito frio, verdade, longe, perto, rápido, vagaroso, médio;	O uso de variáveis lingüísticas nos deixa mais perto do pensamento humano;	Necessitam de mais simulação e testes;
Possui vários modificadores de predicado, tais como: muito, mais ou menos, pouco, bastante, médio;	Requer poucas regras, valores e decisões;	Não aprendem facilmente;
Possui também um amplo conjunto de quantificadores, como: poucos, vários, em torno de, usualmente;	Simplifica a solução de problemas e a aquisição da base do conhecimento;	Dificuldades de estabelecer regras corretamente;
Faz uso das probabilidades lingüísticas (como, provável e improvável) que são interpretados como números nebulosos e manipulados pela sua aritmética;	Variáveis observáveis podem ser valoradas;	Não há uma definição matemática precisa.
Manuseia todos os valores entre 0 e 1, tomando estes, como um limite apenas.	Mais fácil de entender, manter e testar; São robustos. Operam com falta de regras ou com regras defeituosas; Acumulam evidências contra e a favor.	

### 5.6 Buscas usando conceitos e funções nebulosas.

O processo de busca, é responsável pela tradução de uma consulta construída usando-se termos vagos, onde será utilizada a linguagem SQL padrão, para a qual serão transferidos os conceitos da “lógica nebulosa” e que é denominado de SQL Nebuloso. Essa tradução, a seguir, é enviada para o banco, o qual realiza a busca, e seus resultados retornam ao programa que os interpreta e fornece uma resposta ao usuário. Para que o processo de busca, possa interpretar essas consultas, contudo, é preciso que se forneça um arquivo de definições, o qual contém dados quantitativos sobre os conjuntos nebulosos a fim de que seja possível ao interpretador em relação ao banco de dados, montar um esquema de tradução adequado às expectativas do usuário. Neste arquivo estarão as definições das variáveis nebulosas, de seus respectivos rótulos e funções de cujas informações serão extraídos dados necessários para a realização bem sucedida da tradução.

Ao se utilizar o SQL padrão as buscas em relação aos conjuntos:

$$A = \{a, b, c\}$$

$$B = \{b, c, d\}$$

Teríamos como resultados para as funções União e Intersecção

$$C = A \text{ união } B = \{a, b, c, d\}$$

$$D = A \text{ intersecção } B = \{b, c\}$$

Mas com os conceitos da “lógica nebulosa”, ao se utilizar o SQL Nebuloso, em relação aos conjuntos:

$$A = \{(a,0.5); (b,0.8); (c,0.75); (d, 0.2); (e,0.0)\}$$

$$B = \{(a,0.0); (b,0.85); (c,0.6); (d, 1.0); (e,0.15)\}$$

Teríamos como resultados para as funções União e Intersecção

$$C = A \text{ união } B = \{(a,0.5); (b,0.85); (c,0.75); (d, 1.0); (e,0.15)\}$$

$$D = A \text{ intersecção } B = \{(a,0.0); (b,0.8); (c,0.6); (d, 0.2); (e,0.0)\}$$

A consulta nebulosa constitui, na realidade, uma versão modificada da linguagem SQL padrão, na qual usamos termos vagos no lugar de expressões booleanas. Por exemplo, ao invés de usarmos uma expressão do tipo, “altura maior que 1,65 e altura menor que 1,75”, passamos a usar uma expressão do tipo “**onde a altura é a média**”, lógico, respeitando as unidades para a altura e a sintaxe para acesso ao banco de dados.

No nosso exemplo anterior, “altura” seria uma variável nebulosa e “média” é, uma de suas possíveis funções de cálculo de aproximações para a busca.

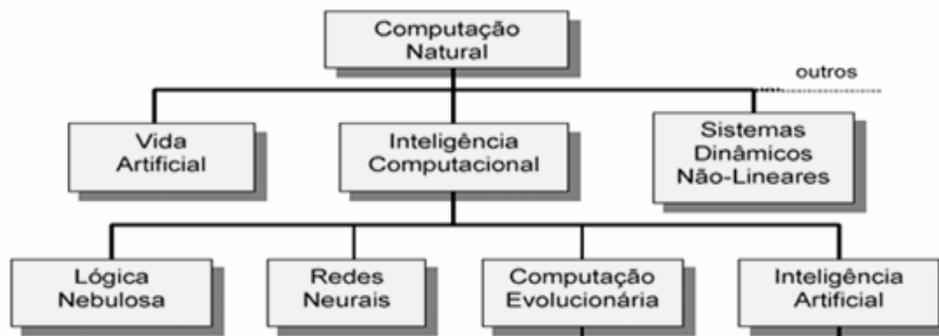
Para esse procedimento de busca deveremos, analisar a sintaxe da consulta, lançando mão, para isso, de uma gramática previamente construída, a qual prevê a construção formal necessária para que a tradução possa ser realizada corretamente. Analisamos, também, a semântica utilizada, a fim de evitar o uso sem sentido de termos, como seria o caso de consultar altura é jovem. Após ter passado com sucesso pela análise, a consulta é, finalmente, traduzida. Este passo é comumente chamado de “defuzzyficação”, uma vez que é responsável pela transformação de termos nebulosos (fuzzy) em termos usados em consultas padrão. Isto é necessário já que a linguagem SQL, usada como interface para o banco de dados, é construída baseada justamente na teoria dos conjuntos.

Como já mencionado anteriormente, a construção do arquivo de definições é que determina a forma como será feita a tradução. A consulta traduzida é, a seguir, enviada ao banco de dados que envia uma resposta com os resultados da mesma. Com esses resultados, o programa entra na fase de “fuzzyficação”, que, na verdade, “traduz de volta” os dados obtidos

para a forma nebulosa. Usando as funções fornecidas no arquivo de definições, o sistema calcula o grau de inclusão de cada registro retornado, ordenando-os e expondo a resposta final para o usuário. Como podemos observar, o processo de busca desenvolvido é uma “camada” que se coloca entre o usuário e a base de dados. Para a pessoa que o utiliza, tudo é feito através de termos vagos e nebulosos, ao passo que para o banco, tudo ocorre dentro dos padrões do SQL tradicional.

### 5.7 “Lógica nebulosa” no contexto da computação natural.

Gráfico mostrando a “lógica nebulosa” no contexto da Computação Natural



Breve definição de cada um desses campos classificados e arranjados acima: Quadro parcial retirado da dissertação de mestrado de (Fernandes, (2005)).

**Computação Natural:** Universo científico que incorpora as áreas de pesquisa: Inteligência Computacional, Vida Artificial, Sistemas Dinâmicos Não-Lineares e outros sistemas complexos.

**Vida Artificial:** É um campo de pesquisa que busca aumentar o papel da síntese no estudo do fenômeno biológico. O potencial de VA é atribuído à possibilidade de, através de seu estudo, se desvendar segredos da vida e levantar a discussão sobre uma série de temas científicos, técnicos e psicológicos.

Tópicos neste campo incluem: Auto-Organização e funcionalidade emergente; Definições e Origens da Vida; Auto-Reprodução; Vírus em Computadores; Evolução e Populações Genéticas; Co-evolução e Dinâmica Ecológica; Organização e comportamento de organismos coloniais e sociais; Agentes Autônomos (Robôs Autônomos e Agentes de Software); Inteligência Coletiva; Comportamento Animal; Ecossistemas Locais e Globais; etc.

**Sistemas Dinâmicos Não-Lineares:** São processos em movimento ou processos que mudam no tempo ou no espaço (caos e fractal). SDNLs variam de muito simples a extremamente complexos e são encontrados com facilidade. A motivação para a pesquisa nesta área reside na habilidade em se prever a direção futura ou estado de tais sistemas através do tempo ou espaço. Tipicamente, e sobre certas circunstâncias, mesmo os mais simples sistemas dinâmicos exibem um comportamento imprevisível e quase aleatório. Tal comportamento é denominado caótico.

Exemplos: Movimento dos corpos celestes; Os altos e baixos das bolsas de valores; As condições meteorológicas; etc.

**Inteligência Computacional:** Área da ciência que estuda a teoria e a aplicação de técnicas inspiradas na natureza, como: Redes Neurais; Lógica Nebulosa; Computação Evolucionária.

a) **Redes Neurais:** Redes Neurais são modelos computacionais não lineares, inspirados na estrutura e operação do cérebro humano, que procuram reproduzir características humanas, tais como: Aprendizado; Associação; Generalização e Abstração.

Devido à sua estrutura, as Redes Neurais são bastante efetivas no aprendizado de padrões a partir de dados: Não-lineares; Incompletos; Com ruído e Compostos de exemplos contraditórios.

b) **Lógica Nebulosa :** Lógica Nebulosa (Fuzzy Logic) tem por objetivo modelar o modo aproximado de raciocínio humano, visando desenvolver sistemas computacionais capazes de tomar decisões racionais em um ambiente de incerteza e imprecisão. A “lógica nebulosa” fornece um mecanismo para manipular informações imprecisas, tais como os conceitos de muito, pouco, pequeno, alto, bom, quente, frio, etc, fornecendo uma resposta aproximada para uma questão baseada em um conhecimento que é inexato, incompleto ou não totalmente confiável.

c) **Computação Evolucionária:** É uma área de pesquisa interdisciplinar que compreende diversos paradigmas inspirados no princípio Darwiniano da evolução das espécies, tais como:

d) **Inteligência artificial:** Inteligência Artificial (IA) é uma área da computação que primeiramente trabalha com o processamento simbólico de conhecimento, criando programas que fazem os computadores parecerem inteligentes. As soluções dos problemas são heurísticas e respostas satisfatórias são aceitas. IA é constituída de técnicas próprias para a solução de problemas.

## 5.8 Conclusões do capítulo

Mostramos neste capítulo que a “lógica booleana” está presente em quase todos os sistemas de recuperação de informação, mas tem suas limitações, pois tendem a retornar resultados, muitas vezes, que não satisfazem as necessidades do usuário. Notamos que isso acontece porque o sistema booleano se baseia em operações clássicas de união, interseção, complemento e negação da Álgebra Booleana para decidir a respeito da pertinência de um elemento a um determinado conjunto. Como só existem duas possibilidades, pertence ou não, concluímos que esta aproximação está sujeita a perda de informação, visto que no mundo real ao analisarmos pelo prisma do raciocínio humano, verificamos que valores podem pertencer parcialmente a um conjunto. Estudamos então uma forma de obter estes resultados desprezados pelas buscas da “lógica booleana”. Através dos conceitos e aplicações da “lógica nebulosa” conseguimos associar funções e conceitos do “SQL nebuloso” para trabalhar de forma a considerar a pertinência de elementos aos conjuntos, criando idéias de “grau de pertinência” para através dos mesmos propiciar condições aos tomadores de decisão aceitar as informações como verdadeira ou não.

## CAPÍTULO 6 – Aplicação da “lógica nebulosa” à recuperação de informação em “data warehouse”.

Neste capítulo, será feita uma comparação dos resultados de recuperação de dados através da integração de “data warehouse” com a “lógica nebulosa” e com a “lógica booleana”. Também serão mostrados, os resultados das pesquisas à tomada de decisão.

Tomar decisões é uma a atividade fundamental dos seres humanos, por isso, muitos estudos têm sido desenvolvidos acerca desse tema. O objeto de análise de um processo de decisão é justamente o estudo de quais estratégias serão utilizadas e como escolhê-las da melhor forma ou mais eficazmente. Estudos vêm sendo amplamente aplicados nas áreas de gerenciamento (controle, investimentos, desenvolvimento de novos produtos, alocação de recursos, etc.) onde o processo de decisão tem papel fundamental. Incluem qualquer situação onde uma escolha ou seleção de alternativas se faz necessária, abrangendo desde as ciências tidas como exatas até as humanas.

A aplicação da teoria de conjuntos fuzzy em processos de decisão consiste basicamente na “fuzificação” da teoria de decisão booleana. Enquanto os processos de decisão sobre condições de risco têm sido modelados com teorias de decisão probabilística e teoria de jogos, a teoria de decisão fuzzy procura lidar com as formulações vagas e imprecisas, inerentemente humanas, quanto suas preferências, limitações e objetivos.

Um processo de decisão é dito sob condições de certeza quando o resultado de cada ação pode ser determinado com precisão. Ele é dito sob condições de risco quando o conhecimento disponível acerca do resultado consiste apenas da distribuição de probabilidade condicional, relativo a cada ação. Por outro lado, quando os resultados devidos a cada ação são caracterizados apenas aproximadamente, então, o processo de decisão é dito sobre condições de imprecisão. Este é o caso dos processos de decisão fuzzy. Os conjuntos fuzzy podem ser introduzidos na teoria de tomada de decisão de diversas formas, Zadeh (1975 apud Haack, 2002, p. 65), sugeriu um modelo fuzzy de tomada de decisão nos quais os objetivos e restrições relevantes são expressas e a decisão é determinada a partir de um tipo de agregação apropriada desses conjuntos.

Os sistemas difusos estimam funções com descrição parcial do comportamento deles, onde especialistas podem prover o conhecimento heurístico, ou esse conhecimento pode ser inferido a partir de dados de entrada-saída do sistema. Dessa forma, podemos dizer que os

sistemas difusos são baseados em regras que utilizam variáveis lingüísticas<sup>6</sup>, difusas para executar um processo de tomada de decisão.

Um Sistema de Inferência pode ser composto de: Base de regras (contêm um conjunto de regras/proposições fuzzy); base de dados (define as funções de pertinência do conjunto difuso nas regras fuzzy); unidade de decisão lógica (realiza operações de inferência, para obter, a partir da avaliação dos níveis de compatibilidade das entradas com as condições impostas pela base de regras, uma ação a ser realizada pelo sistema); interface de “fuzificação” (utiliza as funções de pertinência pré-estabelecidas, mapeia cada variável de entrada do sistema em graus de pertinência de algum conjunto difuso que representa a variável em questão), ou seja, na “fuzificação”, é realizada a análise do ambiente, pois é de extrema importância conhecer: o problema a ser manipulado, como o especialista trabalha e como ele entende sua realidade, qual a linguagem usada pelo especialista para tratar os problemas, quais as decisões e quando elas são tomadas. A interface de “defuzificação” (transforma os resultados difusos da inferência em valores de saída ou seja, a defuzzificação consiste em obter a melhor representação para o conjunto de saída *fuzzy*). Calcula a saída com base na inferência obtida no módulo, unidade de decisão lógica, com as funções de pertinência, cujos resultados ao contrário dos resultados da “lógica booleana” que são dicotômicos, ou valor igual a “1” para se o elemento pertence ao conjunto ou “0” se não pertence), no caso da “lógica nebulosa” os resultados estariam num intervalo de valores contínuos entre “zero” e “um”, [0;1], ou seja o grau de pertinência seria um valor contínuo de valores de pertinência e não apenas pertence ou não pertence. Shimizu (2001, apêndice B) define que dado um conjunto A contido em X é denominado conjunto difuso se cada elemento x de X for caracterizado por uma função de pertinência  $u_a(x)$  cujo valor indica o grau de pertinência ou possibilidade de ocorrência desse elemento no conjunto A.

Por exemplo,

x possui pertinência  $u_a(x) = 1$ , se tivermos certeza que pertence ao conjunto A;

x possui pertinência  $u_a(x) = 0$ , se tivermos certeza de que não pertence ao conjunto A;

x possui pertinência  $u_a(x)$ , com valor entre zero e um, conforme pertença ao conjunto A com maior ou menor precisão.

---

<sup>6</sup> Variáveis lingüísticas difusas seriam: frio, normal, quente e etc.

Exemplo: idade das pessoas por meio de conjunto difuso

Seja:

$X = \{\text{idade entre 1 e 100 anos}\}$  e o conjunto difuso  $A = \{\text{jovem, meia-idade, idoso}\}$ .

Alguns valores da função de pertinência poderiam ser considerados:

- $u_a(\text{jovem com 10 anos}) = 1$  (pessoa com 10 anos é jovem com certeza).  
 $u_a(\text{jovem com 25 anos}) = 0.8$  (pessoa com 25 anos ainda é bastante jovem).  
 $u_a(\text{jovem com 35 anos}) = 0.3$  (não é jovem aos 35 anos).  
 $u_a(\text{idoso com 35 anos}) = 0$  (aos 35 anos ninguém é idoso).  
 $u_a(\text{meia-idade com 35 anos}) = 0.7$  (com 35 anos está entrando na faixa de meia-idade).

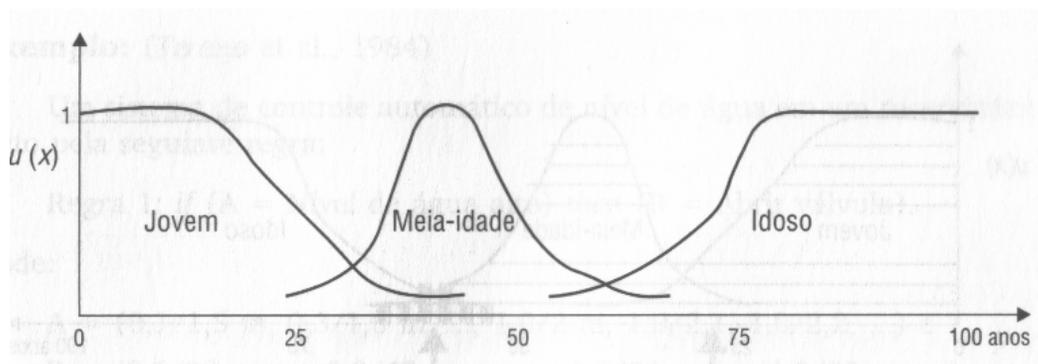


Figura 6.01: Idade das pessoas representadas por conjunto nebuloso.

Outras regras e funções podem ser criadas e fornecidas por especialistas, em forma de sentenças lingüísticas, e se constituem em um aspecto fundamental no desempenho de um sistema de inferência fuzzy. Extrair regras de especialistas na forma de sentenças do tipo “se ... então”, pode não ser uma tarefa fácil, por mais conhecedores que eles sejam do problema em questão. Alternativamente ao uso de regras “se..então”, existem métodos de extração de regras de dados numéricos.

No exemplo a seguir, criamos uma função denominada “grau de pertinência”, para ser verificada com comandos de busca em banco de dados, chamada SQL, cujos passos serão explicados a seguir.

Função “graudepertinência”:

**Passo p1**

```
Create or replace function graudepertinencia (valor_i in number, valor_f in number,
numero in number)
return varchar2 is
grau number (7,2);
dif number(7,2);
```

**Passo p2**

```
begin
if numero between valor_i and valor_f then
    grau:= 1;
```

**Passo p3**

```
    else
if numero < valor_i then
    dif := (valor_i - numero) * 100;
    grau:= 100 - dif;
else
    if numero > valor_f then
    dif := (numero - valor_f) * 100;
    grau := 100 - dif;
    end if;
end if;
end if;
```

**Passo p4**

```
-- retorna mensagem
if grau <= 0 then
    return 'o grau de pertinência é zero - número muito distante do intervalo';
else
    return 'o grau de pertinência é '||grau;
end if;
```

```
end;
```

```
/
```

**Final**

A função “graudepertinência” recebe três parâmetros: valor\_i, que corresponde ao valor inicial de um intervalo de números; valor\_f, que corresponde ao valor final do intervalo; e número, é o valor que será verificado quanto à sua aproximação ou pertinência no intervalo. **(passos p1 a p2).**

A seguir testa-se se o valor está compreendido no intervalo, se isto ocorrer o grau de pertinência será um (1 ou 100%), pois é certeza que o número está no intervalo verificado.

Exemplificando: consideremos, agora, o intervalo numérico de valores compreendidos entre 3 e 7. Na função teríamos valor inicial 3 e valor final 7, **(passos p2 a p3)**. Para valores fora deste intervalo, verifica-se a “distância” do número a pesquisar está “próximo” ou não dos extremos do intervalo. Determinamos um algoritmo muito simples. Para o valor mínimo 3, e o valor a ser pesquisado, por exemplo, 2.9, faríamos o seguinte cálculo: 3 menos o 2.9 resultaria o valor “0.1”, que será transformado em 10% e ao calcular-se  $100\% - 10\%$  obteríamos 90%, cujo valor representará o grau de pertinência, igual a 90%. Portanto, para o valor 2.99, com os mesmos cálculos, obteríamos um grau de pertinência de 99%. **(passos p3 a p4).**

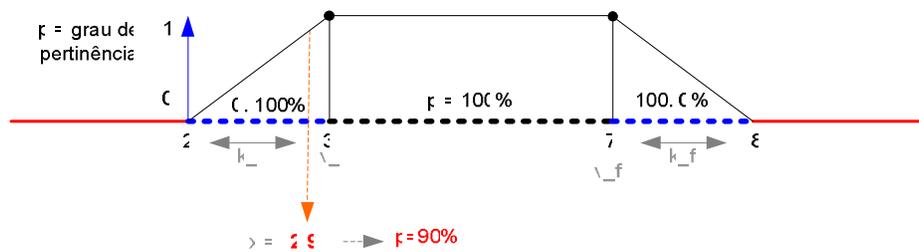
O mesmo procedimento poderia ser considerado para valores maiores que o extremo superior.

Se o valor pesquisado está a uma distância “grande” em relação aos extremos e fora do intervalo, o resultado seria negativo e o grau de pertinência seria considerado zero (0).

Por exemplo, para pesquisar o número 1, em relação ao mesmo intervalo e considerando o extremo inferior, ao fazer-mos o cálculo 3 menos o 1 teríamos o valor 2 que seria transformado em 200% e o resultado seria  $100\% - 200\%$  onde, obteríamos resultado negativo, então o grau de pertinência é considerado igual a zero (0). **(passo p4 até o final).**

Abaixo, resultados obtidos, executando no Oracle (Sistema gerenciador de banco de dados), a função graudepertinencia:

A figura abaixo representa a função relativa ao primeiro exemplo



a) Intervalo de [3 a 7] e valor procurado 2.9;

```
SQL> select graudepertinencia(3,7,2.9) from dual;
```

```
TESTE(3,7,2.9)
```

**O grau de pertinência é 90%.**

b) Intervalo de [3 a 7] e valor procurado 7.5;

```
SQL> select graudepertinencia(3,7,7.5) from dual;
```

```
TESTE(3,7,7.5)
```

**O grau de pertinência é 50%.**

c) Intervalo de [3 a 7] e valor procurado 8.5;

```
SQL> select graudepertinencia(3,7,8.5) from dual;
```

```
TESTE(3,7,8.5)
```

**O grau de pertinência é zero - número muito distante do intervalo**

d) Intervalo de [3 a 7] e valor procurado 1;

```
SQL> select graudepertinencia(3,7,1) from dual;
```

```
TESTE(3,7,1)
```

**O grau de pertinência é zero - número muito distante do intervalo**

No apêndice A, apresentamos um exemplo da utilização da “lógica nebulosa”, através de uma simulação da venda e compra de carros usados.

6.1 Porque a escolha do "data warehouse" para se utilizar a "lógica nebulosa" para tomada de decisão.

Segundo Chi e Turban (1995) a geração atual de Sistemas de informação executiva é composta por vários componentes de software, que possuem muitas características como: capacidade de "drill down" (detalhamento sucessivo das informações); acesso a dados históricos e correntes; uso extensivo de dados externos; sinalizadores gráficos; indicador de tendências, taxas e desvios; análises esporádicas ("ad hoc"); incorporação de gráficos e textos na mesma tela e capacidade de previsão (projeções e simulações).

O "data warehouse" tem todas essas características e elimina a possibilidade de uma diferença de comunicações entre tomadores de decisão e analistas onde se, por sua vez, o analista não entender as reais necessidades dos tomadores de decisão, os resultados das buscas de informações serão catastróficas.

O "data warehouse" permite interação fácil com os tomadores de decisão, e a seleção dos dados da "busca" é feita com o uso do mouse apontando o dado desejado, podendo-se organizar as informações disponíveis; a construção de expressões de filtros e cálculos que é feita por meio de um construtor inteligente de expressões orientado ao contexto e que permite seleção visual de: comandos, funções, sumariações, dados, constantes, etc, como também construções dinâmicas de diversos tipos de gráficos, é sistema desenvolvido para disponibilizar os dados históricos da empresa acrescidos das informações externas, permitindo o cruzamento de informações de dentro e externa à empresa.

O tomador de decisão em nenhum momento precisa escrever sentenças, o que torna o processo muito simplificado, intuitivo e com baixa probabilidade de erros.

Mostraremos no anexo B, algumas interfaces de um sistema com características do "data warehouse", sistema este, desenvolvido por um grupo de alunos da Universidade de Sorocaba, na disciplina, Trabalho de Conclusão de Curso, na qual sou um dos professores responsável pela orientação.

Para Inmon (1997) a extração de grandes volumes de dados acarreta um efeito benéfico em vários sentidos, pois o ambiente de produção tornar-se-á mais fácil de corrigir, reestruturar, monitorar e indexar, considerando os tipos de funcionalidade necessária, do ambiente operacional herdado para o ambiente de "data warehouse".

Na arquitetura do "data warehouse", ao representar a estrutura de dados, deve-se analisar a camada externa da base de dados transacional, que é representada por todos os

dados operacionais, que são os resultados das transações efetuadas diariamente pela organização.

Como essas bases de dados são muitas vezes antigas e projetadas para trabalhar no nível transacional, não suportam o acesso dos dados para fins de análise. O objetivo do "data warehouse" é liberar as informações que estão anexadas à base de dados transacionais e juntá-las às informações de outras fontes de dados da própria organização ou, mesmo, de fontes externas, como já discutido no capítulo 4. Os sistemas de "data warehouse" revitalizam os sistemas da empresa, pois, permitem que sistemas mais antigos continuem em operação, consolidam dados inconsistentes dos sistemas mais antigos em conjuntos coerentes, extraem benefícios de novas informações oriundas das operações correntes e também provêem ambiente para o planejamento e arquitetura de novos sistemas de característica operacional.

O tratamento de informações consiste em recuperar o que realmente interessa ao "data warehouse", portanto, deve-se iniciar a extração das informações das bases de dados existentes, com a preocupação da transformação, limpeza, observando as padronizações, integração dos dados de diferentes aplicações e, ao extrair dados operacionais, levá-los ao "data warehouse".

O "data warehouse", ao contrário de modelos operacionais, não necessita ser normalizado<sup>7</sup> e sempre que possível, as informações devem estar agregadas, para facilitar o entendimento do usuário final, e a desnormalização reduz a quantidade de ligações necessárias para elaboração de buscas e permitem para a "lógica nebulosa" a possibilidade de agregar nos resultados das buscas, informações próximas com seus graus de verdade através de algoritmos matemático e estatísticos.

Em relação aos dados históricos o seu carregamento ao "data warehouse" permite que se consiga fazer comparações entre vários períodos, agregando a essas comparações também valores próximos às buscas efetuadas.

A principal escolha pelo "data warehouse" é a sua não volatilidade, pois os dados ficarão em suas bases de dados por tempo necessário, sem alterações fazendo que, com isso as buscas de informações estarão visualizando dados com valores fixos, e isto seria impossível na base de dados transacional, pois seus valores são alterados a todo instante, permitindo aos tomadores de decisão, efetuarem comparações sem se preocuparem com as atualizações dos dados, fato esse que seria complicado guardar resultados em função das suas pesquisas com dados alterados a todo instante. Além de levarmos em consideração alguns critérios (efetuar a

---

<sup>7</sup> Normalizar: Date(86) processo usado para eliminar a redundância de dados.

condensação dos dados; a não extração de informações; pesquisar dados que apresentem marcas de tempo; necessidades de adicionar atributos e a não normalização) que já foram citados no cap. 4 item 4.3.1, devemos levar em consideração alguns critérios, para o melhor aproveitamento da “lógica nebulosa” em consideração à tomada de decisão quando executamos a extração de dados do sistema transacional para o “data warehouse”.

Nesta fase podemos avaliar a necessidade de se adicionar atributos que sejam calculados a partir de outros e da própria dimensão, de tal maneira que propiciem cálculos de grau de pertinências, atribuindo e criando as funções nebulosas que serão utilizadas nas buscas de informações, pertinentes conforme os algoritmos baseados em necessidades e confiabilidade.

Aliado a estas vantagens em pesquisar com “lógica nebulosa” em data warehouse e para atingir um grau de segurança na confiabilidade dos dados, devemos catalogar no Metadados todas as regras de pertinência e algoritmos de busca que levarão em consideração seus graus de pertinência, não esquecendo também: origem de dados, fluxo dos dados, formato de dados, definições de negócios, atualização dos dados históricos quanto a atualidade dos dados, nome e identificador de cada coluna e os critérios de classificação e atualização das agregações.

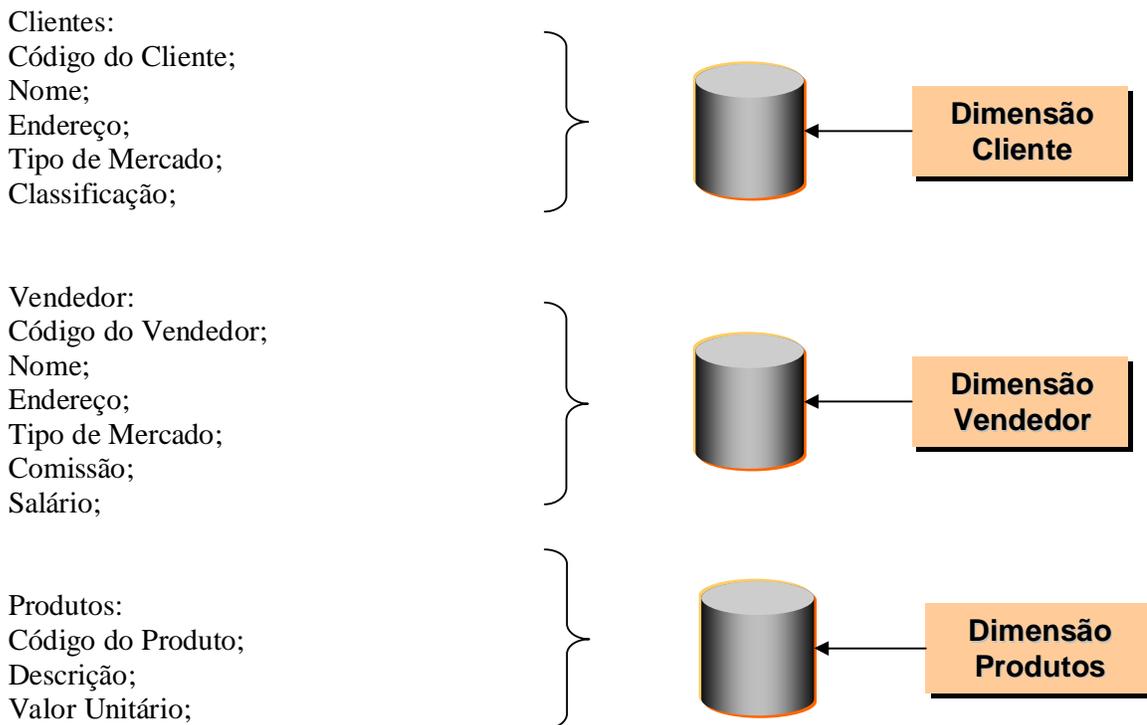
#### 6.1.1 Processos que serão levados em consideração da carga dos dados no “data warehouse”.

Processo	Descrição	Entradas
Extrair Dados (Externos e Históricos)	Extração dos Dados da base OLTP; Extração dos Dados Históricos; Extração dos Dados Externos;	Dados Extraídos
Armazenar Dados	Os Dados extraídos serão armazenados nas tabelas do Data Mart (Sistema Multidimensional)	Dados armazenados nas Tabelas.
Pesquisar Dados	Será efetuada a pesquisa dos dados requisitados pelo usuário na forma de consultas	Dados pesquisados à partir dos parâmetros
Apresentar dados	Serão apresentados os dados pesquisados em forma de consulta Esses dados poderão gerar relatórios e gráficos	Dados apresentados em forma de consulta
Gerar Relatórios	Serão impressos os dados apresentados na consulta na visão selecionada pelo usuário	Relatórios

Gerar Gráficos	Serão gerados gráficos à partir dos dados da consulta. Poderá ser de vários tipos: Barras, Pizza.	Gráfico
Atualizar Dados	Atualização dos dados do Data warehouse, que será realizada após requisição do usuário	Data warehouse Atualizado
Efetuar Backup	Será efetuada uma cópia de segurança dos dados do Data Warehouse a cada atualização. Restauração: Será efetuada uma cópia do backup para o Data Warehouse.	Backup dos dados do Data Mart (Vendas-OLAP) e restauração dos dados.
Funções de grau de pertinência	Construção de algoritmos para cálculo do grau de pertinência	

Exemplificamos a seguir, as dimensões que podem ser consideradas em relação à extração de dados dos sistemas tradicionais para o “data warehouse”, considerando a área de vendas.

Fluxos que entram no Processo de Extrair Dados, considerando, área de vendas.



Regiões:  
Código da Região;  
Descrição;  
Cidades;



**Dimensão  
Regiões**

Cidades:  
Código da Cidade;  
Nome;  
Quantidade de habitantes;  
Quantidade de estabelecimentos de cada tipo de mercado;  
Mercado de Alto Consumo;  
Mercado Tradicional;  
Mercado Frio;  
Dados sobre a situação econômica da população da cidade em questão;



**Dimensão  
Cidades**

Influência da Concorrência:  
Nome do Concorrente;  
Pontos onde vende;  
Método de Vendas;  
Perfil dos Vendedores;  
Método de Entrega;  
Qualidade dos produtos comercializados;



**Dimensão  
Concorrência**

Satisfação do Cliente:  
Nível de satisfação na venda;  
Nível de satisfação na entrega;  
Nível de satisfação pós-venda;



**Dimensão  
Satisfação**

Influências e Oportunidades na Situação Econômica do Brasil:  
Taxa da Inflação;  
Taxas de juros dos tipos de empréstimos disponíveis;  
Taxa para importação dos produtos;  
Taxa para exportação dos produtos;



**Dimensão  
Economia**

Influência e Oportunidades do câmbio de Moedas:  
Valorização da moeda nacional (R\$) em relação a moeda dos outros países;



**Dimensão  
Moeda**

Notamos nesta estruturação dos dados, que a composição dessa nova base de dados teria apenas as informações relevantes para o tomador de decisão.

## 6.2 Comparação de Resultados entre as lógicas booleana e nebulosa.

Precisamos, a fim de comprovar os benefícios do SQL modificado (ferramenta de busca já apresentado no capítulo 3, item 3.4), usando a “lógica nebulosa”, comparar resultados de consultas feitas a bancos cuja incerteza nos impele a adotar este tipo de estratégia. De fato, podemos pensar em uma base de dados de forma que os tipos de perguntas feitas à mesma tenham uma dose de incerteza que viria a prejudicar a confiabilidade final dos resultados retornados usando-se técnicas tradicionais.

Antes de tudo, portanto, é necessário definir os dados que serão utilizados em nossa comparação de resultados. Para tal, ilustraremos com um exemplo simples, retirado da dissertação de mestrado (SANTOS, Germano, 2004) que é um cadastro de pessoas contendo nome, idade e altura. Consultas pertinentes aos dois últimos campos são bastantes vagas, como é possível observar, já que termos relativos como “jovem”, “idoso”, “alto” e “baixo” pode admitir diferentes interpretações.

Supondo a seguinte tabela de funcionários (“EMPREGADOS”):

<b>Nome</b>	<b>Altura</b>	<b>Idade</b>
André	1.72	17
Ricardo	1.72	41
Paulo	1.70	14
José	1.66	19
Érico	1.80	30
Felipe	1.81	20
Fernando	1.55	20

Figura 6.02: Dados dos funcionários

<b>Nome</b>	<b>Altura</b>	<b>Idade</b>
Flávio	1.74	21
Gustavo	1.80	40
Rafael	1.59	19
Diogo	1.71	22
Gabriel	1.79	23
Fábio	1.71	58
Gilberto	1.95	43

Figura 6.03: Dados dos funcionários (cont.)

Digamos que queiramos selecionar os funcionários jovens e de estatura mediana.

Pelo SQL padrão, poderíamos formular uma busca como a que se segue: ***SELECT \* FROM EMPREGADOS WHERE HEIGHT BETWEEN 1,65 AND 1,75 AND AGE BETWEEN 15 AND 30.***

Estes valores foram obtidos usando-se, unicamente, o bom senso, e eles são dependentes do contexto (cultura, país etc.), ou seja, são relativos. Além do mais, para dificultar a questão, nem sempre apenas a idade pode determinar o quanto uma pessoa é jovem, uma vez que outros fatores podem contribuir. Neste caso, obviamente, uma aproximação genérica deve ser feita.

Os resultados da consulta em questão na tabela apresentada anteriormente são os seguintes:

Resultado da consulta usando o SQL padrão:

<b>Nome</b>	<b>Altura</b>	<b>Idade</b>
Diogo	1.71	22
Flávio	1.74	21
José	1.66	19
André	1.72	17

Figura 6.04: Resultado da consulta usando o SQL padrão

Aos dados retirados da dissertação de mestrado (SANTOS, Germano, 2004), passa-se, agora, a aplicação da “lógica nebulosa”.

Definindo-se a altura média (*medium*) como uma função, com parâmetros (1.50, 1.65, 1.75, 1.80) e idade jovem (*young*) como uma função também, cujos parâmetros são (13, 15, 23, 25), usamos o SQL Nebuloso para formular a consulta anterior da seguinte maneira: ***SELECT \* FROM EMPREGADOS WHERE HEIGHT IS MEDIUM AND AGE IS YOUNG.*** Após a tradução realizada pelo sistema utilizando os dados das funções previamente definidas e a conseqüente interpretação dos resultados, obter-se-ia a tabela que se segue:

Resultados da busca usando o SQL nebuloso

**Nome      Altura    Idade    Grau de pertinência**

Diogo	1.71	22	1.00
Flávio	1.74	21	1.00
André	1.72	17	1.00
José	1.66	19	1.00
Rafael	1.59	19	0.70
Paulo	1.70	14	0.50
Fernando	1.55	20	0.33
Gabriel	1.79	23	0.20.

Figura 6.05: Resultado da consulta usando o SQL nebuloso

Observa-se facilmente que na primeira consulta, com operadores booleanos, houve uma “perda” de informações, as quais poderiam ser relevantes. A primeira característica que chama a atenção é a quantidade maior de registros retornados pelo SQL nebuloso. Essa é uma conseqüência direta de sua fundamentação teórica, pois se tende a não desprezar dados, que podem ser aproveitados, aumentando, assim, o leque de possibilidades.

Outro ponto a observar é o acréscimo de mais um campo aos resultados: o **grau de pertinência**. Este é, justamente, o grau de pertinência já apresentado no capítulo 5. Dessa forma temos mais uma vantagem em relação ao SQL padrão, visto que nossos resultados podem ser ordenados decrescentemente de acordo com esse valor. Analisemos o caso

específico do funcionário Paulo, que foi desprezado na primeira consulta por causa de um ano. Muito embora ele pareça ser mais jovem que o desejado, talvez ele esteja perto de completar 15 anos, ou até mesmo pode não estar tão fora dos limites procurados. Concluimos, na verdade, pela sua classificação (grau 0.70), que ele não satisfaz completamente os requisitos buscados, mas, ainda assim, pode ser aproveitado. Este mesmo raciocínio pode ser estendido para os demais casos omitidos no primeiro resultado, mas que foram retornados no segundo. Em certas circunstâncias, pode ser perfeitamente aceitável o grau de pertinência de 70%.

### 6.3 Conclusões das análises entre “lógica nebulosa” e “lógica booleana”

Após analisarmos as diferenças de resultados e a forma que cada método em questão utiliza para tratar o mesmo problema, podemos perceber que fazendo uso dos conceitos da “lógica nebulosa” obtivemos respostas mais próximas das expectativas do usuário. O método desenvolvido no trabalho, apesar de todas as vantagens já apresentadas, pode vir a não corresponder plenamente aos objetivos da consulta, caso o arquivo de definições não tenha sido criado de forma apropriada, ou seja, sendo inconsistente, o que pode gerar resultados incoerentes. Logo, a eficiência do trabalho desenvolvido está intimamente ligada à precisão com que o usuário descreve as variáveis nebulosas e seus respectivos rótulos.

Isto posto, concluimos que, muito embora tenhamos feito uso de alguns exemplos simples, pudemos ter uma noção de todo o potencial que as consultas nebulosas possuem. Caso haja necessidade de se representar um modelo mais complexo no qual a incerteza é um fator preponderante, a tendência de se obter resultados precisos cresce de forma significativa. Apesar de ser um campo relativamente novo, encontramos cada vez mais aplicações que utiliza “lógica nebulosa”, e acreditamos que há muito a se pesquisar.

## CAPÍTULO 7 – Conclusões e considerações finais

Inicialmente quando pensamos na informação enquanto objeto de investigação da Ciência da Informação, pudemos constatar que a mesma pode ser identificada em diversos momentos, desde o processo mental que o indivíduo realiza para transformar dados em informações até o ato de comunicar e representar tal informação e os conhecimentos produzidos.

Isso significa que a área investiga os processos relativos à informação, desde sua geração e organização até sua recuperação, interpretação, transmissão, transformação e uso.

Essa visão também pôde ser percebida e reforçada por meio da análise dos dados coletados, que abordaram a informação enquanto conteúdos significantes capazes de gerar conhecimento no indivíduo. Em vários artigos pesquisados na área, a informação é considerada um estímulo para uma ação, que resulta na transformação de seu meio ou na tomada de uma decisão.

Mas o que se pôde perceber de mais concreto é a abordagem que se faz da informação enquanto matéria-prima na geração de conhecimento, e essa intensificação estejam relacionadas ao fato de estarmos vivendo em um período denominado Sociedade da informação.

Recuperar informação implica operar seletivamente um estoque de informação, o que envolve processos cognitivos que dificilmente podem ser formalizados através de um algoritmo. Mesmo que um modelo computacional de recuperação da informação tenha como base algum tipo de vocabulário e organização lógica, a equiparação dos significados supostamente implícitos pelos significantes depende de uma análise intelectual.

O estudo permitiu concluir que o “data warehouse”, é propício para o processamento de consultas, em oposição ao processamento de transações, de sistemas OLTP, que não podem ser repositórios de fatos e dados históricos para análises para tomada de decisão, pois estes últimos não podem responder rapidamente a consultas pois não permite uma recuperação rápida dos dados. Os dados estão em constante mudança, e existe uma ausência de dados históricos que são necessários para a análise de tendências. Basicamente, a OLTP oferece grandes quantidades de dados que estão disponíveis para o controle do dia-a-dia mas que não são adequados para tomada de decisão. O “data warehouse” oferece o potencial de recuperação e análise de informação rápida e fácil, pois os dados são armazenados em diferentes níveis de detalhamento, permitindo o acesso rápido aos mesmos e pode ser

explorado para a tomada de decisão pelos usuários finais sem dependerem de técnicos de informática para esse fim.

Concluimos também que o metadado, ou informação sobre os dados, é um elemento crítico em um sistema “data warehouse”. É a chave para entender o conteúdo e a estrutura desse tipo de armazenamento de dados, e é por meio dos metadados que um tomador de decisão acessa e manipula dados da empresa utilizando seus próprios termos e não o vocabulário técnico.

Mostramos que as consultas a banco de dados, quando realizadas utilizando lógica convencional, como no exemplo de utilização da linguagem padrão de acesso a banco de dados, *Structured Query Language* (SQL), tendem a retornar resultados, muitas vezes, distantes da realidade ou que não satisfazem as necessidades do tomador de decisão.

Notamos que isso acontece porque este sistema se baseia em operações clássicas de união, interseção, complemento e negação da álgebra booleana para decidir a respeito da pertinência de um elemento a um determinado conjunto. Pelo fato de só existirem duas possibilidades, pertence ou não, percebeu-se que esta aproximação está sujeita “perdas” de informação, e podem vir a comprometer a confiabilidade do resultado final, visto que no mundo real, na maioria das vezes, elementos podem pertencer parcialmente a um conjunto.

Neste trabalho apresentamos uma solução simples para este problema, incorporando à linguagem SQL os conceitos oriundos da “lógica nebulosa”. Observou-se facilmente que nos exemplos apresentados na consulta com “lógica booleana” há perdas de informações, as quais poderiam ser relevantes. A primeira característica que chamou a atenção foi a quantidade maior de registros retornados pelo SQL Nebuloso. Essa é uma consequência direta de sua fundamentação teórica, pois tende-se a não desprezar dados que podem ser aproveitados, aumentando, assim, o leque de possibilidades.

Outro ponto observado, através de algoritmos, é o retorno do valor do “grau de pertinência” que dá a possibilidade ao tomador de decisão optar em aceitar ou não a informação encontrada. Dessa forma temos mais uma vantagem em relação ao SQL booleano, visto que nossos resultados são acompanhados desse valor.

Após analisarmos as diferenças de resultados e a forma que cada método em questão utiliza para tratar o mesmo problema, podemos perceber que fazendo uso dos conceitos da “lógica nebulosa” obtivemos respostas mais próximas das expectativas do tomador de decisão.

O método desenvolvido no trabalho, apesar de todas as vantagens já apresentadas, pode vir a não corresponder plenamente aos objetivos da consulta, caso o arquivo de

definições não tenha sido criado de forma apropriada. Logo, a eficiência do trabalho desenvolvido está intimamente ligada à precisão com que se descrevem os algoritmos para definir os “graus de pertinência”.

Concluimos que, muito embora tenhamos feito uso de alguns exemplos simples, pudemos ter uma noção de todo o potencial que as consultas utilizando da “lógica nebulosa” possuem.

Concluimos também, que as vantagens do uso da “lógica nebulosa”, com o uso de variáveis lingüísticas nos deixam mais perto do pensamento humano, pois requerem poucas regras e decisões que simplificam a solução de problemas com variáveis imprecisas, simplificando a aquisição da base do conhecimento. As vantagens, portanto, superam as desvantagens de se ter mais simulações, testes e dificuldades de estabelecer regras corretamente por não ter uma definição matemática precisa.

Apesar de ser um campo relativamente novo, encontraremos cada vez mais aplicações que utilizam “lógica nebulosa”, e acreditamos que há muito a se pesquisar.

## 7.1 Perspectivas

A área de banco de dados passou, na última década, por algumas reformulações desde a transformação do relacional para os "data warehouse" de hoje. O mercado de "data warehouse" é um dos mais ativos no momento, e as empresas especializadas em construção e fornecimento de softwares, que são gerenciadores de banco de dados, estão lançando produtos novos e mais sofisticados em quantidade considerável. Com tais produtos espera-se que ocorram grandes avanços na área, especialmente em relação às melhorias de desempenho e qualidade das ferramentas para o usuário final.

Alguns pontos deverão melhorar significativamente, nos próximos anos, a tecnologia "data warehouse": 1) ferramentas para projeto de banco de dados multidimensionais; 2) ferramentas para extração de dados do "data warehouse" e sua administração; 3) ferramentas do tipo "data mining" para o usuário final; 4) desenvolvimento de servidores paralelos de banco de dados, que viabilizará um maior suporte a "data warehouse", permitindo um processamento mais rápido em suas consultas. Inmon (2004).

O armazenamento de dados pode auxiliar as empresas a disponibilizar os dados obtidos aos responsáveis pelas decisões, direcionando o rumo a ser tomado pelo empreendimento, porém só se percebe a verdadeira importância dos dados através da comunicação e colaboração entre indivíduos. Com a INTERNET, criou-se o maior depósito

de informações do planeta e isto mudou, de maneira dramática, a forma pela qual as pessoas se comunicam e trocam informações. O baixo custo por usuário, em tese, e o acesso da WWW (World Wide Web) por muitos já atingiu as mais diversas áreas de aplicação, verificando-se a possibilidade de viabilizar o acesso a "data warehouse" via Web; isto significa que uma empresa não precisaria investir em uma nova infra-estrutura, cada vez que desejasse dar acesso a um novo usuário. Certamente a Web vem impor novas demandas às tecnologias do "data warehouse", que deverão oferecer acesso eficiente a um grande número de usuários, garantindo segurança de acesso e facilidades gerenciais das informações.

Diversas áreas estão sendo beneficiadas pela tecnologia decorrente da "lógica nebulosa". Dentre essas áreas podem ser citadas algumas que tiveram relevância no avanço tecnológico e que merecem destaque. O Controle de processos industriais foi a área pioneira, sendo as primeiras experiências, quando foi demonstrado no Queen College, em Londres, que um controlador fuzzy muito simples conseguiu controlar eficientemente uma máquina a vapor.

A primeira aplicação industrial significativa foi desenvolvida pela indústria de cimento F.L.Smith Corp. da Dinamarca. Hoje em dia, uma grande variedade de aplicações comerciais e industriais está disponível, destacando-se neste cenário o Japão e mais recentemente, os EUA e a Alemanha. Dentre os exemplos típicos incluem produtos de consumo tais como geladeiras, ar condicionado, câmeras de vídeo, máquinas de lavar roupa, aspiradores de pó, etc.

Na indústria automotiva destacam-se transmissões automáticas, injeção eletrônica, suspensão ativa e etc. Sistemas industriais incluem controle de grupo de elevadores, veículos autoguiados e robôs móveis, controle de: motores, de tráfego urbano, de parada e partida de trens de metrô.

De fato, nos últimos anos o potencial de manuseio de incertezas e de controle de sistemas complexos tornados possíveis pela "lógica nebulosa", estão sendo combinados com Redes Neurais artificiais, que por sua vez, possuem características de adaptação e aprendizagem. A palavra certa para isto é simbiose, que vem gerando novas classes de sistemas e de controladores neurodifusos, combinando desta forma os potenciais e as características individuais em sistemas adaptativos e inteligentes. Gomide, (1998).

Aplicação da "lógica nebulosa" tem rapidamente se mostrada ser pertinente para desenvolver sistemas de controle sofisticados. Com a ajuda desses sistemas de controle, requisitos complexos podem ser implementados em controladores incrivelmente simples e fáceis de serem mantidos. A mesma "lógica nebulosa", na forma de argumentos aproximados,

está também ressurgindo na tecnologia de informação, onde fornecem apoio (suporte) nas decisões e sistemas especializados com poderosa argumentação, mas, com regras mínimas.

Neste instante, temos consciência de que o presente trabalho está longe de qualquer pretensão de ser exaustivo, sendo, pelo contrário, visto por seu autor como um possível ponto de partida para o aprofundamento e a ampliação da pesquisa em sistemas de informação e as tecnologia "data warehouse" com "lógica nebulosa", ampliando-se o estudo não só para segmentos da indústria e setor de serviços, como também para os sistemas de Internet e Intranet.

Finalmente, é importante destacar que este trabalho contribuiu muito para os conhecimentos do autor em relação à recuperação de informação, às teorias de "data warehouse", às teorias de "lógica nebulosa" e ao ambiente de suporte à decisão.

**Referências:**

ACKOFF, Russel L. **Management Science**. Vol. 14, n 4, Dezembro de 1967.

ALBERTIN, Alberto Luiz. **Administração de Informática**. Funções e Fatores Críticos de Sucesso. São Paulo: Atlas, 1996.

ALENCAR, Edgard de. **Teoria elementar dos conjuntos**. São Paulo: Livraria Nobel S. A. 1968.

ANSOFF, H. Igor. **Estratégia Empresarial**. São Paulo: McGraw-Hill do Brasil, 1977.

BARRETO, Aldo de A. **A eficiência técnica e econômica e a viabilidade de produtos e serviços de informação**. Mar 1996. <http://www.alternex.com.br/~aldoibct/pesquisa/>; data de acesso: 20 nov. 2004.

\_\_\_\_\_. **A questão da informação**. São Paulo em Perspectiva, v.8, n.4, p.3-8. 1994.

BISPO, Carlos Alberto. Transformando dados em informação via Data Mining. Developers, jan.1999.

BORKO, H. **Information Science: What is it? American Documentation**, v. 19, n. 1, p.3-5., 1968.

BUCKLAND, M. K.. **Information and Information Systems**. New York: Greenwood, 1991.

BUSH, V. (1945). **As we may think**. **The Atlantic Monthly**, v. 176, n. 1; pp 101-108., 1945. Disponível <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>>. Acessado em 10 nov. 2004.

CAMARA, Luiz. **Data Warehouse uma experiência no dia a dia**. Developers, fev.1997.

CAMPOS, Maria Luiza & ROCHA, Arnaldo V. **Desenvolvendo um data warehouse**. Rio de Janeiro, 1998.

CASTELLS, M. **A sociedade em rede**. 2ª edição. São Paulo: Paz e Terra, 1999.

CHI, R. T & TURBAN, E. **Distributed Intelligent Executive Information Systems**. Decision Support Systems. São Paulo: Atlas, 1995.

CHU, Shao Yong. **Banco de Dados: Organização, Sistemas e Administração**. São Paulo: Atlas, 1983.

COX, Earl. **The fuzzy Systems**. London. Ed. Chapman and Hall, 1988.

\_\_\_\_\_. **Fuzzy Logic for Business and Industry**. Massachusetts, 1995.

COREY, Michael J. & ABBEY, Michael. Oracle. Ware Warring. California, USA: McGraw – Hill, 1997.

CRUZ, Tadeu. **Sistemas de Informações Gerenciais. Tecnologia da Informação e a Empresa do século XXI**. São Paulo: Atlas, 1998.

DAGHLIAN, Jacob. **Lógica e Álgebra de Boole**. São Paulo: Atlas, 1995.

DAL'ALBA, Adriano. **Um Estudo sobre data warehouse**. Dissertação de Mestrado. Disponível no site ([http://www. Geocities.com/Silicom .Valley](http://www.Geocities.com/Silicom_Valley)), Rio de Janeiro acesso em 20 nov. 2004.

DATE, C. J.. **Banco de Dados: Tópicos Avançados**; Tradução Newton Dias Vasconcelos. Rio de Janeiro: Campus, 1988.

\_\_\_\_\_. **Guia para o padrão SQL**. Rio de Janeiro: Campus, 1989.

DAVIS, Willian S.. **Análise e Projeto de Sistemas. Uma Abordagem Estruturada**. Rio de Janeiro: Livros Técnicos e Científicos S.A., 1994.

ELMASRI. Ramez. **Sistemas de Bancos de Dados**. São Paulo: Pearson. 2005.

FERNANDES, M.C. & SANTOS, R.H. **Lógica Nebulosa X Lógica Paraconsistente**. Universidade de São Carlos. Disponível em: [www.dc.ufscar.br/~fernandes/trabalhos](http://www.dc.ufscar.br/~fernandes/trabalhos). Acessado em 10 jan.2005.

FERNEDA, E. **Construção automática de um thesaurus retangular**. Campina Grande. Dissertação (Mestrado em Informática), Universidade Federal da Paraíba., 1997.

\_\_\_\_\_. **Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. São Paulo. Tese de Doutorado. Universidade de São Paulo, 2003.

FONTES, Edson. **Protegendo a Informação. Fatores crítico para o negócio**. Developers, fev. 1998.

FOURNIER, Roger. **Guia Prático para o Desenvolvimento e Manutenção de Sistemas Estruturados**. São Paulo: Makron Books,1994.

FURLAN, José David. **Planejamento Estratégico de Sistemas de Informações Executivas**. São Paulo: Makron Books,1994.

GANES, B. **Foundationsof fuzzy reasoning, International journal of Man-Machine. Studies**,1976.

GATES, Bill et allii. **A estrada do futuro**. Trad.Beth Vieira et al.São Paulo: Companhia das Letras, 1995.

GOMIDE, Fernando. **An Introdution to fuzzy sets: analysis and design**. London, Bradford Books, 1998.

HABERKORN, Ernesto. **O computador na administração de empresas**. São Paulo: Editora Atlas, 1985.

HAACK, Susan. **Filosofia das lógicas**. São Paulo. Editora Unesp, 2002.

HAMEL, C. K. Gary. **Competindo pelo Futuro**. Rio de Janeiro. Campus, 1995.

HARJINDER, G. & RAO, P. C.. **The Official Guide to Data warehousing**. Que: Corporation, 1996.

HELLER, Robert. **Os Tomadores de Decisão**. Rio De Janeiro: Makron Books,1991.

HERRERA, Antonia Heredia. **Archivística general. Teoría y Práctica**. Sevilla: Diputación Povincial, 1991.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados**. Porto Alegre: Sagra Luzzatto. 1998.

INMON, William H.. **Data warehousing. Como transformar informações em oportunidades de negócios**. São Paulo: Berkeley, 2004.

\_\_\_\_\_. **Como Construir o data warehouse**. Rio de Janeiro: Campus, 1997.

JANNUZZI, Celeste A. S. Correa. Gesta. **Gestão da informação nas empresas: uma abordagem conceitual para a Ciência da Informação**. São Paulo. USP, Tese de Doutorado, 2004.

KIMBALL, Ralph. **The data warehouse Toolkit**. New York: Jhon Willey & Sons, 1997.

KROENKE, David M.. **Banco de Dados**. Fundamentos, Projeto e Implementação Rio de Janeiro: Livros Técnicos e Cientificos Editora S.A.1998.

LAUDON, K. C. & LAUDON, J. P.. **Sistemas de informação Gerencial**. Administrando a Empresa Digital. 5a. Edição, São Paulo. Prentice Hall, 2004.

\_\_\_\_\_. **Management Information Systems. New Approches to Organization and Technology**. New Jersey: Prentice Hall, 1998.

\_\_\_\_\_. **Management Information Systems. Transforming Business and Manegement** . New Jersey: Prentice Hall, 1998.

LE COADIC, Y-F. **A Ciência da Informação**. Brasília: Briquet de Lemos, 1996.

LÉITE, Leonardo L. Pereira. **Introdução aos Sistemas de Gerência de Banco de dados.** São Paulo: Editora Edgard Blücher Ltda, 1980.

LÉVY, Pierre. **As Tecnologias da Inteligência:** o futuro do pensamento na era da Informática. Trad. Carlos Irineu da Costa. Rio de Janeiro: Editora 34, 1993.

LUGO, Gustavo A. G. **Recuperação de Informação voltada a Computação Nebulosa.** São Paulo, 2003.

MCGARRY, K. **O contexto dinâmico da informação.** Brasília: Briquet Lemos, 1999.

MEIRELLES, Fernando de Souza. **Informática. Novas Aplicações com Microcomputadores.** São Paulo: Makron Books, 1994.

MELO, Ivo Soares. **Administração de Sistemas de Informação.** São Paulo, Pioneira Thomson, 2002.

MELO, Rubens Nascimento e TANAKA, Asterio K.. **Banco de dados em aplicações cliente/servidor.** Rio de Janeiro: Infobook. 1997.

MOOERS, C. **Zatocoding applied to mechanical organization of knowledge.** American Documentation, v. 2, n. 1, p.20-32, 1951.

MULLER, Robert. J. **Projeto de Banco de Dados. Usando UML para modelagem de dados.** São Paulo: Burkeley 2002.

NIMER, Fernando. **Obtendo vantagem competitiva com o uso de Datamining.** Developers, fev. 1998.

OLIVEIRA, Adelize Generini de. **Data Warehouse Conceitos e Soluções.** Florianópolis. SC: Advanced, 1998.

OLIVEIRA, Djalma de P. R.. **Sistemas de Informações Gerenciais. Estratégias, Táticas, Operacionais.** São Paulo: Atlas, 1998.

PORTER, Michael E. **Estratégia Competitiva. Técnicas para Análise de Indústria e da Concorrência.** Rio de Janeiro: Campus, 1986.

PRESSMAN, Roger S. **Engenharia de Software.** Rio de Janeiro: McGraw-Hill, 2002.

RABUSKE, Renato Antonio. **Inteligência Artificial.** Florianópolis. Ed. da UFSC, 1995.

ROBREDO, Jaime. **Da Ciência da Informação revisitada aos sistemas humanos de informação.** Brasília: Thesaurus. 2003. p. 1-102.

\_\_\_\_\_. **Documentação de hoje e de amanhã: uma abordagem informatizada da biblioteconomia e dos sistemas de informação.** São Paulo: Global, 1994.

ROSINI, Alessandro Marco. **Administração de Sistemas de Informação e a Gestão do Conhecimento.** São Paulo: Pioneira Thomsom, 2003.

SALTON, G. e MCGILL, M. J.. **Introduction to Modern Information Retrieval.** Rio de Janeiro: Ed. McGraw Hill, 1983.

SANTOS, Germano. J. Carvalho. **Lógica Fuzzy.** Dissertação de Mestrado. Disponível no site (<<http://www.uesc.br/arbels/dme.news.html>>), Universidade Estadual de Santa Cruz, Bahia. acesso em 15 nov. 2004.

SANTOS, Luiz Alberto A. **Planejamento e Gestão Estratégica nas Empresas.** São Paulo: Atlas, 1992.

SARACEVIC, Tefko. **Interdisciplinary nature of information science. In:\_\_\_\_\_ Information Science: origin, evolution and relations.** Los Angeles. CA Taylor Graham, 1992. p. 36-41.

\_\_\_\_\_. **Information Science. Journal of American Society for Information Science.** 1999.

SAVIANI, José Roberto. **O Analista de Negócio e da Informação.** São Paulo: Atlas, 1992.

SHANNON, C. e WEAVER, W. **The Mathematical theory of communication**. University of Illinois Press, 1949.

SHERA, J.H e CLEVELAND, D.B. **History and foundations of Information Science. Annual Review of Information Science and Technology**, v. 12, p.249-275, 1977.

SHIMIZU, Tamio. **Decisão nas organizações: introdução aos problemas de decisão encontrados nas organizações e nos sistemas de apoio à decisão**. São Paulo: Atlas, 2001.

SOARES, Vânia Jesus de Araujo. **Detalhamento e Evolução das Dimensões de um Data warehouse**. Rio de Janeiro: McGraw-Hill, 1998.

STAIR, Ralph M. **Princípios de Sistemas de Informação**. Rio de Janeiro: Livros Técnicos e Científico Editora S. A., 1998.

TANLER, Rick. **Intranet data warehouse**. São Paulo: Infobook, 1998.

TOFFLER, Alvin. **Power Shift. As Mudanças do Poder**. Rio de Janeiro: Record, 1995.

ZADEH, L.A. **Fuzzy sets. Information and Control**. v. 8, n. 3, p.338-353, 1965.

\_\_\_\_\_. **Fuzzy logic and approximate reasoning**, Synthese, 30.

## Apêndice A

### Exemplo do uso da “lógica nebulosa” para venda/compra de carro usado.

**Sistema de Avaliação de Veículos**

Preço Veículos | Avaliações | Configurações | Sobre...

**Seleção a Marca**  
 Seleção a Marca

**Seleção o Modelo**  
 Seleção o Modelo

**Seleção o Ano**  
 Seleção o Ano

➔ Avaliar Veículo

✎ Nova Consulta

**Avaliação Total do Veículo**  
**Valor Tabela:**  
**Valor Depreciação:** Lbl Total  
**Valor Total:** ValDepre

Figura A,01: Selecionar marca, modelo e ano do carro.

**Sistema de Avaliação de Veículos**

Preço Veículos | Avaliações | Configurações | Sobre...

**Seleção a Marca**  
 Seleção a Marca

RENAULT  
 Seat  
 Subaru  
 Suzuki  
 TOYOTA  
 Volkswagen  
 Volvo  
 VW

➔ Avaliar Veículo

✎ Nova Consulta

**Avaliação Total do Veículo**  
**Valor Tabela:**  
**Valor Depreciação:** Lbl Total  
**Valor Total:** ValDepre

Após selecionarmos: marca, modelo e ano do carro necessitam, fazer avaliação das do veículo, considera-se a quilometragem (que é um valor conhecido), pois se encontra no tacógrafo e segue a avaliação, então associa-se a “lógica nebulosa”, pois, ao se avaliar os

itens: parte elétrica, funilaria, motor, etc. os valores não são exatos e a avaliação é feita considerando por ex. os termos: “quase bom”, “acima do regular”, “perto do bom”, “ótimo”, palavras que usamos no dia a dia, e assim será feito a avaliação dos demais itens. (Figura A, 02).

The screenshot shows a software interface titled "Sistema de Avaliação de Veículos" with a menu bar containing "Preço Veículos", "Avaliações", "Configurações", and "Sobre...". The main area is divided into several sections for evaluation:

- Kilômetragem:** Radio buttons for "1 à 25.000 KM" (selected), "25.001 à 50.000 KM", "50.001 à 75.000 KM", "75.001 à 100.000 KM", and "acima de 100.000 KM".
- Elétrica, Funilaria, Motor, Pintura:** Each has a vertical slider with four levels: Ruim, Regular, Bom, and Ótimo. The sliders are positioned between Regular and Bom.
- Pneu, Suspensão, Tapeçaria, Transmissão:** Each has a vertical slider with four levels: Ruim, Regular, Bom, and Ótimo. The sliders are positioned between Bom and Ótimo.
- Vidros:** Has a vertical slider with four levels: Ruim, Regular, Bom, and Ótimo. The slider is positioned between Regular and Bom.
- Confirma:** A button located at the bottom left of the evaluation area.

Figura A, 02: Avaliar itens.

Neste sistema esses “valores” seriam transformados em graus de pertinência entre [0 e 1] e através de algoritmos, levando em consideração todos os valores pertinentes a venda/compra de carros, conseguiria o valor da depreciação e teríamos o valor de venda/compra do carro avaliado.

Os valores considerados nesse exemplo, são para fins didáticos.

**Sistema de Avaliação de Veículos**

Preço Veículos | Avaliações | Configurações | Sobre...

**Selecione a Marca**  
VW

**Selecione o Modelo**  
Gol 1.0 4p.

**Selecione o Ano**  
2001

➔ Avaliar Veículo

🔍 Nova Consulta



**Avaliação Total do Veículo**

<b>Valor Tabela:</b>	18487
<b>Valor Depreciação:</b>	5083.925
<b>Valor Total:</b>	13403.075

Figura A, 03: Resultado da avaliação.

Após os cálculos para esse exemplo teríamos valores:

Valor tabela: R\$ 18487,00

Valor depreciação: R\$ 5083,92

Valor de Venda/Compra: R\$ 13403,00

## Apêndice B

Nesta função do sistema, estarão disponibilizados todas as opções de acordo com os dados inseridos no “data warehouse” do setor de “Vendas”, para que o usuário efetue buscas de informação para sua tomada de decisão.

Na busca das informações, será usado internamente comandos de banco de dados que aparecem em nosso exemplo, mas que seriam transparentes para o usuário final, para cada busca nova será montado um novo “select” (comando de busca em banco de dados), cujo comando será montado dinamicamente (em tempo de execução do sistema, sem a interferência de programadores ou analistas do Centro de Processamento de Dados da Organização), e sim de acordo com o que o usuário final selecionar.

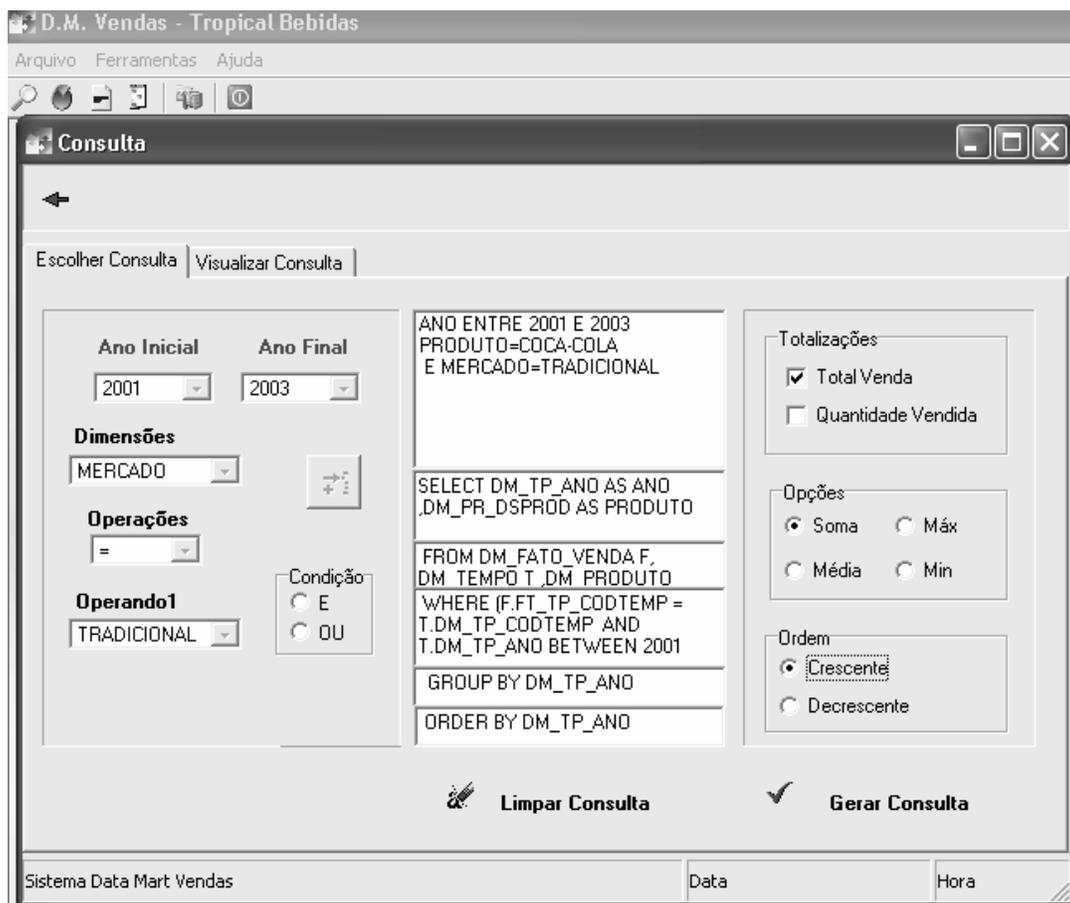


Figura B, 01: Módulo gera consulta

Será obrigatória a escolha de ano inicial e ano final no início de cada consulta.

A caixa de condição só ficará ativa quando já tiver uma dimensão inserida nos memos de visualização.

Para gerar consulta, o usuário deverá escolher uma totalização, uma opção de totalização e a ordem de visualização.

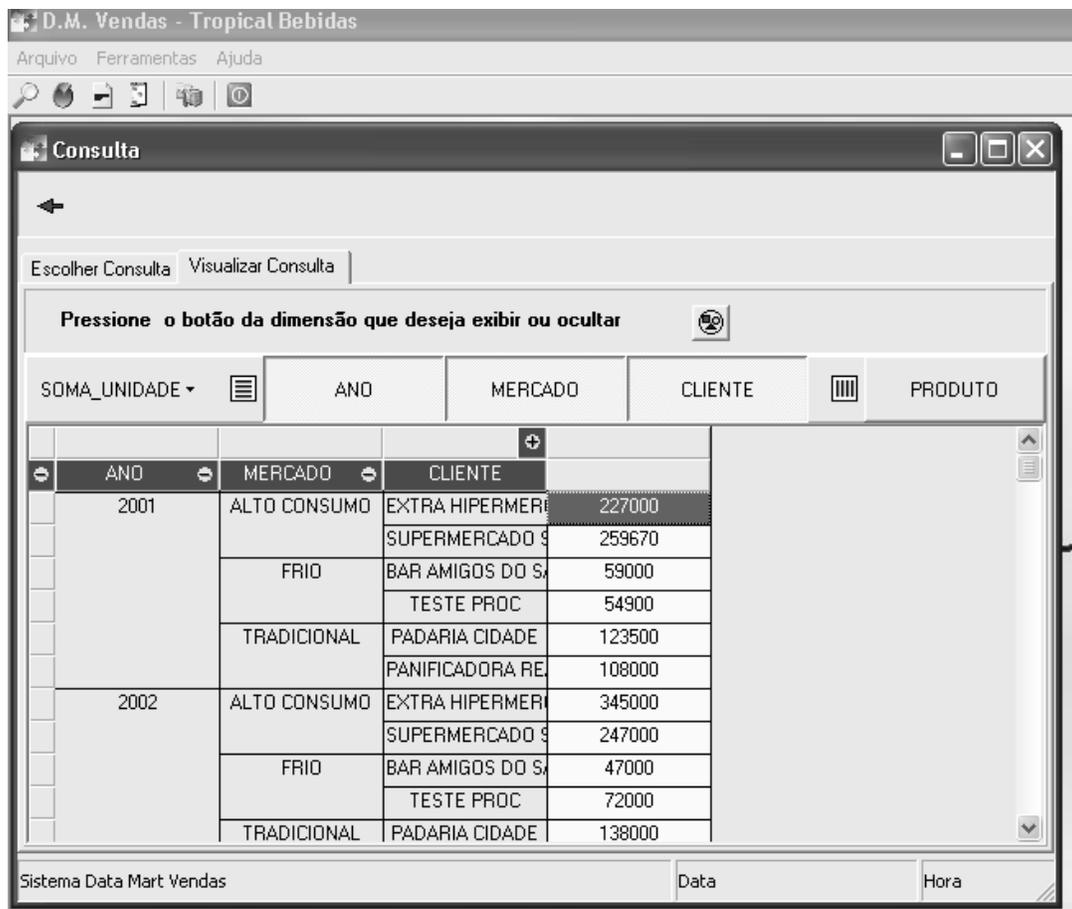


Figura B, 02: Módulo consulta

Este módulo permitirá a quaisquer usuários a geração de consultas dinâmicas com o cruzamento de várias informações disponíveis no “data warehouse” de Vendas.

Nesta tela serão visualizados os dados resultantes das opções de consultas que serão gerados de acordo com a escolha do tomador de decisão, a partir de então, poderão ser gerados gráficos dinâmicos e relatórios. Nesta opção serão efetuadas as operações do “data warehouse”, como: drill down; dril up; slice e dice, através dos botões: (ano, mercado, cliente e produto)

Com a consulta serão mesclados dados do Sistema da Empresa, dados externos e dados dos concorrentes, permitindo comparações sobre diversos aspectos.

No módulo de consultas, na aba “escolher consulta” o usuário determinará o tipo de consulta que deseja fazer, não ficando restrito a parâmetros, que possam limitar o tipo e a densidade da informação coletada.

Para a formulação das consultas o usuário deverá obrigatoriamente escolher um “ano para início e término”, sendo este o único parâmetro que é necessário, pois todas as consultas têm como base principal o tempo, pois os fatos ocorrem em uma determinada data, para comparações com dados futuros.

Após a determinação do intervalo de anos o usuário terá disponível a escolha das dimensões como (Produtos, Clientes, Mercado) e logo depois a operação desejada (=, <>), e na seqüência a próxima escolha será o tipo de Produto, Mercado ou Cliente dependendo do que foram escolhidas na caixa dimensões.

Logo após a escolha da Dimensão, Operante, Operação 1 o usuário deverá clicar em um botão que validará a escolha e se o usuário desejar escolher mais alguma dimensão deverá selecionar na condição (E, OU) e repetir a operação de escolha das dimensões.

O botão está disponível e habilitado logo após o usuário escolher a Dimensão, Operando e Operação 1 e ao ser clicado efetivará a escolha feita pelo usuário adicionando o que foi escolhido, a caixa de texto ao lado.

Na aba “totalizações” o usuário poderá escolher entre Total de Vendas e Total de Quantidade Vendida, ou até mesmo as duas totalizações para serem efetuadas na consulta.

Com a escolha dos dados, o usuário poderá escolher as totalizações como Total de Vendas e Quantidade Vendida, tendo seguidamente as opções de agrupamento com (SOMA, MÉDIA, MÍNIMO, MÁXIMO) e a ordem da consulta com os dados crescente ou decrescente.

Na aba “ordem” o usuário poderá escolher a ordenação dos dados da consulta tanto como crescente ou decrescente.

No botão “Limpar Consulta” o usuário poderá cancelar a consulta feita limpando as caixas disponíveis na tela, para ser efetuada uma nova consulta.

Ao clicar no botão “Gerar Consulta” ocorrerá à visualização prévia dos dados pesquisados com sua disposição.

D.M. Vendas - Tropical Bebidas

Arquivo Ferramentas Ajuda

Consulta

Escolher Consulta Visualizar Consulta

Pressione o botão da dimensão que deseja exibir ou ocultar

SOMA\_UNIDADE ▾ PRODUTO MERCADO CLIENTE ANO

			ANO		
PRODUTO	MERCADO	CLIENTE	2001	2002	2003
CERVEJA KAISER	ALTO CONSUMO	EXTRA HIPERMER	119000	176000	206000
		SUPERMERCADO S	138000	126000	162000
	FRIO	BAR AMIGOS DO S	34000	25000	55000
		TESTE PROC	26900	39000	34000
	TRADICIONAL	PADARIA CIDADE	68000	71000	115000
		PANIFICADORA RE	56000	90000	131000
COCA-COLA	ALTO CONSUMO	EXTRA HIPERMER	108000	169000	199000
		SUPERMERCADO S	121670	121000	159000
	FRIO	BAR AMIGOS DO S	25000	22000	55000
		TESTE PROC	28000	33000	34000

Sistema Data Mart Vendas Data Hora

Figura B, 03: Módulo consulta

Nesta tela serão visualizados os dados resultantes das opções de consultas que serão gerados de acordo com a escolha do tomador de decisão, igual a consulta anterior usando os conceitos: drill down e drill up.

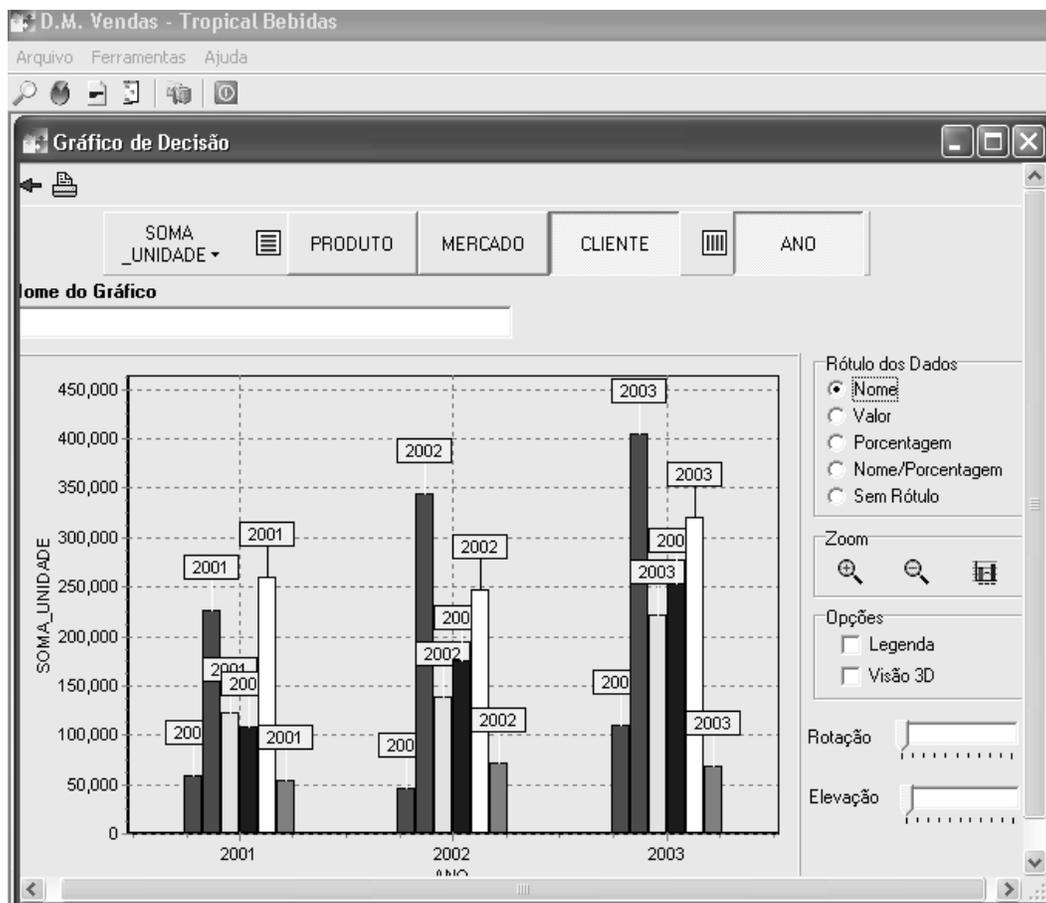


Figura B, 04: Módulo gráfico

O módulo de Gráfico será um complemento para a visualização mais detalhada da consulta. Com o Gráfico a consulta será feita visualmente, através de cores e formas que identificarão de uma melhor maneira o comportamento de vendas e de consumo dos clientes/empresas.

Na caixa “campos para agrupamento” o usuário terá a opção de escolher dentre os campos pesquisados na consulta, qual será a base para a visualização do gráfico.

Na caixa “valor numérico”, deverá ser escolhido os totais disponíveis no sistema para montar o eixo de visualização e perspectiva do gráfico, podendo escolher total em reais de vendas ou total de quantidade neste sistema.

Com a caixa “especificar ano”, assim como no relatório o usuário poderá fazer uma consulta com um intervalo de anos e no gráfico mostrar somente um dos anos do intervalo para uma definição mais clara e específica da consulta.

Com a caixa “especificar mercado”, o usuário poderá fazer uma consulta com todos os mercados e no gráfico mostrar somente um deles, para uma definição mais clara e específica da consulta.

O usuário terá a opção de digitar um título para o gráfico na Caixa de Texto “nome do gráfico” para sua personalização de acordo com o que o usuário deseja.

Em “tipo de gráfico” ter-se-á várias opções de gráfico para que o usuário escolha aquele que seja mais eficaz e mostre com mais definição a oscilação da informação pesquisada, tendo como opções: Área, Barra, Barra Horizontal, Linha E Pizza.

“Rótulo do gráfico” permite a identificação das informações que serão mostradas ao usuário, funciona como uma identificação das informações sobre a pesquisa utilizada naquele instante. As opções de Rótulo são: Nome, Valor, Porcentagem, Nome/Porcentagem, Nome/Valor, Sem rótulo.

O “zoom” permitirá a variância de tamanho em que o gráfico será mostrado, podendo ser aumentado, diminuído ou mostrado em tamanho normal.

Na aba “opções” mais uma vantagem que o usuário poderá usufruir com a escolha da exibição de legenda ou com a visualização do gráfico em 3D, para uma melhor formatação da informação.

O gráfico será exibido conforme as especificações desejadas pelo usuário.

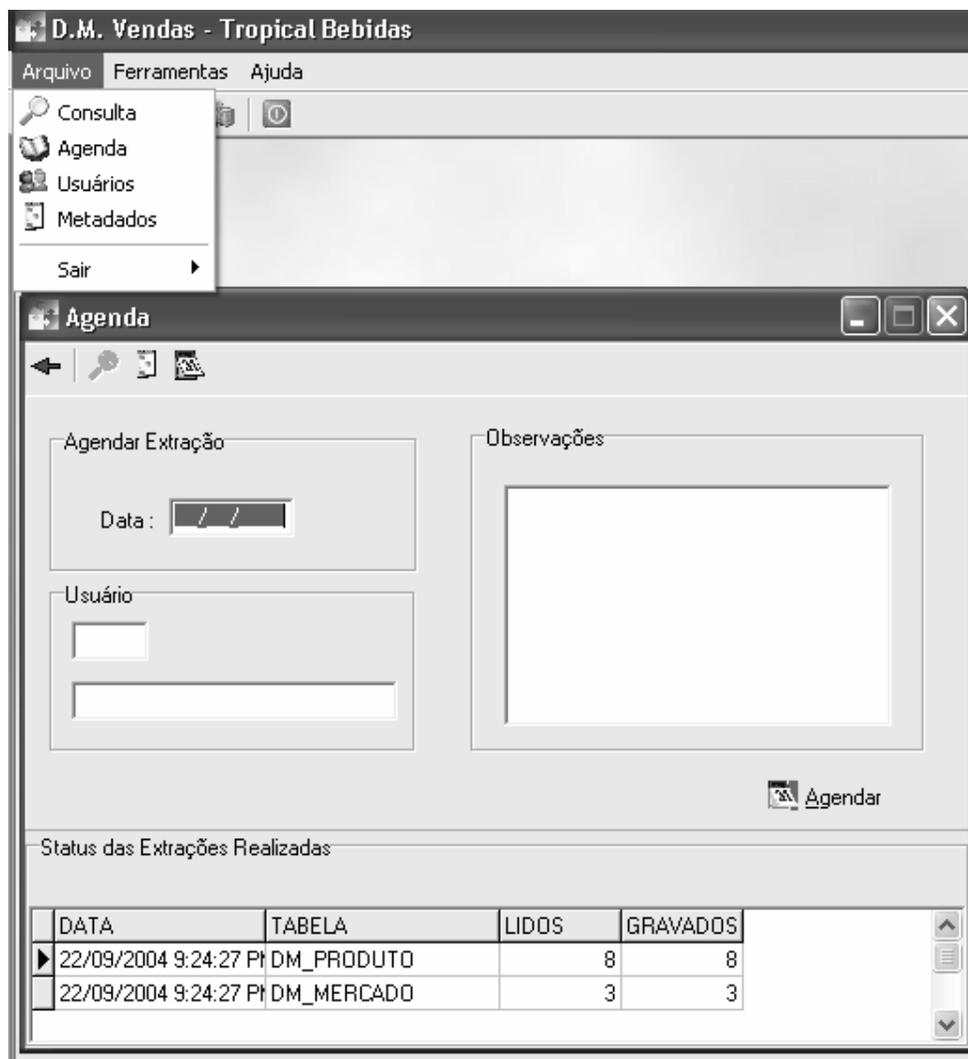


Figura B, 05: Módulo agenda

Esta tela permite ao usuário escolher um novo dia para ser refeita a extração dos dados internos e externos para atualizar o “data warehouse”.

Na aba “agenda extração” o usuário poderá agendar uma data para uma extração que deseje fazer sem ter sido previamente agendado.

Na aba “observações” o usuário poderá digitar observações que queira adicionar quando for agendar uma data para a extração.

No botão Agendar o usuário efetuará o agendamento que foi pré-definido nos passos anteriores.

Em “status das extrações realizadas” o usuário terá a descrição de todas as extrações que foram feitas até então.

Ao clicar neste botão o usuário terá a informação de todos os agendamentos que estão feitos, para não marcar outro na mesma data.

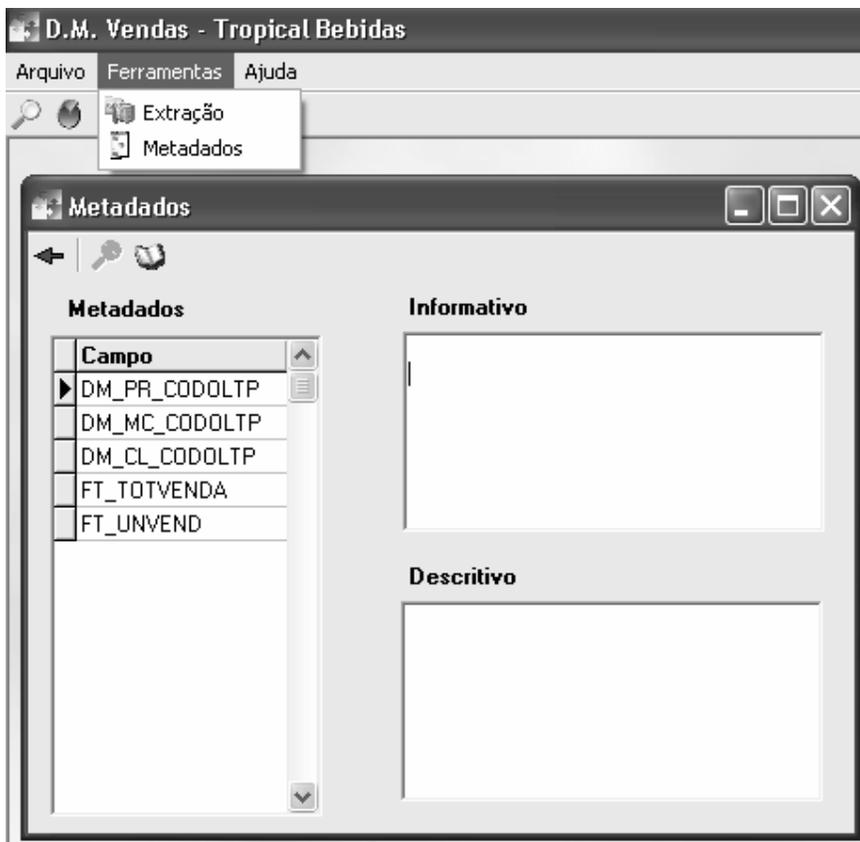


Figura B, 06: Módulo metadados

Esta mostra explicações de cada campo, para o caso de esquecermos algum detalhe do sistema.

Na aba “metadados” estão disponíveis todos os campos utilizados no sistema.

Na aba “informativo” as definições a respeito do local onde este campo está alocado e sua utilização.

Na aba “descritivo” relata o que significa cada campo utilizado com sua descrição.

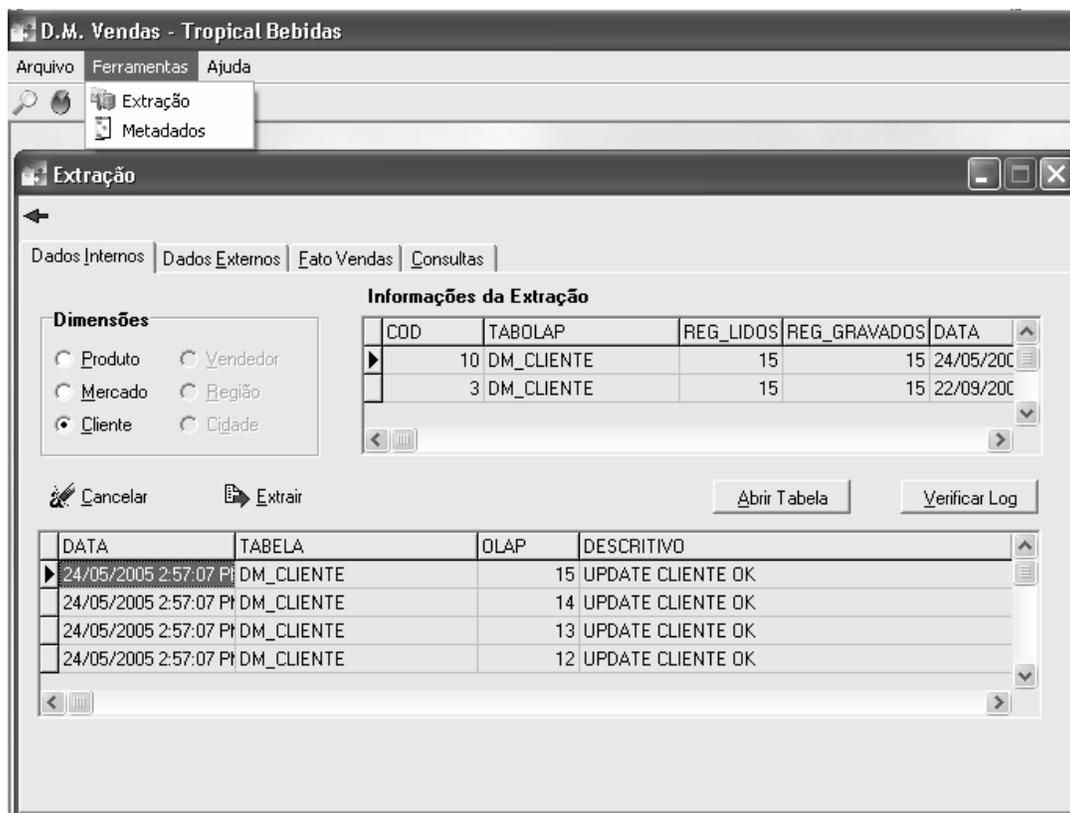


Figura B, 07: Módulo extração

Esta função do sistema permitirá a extração dos dados do Sistema Transacional, que serão armazenados e agregados aos dados no “data warehouse”, sobre as vendas dos produtos; dados que são, denominados dados históricos.

Na aba “agenda extração” o usuário poderá agendar uma data para uma extração que deseje fazer sem ter sido previamente agendado.

Na aba “observações” o usuário poderá digitar observações que queira adicionar quando for agendar uma data para a extração.

No botão Agendar o usuário efetuará o agendamento que foi pré-definido nos passos anteriores.

Em “status das extrações realizadas” o usuário terá a descrição de todas as extrações que foram feitas até então.