UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA

ROBERTA MACÊDO MARQUES GOUVEIA

Dissertação de Mestrado

João Pessoa-PB
Maio-2009

ROBERTA MACÊDO MARQUES GOUVEIA

MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA

Dissertação de mestrado apresentada ao Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, como requisito parcial para obtenção do título de Mestre em Informática (Sistemas de Computação).

Orientadora: Professora Dra. Valéria Gonçalves Soares Elias
Co-orientador: Professor Dr. Heber Pimentel Gomes

João Pessoa-PB
Maio-2009

G719m Gouveia, Roberta Macêdo Marques.
Mineração de dados em data warehouse para sistema de abastecimento de água / Roberta Macedo Marques Gouveia. João Pessoa, 2009.
147f. : il.
Orientadora: Valéria Gonçalves Soares Elias.
Co-orientador: Heber Pimentel Gomes.
Dissertação (Mestrado) – UFPB/CCEN
1. Data warehouse – Banco de dados. 2. Mineração de dados. 3. Tecnologias OLAP.

UFPB/BC CDU: 004.65 (043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado da Roberta Macedo Marques Gouveia, candidata ao Título de Mestre em Informática na Área de Sistemas de Computação, realizada em 29 de maio de 2009.

Aos vinte e nove dias do mês de maio do ano dois mil e nove, às oito horas, na Sala de Reuniões do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar a candidata ao grau de Mestre em Informática, na área de "*Sistemas de Computação*" e na linha de pesquisa "*Computação Distribuída*", a Sra. Roberta Macedo Marques Gouveia. A comissão examinadora foi composta pelos professores doutores: Valéria Gonçalves Soares (DI-UFPB), Orientadora e Presidente da Banca Examinadora, Lucídio dos Anjos Formiga Cabral (DI-UFPB), Ed Porto Bezerra (DI-UFPB) e Heber Pimentel Gomes (UFPB), como examinadores internos e Sônia Virginia Alves França (UFRPE), com examinadora externa. Dando início aos trabalhos, a Profª. Valéria Gonçalves Soares, cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra à candidata para que a mesma fizesse, oralmente, a exposição do trabalho de dissertação intitulado "*MINERAÇÃO DE DADOS EM DATA WAREHOUSE PARA SISTEMA DE ABASTECIMENTO DE ÁGUA*". Concluída a exposição, a candidata foi argüida pela Banca Examinadora que emitiu o seguinte parecer: "*Aprovada*". Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, professor José Antônio Gomes de Lima, membro do Colegiado deste Programa, representando a coordenação do PPGI, lavrei a presente ata que vai assinada por mim mesmo e pelos membros da Banca Examinadora. João Pessoa, 29 de maio de 2009.

José Antônio Gomes de Lima

Profª. Dra. Valéria Gonçalves Soares
Orientadora (DI-UFPB)

Prof. Dr. Lucídio dos Anjos Formiga Cabral
Examinador Interno (DI-UFPB)

Prof. Dr. Ed Porto Bezerra
Examinador Interno (DI-UFPB)

Prof. Dr. Heber Pimentel Gomes
Examinador Interno (UFPB)

Profª. Dra. Sônia Virgínia Alves França
Examinador Externo (UFRPE)

AGRADECIMENTOS

A Deus pelo dom da vida e pelas oportunidades concedidas em minha vida, permitindo-me enveredar pelo caminho da ciência e do saber, e dando-me o alento necessário para prosseguir. Nossa aliança é eterna!

À Nossa Senhora, pelo seu grande exemplo de vida, mostrando-me o caminho da fé, superação, esperança, tolerância, doação e principalmente, seu exemplo de amor.

Aos meus pais que tanto amo, Severino M. Gouveia e Ilsaira M. M. Gouveia, pelo exemplo de dedicação, amizade, amor incondicional e investimento dispensado ao longo da minha formação.

Ao Prof. Dr. Heber Pimentel Gomes um agradecimento muito especial pelas orientações e pela amizade conquistada ao longo desses dois anos juntos ao Laboratório de Eficiência Energética e Hidráulica em Saneamento - LENHS. Seus ensinamentos e motivações foram significantes para a conclusão deste trabalho.

À Prof^{a}. Dra. Valéria Gonçalves Soares Elias pelas orientações sugeridas, das quais foram úteis ao desenvolvimento desta pesquisa.

Aos meus irmãos Bruno M. M. Gouveia e Rafael M. M. Gouveia pelo apoio e harmônica convivência, me incentivando a seguir em frente e fornecendo todo o sustentáculo.

Ao meu amado Alexandre Magno Gurgel Fialho pelo amor, dedicação, apoio, carinho e compreensão em todos os momentos.

Aos meus amigos e colegas da UFPB, em especial a toda equipe do LENHS – dentre eles, Moisés M. Salvino, Paulo Sérgio O. Carvalho, Saulo B. de Tarso, Magno J. G. Silva e Wil L. L. Camboim – pelo auxílio, incentivo e companheirismo.

Ao Governo do Brasil, pelo apoio financeiro concedido através das Centrais Elétricas Brasileiras S.A. (**ELETROBRÁS**), da Financiadora de Estudos e Projetos (**FINEP**) e do Conselho Nacional de Desenvolvimento Científico (**CNPq**).

À Companhia de Água e Esgotos da Paraíba (CAGEPA), em nome dos engenheiros Leonardo L. B. Montenegro e Jaqueline Pequeno, pela disponibilização dos dados necessários ao estudo de caso do trabalho.

À UFPB, instituição que, através de seus docentes e funcionários, foi responsável pela minha formação acadêmica. E aqueles que contribuíram de alguma forma para a realização deste trabalho. **Muito Obrigada!**

RESUMO

Esta dissertação se propõe a utilizar tecnologias de Banco de Dados com a finalidade de oferecer apoio à decisão para os gestores do setor de saneamento, haja vista que os serviços de abastecimento de água para uso da população se constituem em um dos principais indicadores da qualidade de vida da humanidade. A idéia fundamental consiste em coletar os dados operacionais, reduzi-los ao escopo de um problema, organizá-los em um repositório de dados, e finalmente aplicar as tecnologias OLAP e os algoritmos de Mineração de Dados, a fim de obter resultados que proporcionem aos gestores um melhor entendimento do comportamento e perfil da companhia. Para facilitar a aplicação de técnicas de Mineração de Dados é necessário que estes dados estejam armazenados apropriadamente. Neste sentido, uma das alternativas para o aumento da eficiência no armazenamento, gestão e operação dos dados para o suporte a decisão baseia-se no desenvolvimento do *Data Warehouse*. Este ambiente constitui fontes de informações estratégicas do negócio, gerando um diferencial competitivo para a companhia. Diante deste contexto, se fez necessário a implementação do repositório de dados, o *Data Warehouse*, para armazenar, integrar e realizar as consultas multidimensionais sobre os dados extraídos da companhia de abastecimento de água. Portanto, esta dissertação de mestrado tem como objetivos projetar um *Data Warehouse* Departamental referente ao setor comercial, também conhecido como *Data Mart*; aplicar as tecnologias OLAP sobre os cubos de dados multidimensionais; e executar algoritmos de Mineração de Dados visando a geração de um sistema de apoio à decisão para minimização das perdas aparentes no sistema de abastecimento urbano de água.

Palavras chave: *Data Warehouse*, OLAP, *Data Mining*, Sistemas de Abastecimento de Água e Perdas Aparentes.

ABSTRACT

This work propose to use technologies of databases with the aim of providing decision support for managers of sector of sanitation, given that the services of water supply for use of the population are a key indicator of quality of life. The fundamental idea is to collect operational data, reduce them to the scope of the problem, organize them into a repository of data, and finally apply the techniques OLAP and Data Mining algorithms to obtain results that give managers a better understanding of the behavior and profile of the company. To facilitate the application of the techniques of Data Mining is necessary that the data are stored properly. Accordingly, an alternative for increasing the efficiency in storage, management and operation of data to support the decision based on the development of Data Warehouse. This is source of strategic information of the business, creating a competitive differential for the company. In this context, was required to implement the repository of data, Data Warehouse, to store, integrate and carry out consultations on the multidimensional data from the company of water supply. Therefore, this Master's thesis aims to design a Data Warehouse relating to Departmental Business, also known as *Data Mart*; applied the technology on the OLAP multidimensional cubes of data, and run the Data Mining algorithms to the generation of a decision support system to minimize the apparent losses in the urban water supply system.

Keywords: Data Warehouse, OLAP, Data Mining, Water Supply Systems and Apparent Losses.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BI	*Business Intelligence*
CAGEPA	Companhia de Água e Esgotos da Paraíba
DW	*Data Warehouse*
EIS	*Executive Information Systems*
ETL	*Extraction, Transformation and Load*
ID-3	*Iterative Dichotomiser*
JDBC	*Java Database Connectivety*
KDD	*Knowledge Discovery in Databases*
OLAM	*On-Line Analytical Mining*
OLAP	*On-Line Analytical Processing*
OLTP	*On-Line Transaction Processing*
PNCDA	Programa Nacional de Combate ao Desperdício de Água
ROLAP	*Relational On-Line Analytical Processing*
SAD	Sistemas de Apoio à Decisão
SGBD	Sistema Gerenciador de Banco de Dados
SNIS	Sistema Nacional de Informações sobre Saneamento
SQL	*Structured Query Language*
WEKA	*Waikato Environment for Knowledge Analysis*
XML	*Extensible Markup Language*

CAPÍTULO 1

Este capítulo introdutório descreve as principais motivações para realização do trabalho, apresenta os objetivos e a justificativa da pesquisa e, finaliza, expondo a estrutura e organização da dissertação.

1 INTRODUÇÃO

Os sistemas informatizados coletam e armazenam enormes quantidades de dados em seus bancos de dados, aumentando o número de corporações que buscam alternativas para um planejamento, controle e gestão mais eficiente das informações armazenadas, com o melhoramento dos processos de apoio à tomada de decisão e sistemas inteligentes, baseados em descobertas de conhecimento.

Nos dias atuais, com a necessidade de desenvolver sistemas para dar suporte a decisões gerenciais, vem sendo utilizado e aperfeiçoado o *Data Warehouse (DW)*. O DW é um ambiente cuja finalidade é extrair, integrar, limpar e dar consistência aos dados provenientes dos sistemas transacionais da companhia. Além disso, o DW dimensiona e consolida esses dados, organizando-os e melhorando a performance das consultas.

Os primeiros sistemas de suporte à decisão ficaram conhecidos como *Executive Information Systems* (EIS), e tornaram-se muito populares devido à rapidez com que geravam as informações. Contudo, a falta de flexibilidade para realizar consultas *ad hoc* e a necessidade de definição de fórmulas e formatação de novos relatórios por parte do usuário, fizeram com que os EIS ficassem restritos à geração de relatórios corporativos pré-estabelecidos. Visando suprir as necessidades acima citadas surgiram as ferramentas OLAP (*On-Line Analytical Processing*). Elas tornaram viável a construção de um ambiente no qual os analistas de negócio pudessem facilmente navegar pelos dados da companhia, realizando consultas *ad hoc*, fazendo novos cruzamentos entre as dimensões de análise.

Diante deste ambiente empresarial cada vez mais competitivo, a tecnologia da informação, quando bem utilizada, torna-se um importante diferencial entre as empresas que buscam excelência na qualidade do serviço prestado. Neste cenário, surgem as técnicas e aplicações de Mineração de Dados com intuito de descoberta de padrões de comportamento e

de novos conhecimentos sobre os dados armazenados. Portanto, a gestão aleatória baseada na intuição dá lugar a inteligência de negócio.

O presente trabalho expõe uma experiência do Processo de Descoberta do Conhecimento em Banco de Dados, também conhecido com *Knowledge Discovery in Databases* (KDD), a fim de observar a viabilidade e aplicabilidade de um caso real de apoio à decisão. O estudo segue sob a forma da pesquisa bibliográfica, da criação e implementação do *Data Warehouse* Departamental, do uso de tecnologias de análise e recuperação de dados úteis ao processo decisório, conhecidas como OLAP, e da aplicação de técnicas e algoritmos de *Data Mining* para descoberta de novos conhecimento e padrões nos dados.

1.1 OBJETIVOS

Os serviços de abastecimento de água para uso da população continuam sendo um dos indicadores da qualidade de vida da população, sendo de fundamental importância à saúde e à alimentação. Estudos recentes comprovam que a água está se tornando mais escassa, e que menos de 1% (um por cento) da água no mundo está diretamente acessível ao homem. Cerca de vinte países, a maioria deles na África e no Oriente Médio, sofrem de escassez crônica de água, causando danos severos à produção de alimentos e atraso no desenvolvimento econômico (JAMES, et al., 2002).

O estudo proposto por esta dissertação pretende provocar o interesse em pesquisadores envolvidos com a produção, implantação, manutenção, gerência e utilização de Sistemas de Informações Gerenciais ou de Apoio à Decisão. Assim, o resultado desse trabalho terá sua validade para todos aqueles profissionais envolvidos, de alguma forma, em projetos de *Data Warehouse e Data Mining*.

Os objetivos gerais do trabalho são:

- Projetar e desenvolver um Sistema de Apoio à Decisão (SAD);
- Aplicar as tecnologias de Banco de Dados voltadas para projetos de suporte a decisão (modelagem multidimensional);
- Organizar os dados do setor do sistema de abastecimento de água em um *Data Warehouse*, para que eles possam ser analisados por tecnologias OLAP;
- Encontrar padrões e conhecimentos nos dados do setor analisado através dos algoritmos de *Data Mining*.

De acordo com as peculiaridades do setor, os objetivos específicos são:

- Determinar o perfil do setor e do consumidor, por meio da verificação dos consumos de água, valores faturados (conta de água) e pontos de utilização de água;
- Verificar e diagnosticar a situação dos medidores (hidrômetros) presentes nos imóveis;
- Encontrar respostas para as anormalidades e irregularidades praticadas pelos consumidores da qual a empresa de abastecimento de água desconhece;
- Avaliar as inadimplências dispostas no setor selecionado para o estudo de caso.

Este trabalho visa contribuir para o uso racional e eficiente dos recursos hídricos, para isso são aplicadas tecnologias de Banco de Dados como *Data Warehouse*, OLAP e *Data Mining*. Tais tecnologias se propõem em fornecer à entidade gestora de um sistema de abastecimento de água um controle maior do comportamento dos consumidores e imóveis, proporcionando tomadas de decisões eficientes que buscam a redução de perdas de água e das perdas econômicas da companhia de saneamento.

Neste trabalho há a necessidade de conhecimentos envolvendo os dados históricos, tais como o tempo em que o cliente se encontra inadimplente junto à operadora de abastecimento de água; dados históricos das contas e consumos de água e esgoto, histórico do hidrômetro (dados relativos à troca do hidrômetro), etc. Os algoritmos de *Data Mining* com dados que variam com o tempo (séries temporais) são utilizados neste trabalho para prever novos conhecimentos a partir dos dados históricos da série. Tais algoritmos analisam a quantidade de dados existentes e fornecem uma previsão do que pode acontecer nos próximos períodos, levando em consideração os dados passados da base temporal.

As tecnologias de *Data Warehouse* serão utilizadas como parte do processo de descoberta de conhecimento na base de dados do setor de saneamento da cidade de João Pessoa-PB. O ambiente de *Data Warehouse* organizará e disponibilizará os dados, visando facilitar os comandos e execuções OLAP e as consultas para o processo de *Data Mining*.

O termo *Data Warehouse Departamental* é sinônimo de *Data Mart*. Já o termo *Data Warehouse Corporativo* é distinto de ambos. Desta forma, ao longo da dissertação serão encontrados os termos *Data Warehouse*, *Data Warehouse Departamental ou Data Mart*, ambos indicando o mesmo conceito, ou seja, um armazém de dados para o setor de saneamento urbano da cidade de João Pessoa - Paraíba.

O uso das tecnologias OLAP proporcionará as agregações e sumarizações dos dados contidos no *Data Warehouse*, gerando informações úteis ao processo decisório e oferecendo uma análise mais detalhada do setor. A ferramenta OLAP utilizada neste trabalho foi *Pentaho Analysis View*, que por sua vez utiliza a ferramenta *Pentaho Schema Workbench*, ambas serão apresentadas no capítulo 3.

A aplicação do *Data Mining* visa encontrar os consumidores em potencial que apresentam algumas ou todas as características daqueles que já cometeram algum tipo de fraude e/ou inadimplência na rede de distribuição de água, assim como detectar erros e anormalidades na medição do consumo de água por meio dos hidrômetros. Ao constatar tais irregularidades e anormalidades nos consumos e faturas, ações poderão ser tomadas por parte da companhia para eliminá-las, reduzindo o alto índice de perdas de água e consequentemente o alto percentual de perdas de faturamento.

Os resultados obtidos com o *Data Mining* serão utilizados a fim de detectar padrões, descobrir regras significativas e estabelecer relações entre os índices de inadimplências e anormalidades das ligações de água e esgoto dos consumidores, na tentativa de reduzir os índices de perdas aparentes na distribuição de água.

Os dados serão extraídos do *Data Warehouse* Departamental para em seguida alguns algoritmos de *Data Mining* serão aplicados sobre esses dados pelo software *Pentaho* WEKA. Os resultados serão analisados com o propósito de obter medidas corretivas e preventivas para minimizar o problema das perdas aparentes nos sistemas de abastecimento de água. Serão utilizados e comparados entre si três algoritmos de mineração de dados do Aprendizado Indutivo Supervisionado. Quanto ao Aprendizado Indutivo Não-Supervisionado será aplicado um algoritmo que servirá como complemento no processo de descoberta do conhecimento dos dados contidos no *Data Warehouse* (Os tipos de Aprendizado Indutivo serão explanados na seção 2.5.2).

1.2 MOTIVAÇÃO DA PESQUISA

As companhias de saneamento no Brasil perdem em média 44,18% da água que corre no seu sistema de abastecimento, de acordo com o Programa Nacional de Combate ao Desperdício de Água (PNCDA), (MARCKA, et al., Revisão 2004). Boa parte desta água se perde antes mesmo de chegar aos imóveis e atender a população, isto é, a água que se perde entre as estações de tratamento (ETA) e a rede de distribuição do consumidor final.

Segundo o Ministério das Cidades, além dos impactos negativos que as perdas hídricas provocam nos custos operacionais, ampliando a necessidade de investimento em novas instalações de produção e tratamento, elas também causam danos à natureza, pelo aumento da demanda, e geram prejuízos à distribuição regional, principalmente para áreas do Nordeste, onde há escassez de recursos hídricos, e também do Sudeste, cuja região concentra a maior parte da população.

O problema das perdas aparentes em sistemas de abastecimento de água é um assunto que está sempre em foco, visto que o uso correto e consciente da água pela população e pela companhia é significante para o desenvolvimento da humanidade. A detecção das perdas aparentes tem sido de grande interesse para diversas companhias de abastecimento de água, uma vez que representam um fator negativo, tanto financeiro quanto ambiental. Foi desta forma que surgiu o interesse de aprofundar nesta área e desenvolver este trabalho de mestrado.

Portanto, a motivação da presente dissertação surge do interesse de investigar mais detalhadamente se as perdas aparentes de água estão distribuídas proporcionalmente pela cidade ou se estão concentradas em áreas específicas, como por exemplo, nos setores onde o poder aquisitivo dos consumidores é baixo. Para o estudo de caso, serão utilizados dados de um setor do saneamento da cidade de João Pessoa - Estado da Paraíba.

A Companhia de Abastecimento de Água da Paraíba (CAGEPA) disponibilizou o setor 64, na cidade de João Pessoa-PB, para o estudo de caso da presente pesquisa. Este setor corresponde ao sistema de abastecimento urbano de água do bairro e comunidade de Miramar e suas proximidades. Ele apresenta realidades sociais distintas, contemplando população de classe alta, média e a população de baixa renda (habitações populares), além de dispor de diversos tipos de estabelecimentos (comercial, público, industrial, residencial, etc.). Este setor possui aproximadamente 17.800 pontos de utilização e 1.300 consumidores.

A solução desenvolvida nesta dissertação poderá ser aplicada para os demais setores da cidade, trazendo como resultado futuro, uma visão geral dos consumidores de todo o setor de saneamento de João Pessoa. A idéia fundamental desta pesquisa de mestrado é traçar e analisar o perfil dos consumidores e dos imóveis quanto à medição e às perdas aparentes em um determinado período de referência contínuo.

1.3 JUSTIFICATIVA DO TRABALHO

As perdas de água em sistema de abastecimento de água correspondem ao volume de água retirado dos mananciais, e que se encontra na Estação de Tratamento de Água (ETA), subtraído dos volumes de água medidos nos hidrômetros. As ações que visam o controle e a redução de perdas de água delineiam-se na melhoria da qualidade da operação e gestão dos sistemas de abastecimento de água e, consequentemente, inserem-se no contexto do uso racional da água.

1.3.1 Perdas em Sistemas de Abastecimento de Água

Segundo (MARQUES, et al., 2006), o volume de água computado pela companhia de abastecimento de água que não foi faturado corresponde ao índice de perda do sistema. Estas perdas podem ser geradas por vazamentos nas tubulações da rede de distribuição, erros de medição, fraudes nos hidrômetros, erros cadastrais, inadimplências ligações clandestinas de água etc. As perdas são de dois tipos: Reais e Aparentes.

1.3.1.1 Perdas Reais

Segundo (GOMES, et al., 2007), as perdas físicas de água, também chamadas de Perdas Reais, ocorrem em todo o sistema de abastecimento, desde o ponto de captação até os de consumo, passando pela estação de tratamento, de bombeamento, reservatórios, rede de distribuição e ligações prediais. Elas representam a água que efetivamente não chega ao consumidor, em decorrência de vazamentos nas redes de distribuição e seus ramais provocados por deficiência nos equipamentos, envelhecimento das tubulações e conexões, e operação e manutenção inadequada em todo o sistema.

1.3.1.2 Perdas Aparentes

De acordo com a *International Water Association* (IWA), as Perdas Aparentes, também chamadas de Perdas Não Físicas ou Comerciais, referem-se a toda água que não é medida ou que não tenha o seu uso definido. Ocorre com a água que é tratada e fornecida pela companhia, e consumida pelos clientes, porém não é corretamente medida e, portanto não é faturada, nem gera arrecadação correspondente. Estão relacionadas às ligações clandestinas e/ou irregulares, fraudes nos hidrômetros, erros de micro e macromedição, política tarifária, erro cadastral (desatualização do cadastro, inatividade em ligação ativa, ligação não cadastrada por descuido), erro de leitura, etc.

Para (JAMES, et al., 2002), algumas das causas para as Perdas Aparentes são os erros e desatualizações no cadastro de clientes; Fraudes, violação ou danificação de medição nos hidrômetros[1]; e Ligações Clandestinas ou Ligações não Cadastradas.

Segundo estima (QUEYROI, 2007), metade dos problemas no segmento de saneamento estão ligados a vazamento, ou seja, perdas físicas, e a outra metade são decorrentes de falhas na medição, ou seja, perdas aparentes.

De acordo com (SNIS, 2007), as regiões Norte e Nordeste são as áreas onde há maior perda de faturamento e são também onde predominam as menores rendas per capta no país. Isto aponta para dois aspectos possíveis de situações de perdas: um relacionado ao baixo poder de consumo destas populações, altos índices de inadimplência e consequentemente lucros menores e outro relacionado às grandes potencialidades de irregularidades nas redes, com perdas de volumes de água tratada em função das ligações clandestinas.

No que se refere aos dados do (SNIS, 2007), o valor médio das perdas de faturamento para todo o conjunto de prestadores de serviços foi de 39,8%. Ressalta-se, segundo o relatório, que os prestadores com maiores perdas concentraram-se nas regiões Norte (53,4%) seguida do Nordeste (45,1%). A região Sudeste possui índices de perdas em torno de 39,8%, Centro-Oeste de 39,2% e Sul de 26,6%.

A Companhia de Água e Esgotos da Paraíba (CAGEPA), utilizada no estudo de caso, obteve um intervalo de perdas de faturamento entre 40,1 e 50,0 %. Este alto índice reflete-se de forma negativa para o Estado, visto que as perdas de faturamento estão diretamente ligadas às perdas reais e aparentes. Estas, por sua vez, acarretam problemas estruturais, ambientais e sociais para toda a população.

É importante reduzir as perdas aparentes para elevar a eficiência do sistema de abastecimento de água. Na tentativa de minimizar e evitar tais desperdícios, este trabalho empenha-se em investigar e detectar perdas aparentes, e para alcançar este objetivo, utilizou-se o processo de descoberta do conhecimento em base de dados, com ênfase no *Data Mining*.

[1] Por exemplo: rompimento do lacre e inversão do hidrômetro; execução de *by pass (i.e.*, desvio feito no aparelho, evitando que ele meça corretamente o volume consumido); colocação de arame para travar a turbina do hidrômetro etc.

A análise de grande volume de dados permitirá que se observem tendências, que se detectem regiões onde as perdas aparentes e inadimplências dos consumidores são mais freqüentes; quais são categorias de consumo mais suscetíveis às perdas, entre outras ações.

1.4 ESTRUTURA DA DISSERTAÇÃO

A presente dissertação está organizada em 7 capítulos, incluindo este introdutório. O Capítulo 2 configura o estado da arte da pesquisa e tem como objetivo apresentar os principais conceitos envolvidos com o tema da dissertação, sob forma de uma revisão bibliográfica.

O capítulo 3 apresenta e caracteriza a companhia de abastecimento de água envolvida no estudo de caso; e relaciona a teoria exposta no capítulo 2 sob a forma de um estudo de caso real. Nele serão discutidas as tecnologias de banco de dados aplicadas ao setor de saneamento, além de descrever os mecanismo de criação e implementação do *Data Warehouse*; a utilização das tecnologias OLAP e de *Data Mining*, apresentando suas principais funções, vantagens e aplicabilidade.

O capítulo 4 apresenta os resultados e discussões do estudo de caso, apresentado as comparações dos algoritmos de *Data Mining* quanto ao seu tipo de aprendizado indutivo.

O capítulo 5 retoma as discussões gerais do trabalho de forma conclusiva, finalizando a dissertação com os resultados e contribuições relevantes, dificuldades encontradas e as indicações para trabalhos futuros. O último capítulo expõe as referências bibliográficas consultadas.

CAPÍTULO 2

Este capítulo configura o estado da arte da dissertação e empenha-se em discutir os assuntos e requisitos relacionados aos Sistemas de Apoio à Decisão, Data Warehouse, OLAP e Data Mining. São apresentados os principais conceitos, o histórico e importância de cada um no processo decisório, mostrando sua relevância para o atual mercado competitivo e tecnológico do Business Intelligence.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SISTEMA DE APOIO À DECISÃO

Os Sistemas de Apoio à Decisão (SAD), ou *Decision Support Systems* (DSS), visam proporcionar uma avaliação crítica das informações dos negócios, auxiliando a gerência a definir tendências, apontar problemas e absorver decisões inteligentes.

De acordo com (DATE, 2004), o processo de tomada de decisão com auxílio de computadores iniciou na década de 70, onde os processos começaram a ser informatizados e as informações passaram a ser pré-definidas e selecionadas por meio dos *Executive Information Systems* (EIS). Na fase atual, os processos de tomada de decisão são totalmente informatizados e o gestor define os atributos mais importantes ao processo decisório, recebendo subsídios e informações processadas pelos Sistemas de Apoio à Decisão, através de ferramentas OLAP, que será discutida na seção 2.4.

Nas décadas anteriores, o foco estava voltado ao crescente aumento da quantidade de informação armazenada em formato eletrônico. Segundo (ZARUR, 2005), estima-se que a quantidade de dados duplica a cada um ano e meio e que o tamanho e número de bases de dados crescem a um ritmo ainda mais elevado. Este grande aumento deve-se essencialmente à constante diminuição do custo de armazenamento dos dados e ao efetivo aumento da eficiência dos computadores em manuseá-los.

De acordo com (ELMASRI, et al., 2005), os Bancos de Dados de apoio à decisão costumam ser extensos, fortemente indexados e envolver uma grande quantidade de

redundância, em especial, sob a forma de replicação e de tabelas de totalização. As chaves costumam envolver um componente temporal e as consultas costumam ser complexas.

Certos aspectos dos sistemas de BD para apoio à decisão os distinguem dos sistemas de BD tradicionais, sendo o principal deles o fato dos BD para apoio à decisão serem quase que exclusivamente para leitura/consultas, e dificilmente para atualizações. Como consequência, observa-se as dificuldades em se trabalhar na prática com um grande número de variáveis, que são os atributos do BD, e a grande quantidade de dados históricos. Em virtude desta complexidade, opta-se por extrair apenas as informações mais relevantes da base de dados transacional.

O bom processamento de extração dos dados é a principal razão para o sucesso na tomada de decisão. Esta extração corresponde à cópia dos dados desejáveis do ambiente operacional para o processamento subsequente. Significa que os usuários podem operar sobre os dados extraídos da maneira como desejarem, sem interferência no ambiente operacional.

Após tantos anos de concentração na obtenção de dados, o problema, agora, passa a ser o aproveitamento deste precioso recurso. Reconheceu-se que estes dados propiciam aos indivíduos responsáveis pelas decisões, o planejamento das ações, a definição de estratégias e a eficácia em suas decisões.

O apoio à decisão se utiliza de várias tecnologias, dentre elas, *Data Warehouse*, *Data Mart*, Sistema Gerenciadores de Banco de Dados, Processamento Analítico On-line (OLAP), Banco de Dados Multidimensionais, Mineração de Dados (*Data Mining*) etc.

As Ferramentas de Apoio à Decisão (FAD) fazem parte do conceito de *Business Intelligence* (BI), ou Inteligência de Negócios, e correspondem ao conjunto de tecnologias que permitem o cruzamento de informações e suporte a análise dos indicadores de desempenho de um negócio (COLAÇO, 2004).

Estas ferramentas são softwares desenvolvidos com objetivo de apresentar graficamente (e não apenas numericamente) as informações do negócio, auxiliando a simulação de ocorrências, fornecendo maior capacidade de análise para o descobrimento de novos conhecimentos e padrões.

2.1.1 Descoberta de Conhecimento em Banco de Dados

O processo de descoberta de conhecimento em banco de dados se propõe em encontrar e interpretar padrões através das análises nas fontes de dados. O objetivo é extrair de grandes bases de dados, sem nenhuma formulação prévia de hipóteses, as informações desconhecidas, válidas e acionáveis, que poderão ser úteis para a tomada de decisão.

Ficou mais conhecido pelo acrônimo KDD, que em inglês significa *Knowledge Discovery in Database*. O processo de KDD foi proposto para determinar as etapas que produzem conhecimentos a partir dos dados e, principalmente, definir a etapa de *Data Mining* (Mineração de Dados), que é a fase que transforma dados em conhecimento (FAYYAD, et al., 1996).

Como ilustra a Figura 2.1, cada fase da execução do processo KDD possui uma interseção com as demais. Deste modo, os resultados produzidos em uma fase podem ser utilizados para melhorar os resultados das próximas fases. Este cenário revela um processo iterativo, que busca sempre aprimorar os resultados a cada iteração.

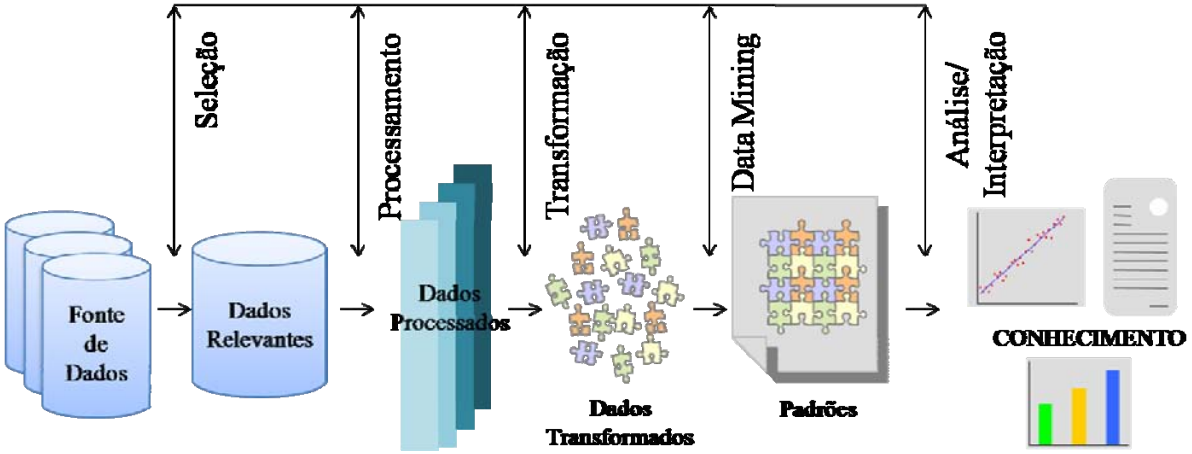


Figura 2.1 - etapas do processo de KDD

Fonte: (Adaptação) (SYMEONIDIS, et al., 2005 p. 14)

O processo de KDD envolve três etapas iniciais: seleção, (pré) processamento e transformação, as quais compõem a preparação dos dados. Em seguida vem a fase de *Data Mining*, considerada essencial ao processo e foco principal deste trabalho. Por fim, o conhecimento gerado é analisado e assimilado, por meio da etapa de análise e interpretação dos resultados, que se encontra no topo do processo.

2.2 DATA WAREHOUSE

Os *Data Warehouses* podem ser traduzidos como Armazéns de Dados e são tipos especiais de banco de dados que se tornaram conhecidos e bastante utilizados a partir da década de 90. Será utilizado o termo em inglês neste trabalho, visto que a maioria dos autores utiliza-o por considerarem mais intuitivo. De acordo com (INMON, 2005), o termo é definido como “um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões da gerência”. Onde não volátil significa que, uma vez inseridos, os dados não podem ser alterados, embora possam ser excluídos. O conceito de armazém de dados surgiu por duas razões: primeiro, pela necessidade de fornecer uma origem de dados única, limpa e consistente para fins de apoio à decisão; segundo, pela necessidade de fazê-lo sem causar impacto sobre os sistemas operacionais.

O processo de desenvolver e gerenciar repositórios de dados a partir de várias fontes com o propósito de obter uma visão detalhada e singular de parte ou todo um negócio, é conhecido como *Data Warehousing*. De acordo com (GARDNER, 1998), a concretização do *Data Warehousing* é considerada um dos primeiros passos para tornar factível a análise de grande quantidade de dados no apoio ao processo decisório.

Segundo (PONNIAH, 2001), o *Data Warehousing* não é um software ou produto de hardware que se adquire para fornecer informações estratégicas. É, sim, um ambiente computacional onde os usuários são colocados diretamente em contato com os dados que necessitam para tomar as melhores decisões.

O produto principal obtido de um projeto de *Data Warehousing* é o seu *Data Warehouse* (DW), e cujo objetivo básico é gerar um repositório que contenha dados limpos, agregados e consolidados, podendo este ser analisado por ferramentas do tipo OLAP (*On-Line Analytical Processing*) e *Data Mining* (assuntos abordados nas seções 2.4 e 2.5, respectivamente).

As bases de dados convencionais (relacionais) possuem algumas características, tais como dinamismo, redundâncias, incompletude e ruídos, tornando-as confusas e não viáveis à extração de informações delas próprias. O *Data Warehouse* surgiu com o objetivo de fornecer os subsídios necessários para a transformação de uma base de dados que utiliza *On-Line Transaction Processing* (OLTP) para *On-Line Analytical Processing* (OLAP). A primeira significa os processamentos que executam as operações do dia-a-dia da organização e a

última, os processamentos que suportam a tomada de decisões. Os termos OLTP e OLAP serão detalhados na seção 2.4.

Alguns problemas são apontados por (KIMBALL, et al., 2002; IMHOFF, et al., 2003) quanto ao uso do modelo relacional pra a realização de consultas complexas. A manipulação dos dados, incluindo as consultas, é muito mais rápida e intuitiva no modelo multidimensional em comparação ao modelo relacional.

Enquanto uma busca no modelo relacional exige a navegação entre diversas tabelas, no modelo multidimensional isto não é necessário, o que o torna mais eficiente e com melhor desempenho. Devido ao grande número de tabelas normalizadas do modelo relacional, torna-se inviável a realização das consultas, já que é preciso fazer um grande número de conexões (*inner join*) entre as mesmas.

Os benefícios da modelagem multidimensional é que ela torna os esquemas de dados mais compreensíveis para os usuários finais, e por outro lado, ela permite usar armazenamento específico e técnicas de acesso que melhoram o desempenho de *queries*. A maneira para obter estes benefícios é a simplificação dos esquemas de dados, de forma que eles só contenham as coisas essenciais (i.e. um fato para ser analisado e suas dimensões de análise).

Constantemente há atualização na base de dados e consequentemente as informações históricas são perdidas. Na projeção de bases de dados para *Data Warehouse*s, deve-se quebrar o paradigma dos modelos de dados normalizados utilizados nos BD tradicionais, e buscar armazenamento histórico/temporal. Ao desnormalizar as tabelas, o projetista do DW busca ganhar desempenho nas consultas, contudo, não se deve introduzir redundância em qualquer lugar do modelo.

A idéia dos *Data Warehouses* geralmente se destina a fornecer uma única origem aos dados para todas as atividades de apoio à decisão. O propósito de construir uma espécie de warehouse limitado e de uso especial, adaptado à finalidade imediata, é uma solução aos problemas encontrados com os *Data Warehouses* coorporativos, visto que desta forma é possível o acesso mais rápido aos dados, ao contrário se eles tivessem que ser sincronizados com todos os outros dados a serem carregados no warehouse completo. Essas considerações levaram ao conceito de *Data Marts*, que será apresentado no próximo Item.

Existem três tipos principais de processamentos usados com o *Data Warehouses* (HAN, et al., 2006):

- Processamento de Informação: suporta consultas, análises estatísticas e relatórios;
- Processamento Analítico: ferramentas OLAP e suas operações;
- Processamento de Mineração de Dados: descoberta de conhecimento automatizada, encontrando padrões escondidos nos dados. Pode-se realizar visualizações dos dados, assim como classificações e predições através das técnicas de *Data Mining*.

2.2.1 Data Mart

De acordo com (KIMBALL, et al., 2002 p. 36):

> *"Um Data Mart é um Data Warehouse de menor capacidade e complexidade usado para atender a uma unidade específica de negócios. Portanto, são tipicamente mais fáceis de construir e manter."*

Um *Data Mart*, segundo (INMON, 2005) é uma coleção de assuntos organizados para dar suporte à tomada de decisão e estão baseados nas necessidades de um determinado departamento. É geralmente descrito como um subconjunto dos dados extraído para um ambiente separado. Eles são úteis nas seguintes condições:

- Os dados devem estar segregados para melhorar o desempenho do sistema do ponto de vista do usuário.
- Deve existir uma cópia dos dados onde apenas pessoas com autorização podem ter o privilégio de acessá-las.
- Em um ambiente corporativo, é importante fortalecer o conceito de propriedade dentro do banco de dados. Diferentes setores (Financeiro, Marketing, Vendas, etc.) serão responsáveis por diferentes *Data Mart*s.

Um *Data Mart* representa uma área específica a partir de um único processo empresarial, sendo considerado a parte de um todo. É por isso que o *Data Mart*, que é uma abordagem descentralizada do conceito de *Data Warehouse*, não é um "pequeno *Data Warehouse*", mas sim uma unidade lógica de um DW, podendo ser qualificado como um *Data Warehouse* Departamental. A Tabela 2.1 relaciona algumas diferenças entre o ambiente de *Data Mart* e o ambiente de *Data Warehouse*.

Tabela 2.1 - diferenças entre *data mart* e *data warehouse*

Data Mart	*Data Warehouse*
Departamental (única área);	Corporativo (múltiplas áreas);
Nível tático;	Nível estratégico;
Otimizado para acesso e análise;	Otimizado para armazenamento e gerenciamento de grandes volumes de dados;
Poucas fontes de dados;	Muitas fontes de dados;
Pequenos estágios de implementação (menor tempo)	Múltiplos estágios de implementação (maior tempo);

Fonte: (INMON, 2005)

Observa-se que as principais diferenças entre *Data Mart* e *Data Warehouse* estão relacionadas ao tamanho e o escopo do problema a ser resolvido. Enquanto um *Data Mart* trata de problema departamental ou local, um *Data Warehouse* envolve o esforço de toda a companhia para que o suporte à decisões atue em todos os níveis da organização. Desta forma, o desenvolvimento de um *Data Warehouse* requer tempo, dados e investimentos gerenciais muito maiores que um *Data Mart*.

De acordo com (INMON, 2005), um dos assuntos em pauta para a área de TI nos últimos anos é decidir qual ambiente de apoio à decisão desenvolver primeiro, o *Data Warehouse* ou os *Data Marts*. A escolha entre um único *Data Warehouse* Corporativo e uma arquitetura consistindo de muitos *Data Mart*s é um ponto de algumas controvérsias entre os pesquisadores. Uma boa parte dos especialistas defende a implementação de *Data Mart*s como passo inicial e existe uma unanimidade de especialistas alertando ao usuário que em momento algum ele pode esquecer o modelo corporativo, sob o risco de obter sérios prejuízos.

Após o levantamento e definição do conjunto de atributos e dados necessários para realização desta pesquisa, optou-se por implementar um *Data Warehouse* Departamental, ou seja, um *Data Mart* do departamento comercial A escolha se deu em virtude dos dados adquiridos corresponderem às informações comerciais dos consumidores e imóveis de um setor da companhia de abastecimento de água. Os resultados obtidos com aplicação das ferramentas OLAP e *Data Mining* sobre o *Data Warehouse* Comercial visam à criação de um novo ambiente computacional com o propósito de fornecer informação estratégica para a companhia de saneamento.

A presença de vários *Data Marts* em uma mesma companhia oferece alto risco de redundância dos dados. Esses ambientes de armazenamento e análises de dados fisicamente distintos trazem benefícios e facilidades, entretanto, existe um preço a se pagar. Desta forma, ao construir *Data Marts* deve-se sempre ter a preocupação de compartilhamento de dados, tabelas e relatórios em comum entre os demais departamentos, consequentemente entre os demais *Data Marts*. Afinal, relatórios em comum não podem possuir valores diferentes entre os departamento.

A separação física dos dados em diferentes grupos, pela presença de vários *Data Marts* em uma única companhia, diminui a habilidade de organização das informações. A dificuldade em evitar a inconsistência dos dados pode ir contra o paradigma de um *Data Warehouse*. Afinal, uma das principais motivações para o surgimento do DW foi eliminar as inconsistências dos dados e agrupá-los em um único ambiente de apoio à decisão.

2.2.2 Propriedades do Data Warehouse

De acordo com (INMON, 2005), o DW deve seguir quatro propriedades fundamentais, são elas: Orientado por Temas, Integrado, Variante no Tempo e Não Volátil.

A propriedade “Orientado por Tema”, (INMON, 2005) refere-se à importância de organizar as informações pelos temas principais. Para o setor de saneamento, que caracteriza o estudo de caso deste trabalho, os principais temas são: perfil dos consumidores e imóveis, serviço prestado e perdas aparentes.

Cada tema pode envolver várias tabelas e atributos e podem existir dados acumulativos e detalhados. Para o tema perfil dos consumidores, por exemplo, os atributos podem ser os dados cadastrais (nome, endereço, telefone, e-mail), dados das contas e consumos de água, etc. Como exemplo de dados acumulativos tem-se a consulta que retorna o somatório dos consumos descendentes, agrupados por clientes no período de 2007 a 2008.

A propriedade “Integrado” presente em um DW mostra a necessidade de acoplar dados de diferentes formatos. Os dados precisam seguir uma convenção padrão para que desta forma eles possam fornecer significados únicos. Um sistema do setor comercial pode codificar o “indicativo de medidor” como SIM ou NÃO. Onde SIM se refere ao consumidor que possui hidrômetro para medição do consumo de água e NÃO caracteriza o consumidor que não possui hidrômetro para medição. Outro setor da companhia de abastecimento pode

codificar 0 (Tem Hidrômetro) e 1 (Não tem Hidrômetro), assim como S (Tem Hidrômetro) e N (Não tem Hidrômetro). Desta forma, é necessário definir uma única codificação dos dados extraídos para o *Data Warehouse*.

A terceira propriedade "Variante no Tempo" em um ambiente de *Data Warehouse* determina que os dados não sejam atualizáveis e que eles possam ser comparados ao longo do tempo. Os dados são atribuídos como retratos da base de dados operacional atual, onde cada ocorrência e cada mudança são consideradas como um novo registro, pois a informação histórica não é perdida.

Contudo, em um Ambiente Transacional[2] a atualização dos dados ocorre em virtude das mudanças ocorridas. Os dados retornados em consultas correspondem à informação no momento da consulta, e neste caso as consultas históricas não são consideradas[3].

Supondo que desejamos recuperar a quantidade de pontos de consumo do consumidor. Em 2007 o consumidor possuía 20 pontos de consumo em sua residência, já em 2008 passou para 23 pontos de consumo. A consulta retornará apenas a estado atual dos pontos de consumo, ou seja, 23. A informação histórica anterior é perdida. Entretanto, no DW ao consultar os pontos de acesso do cliente em 2007, do exemplo acima, o resultado corresponderá ao valor 20.

A última propriedade proposta por (INMON, 2005), que é a "não volatilidade" dos dados, se verifica em banco de dados que é disposto fisicamente para otimizações de inclusões e consultas. Ou seja, não deve ser um banco preparado para atualizações.

O DW consiste em fornecer apenas acessibilidade aos dados, não permitindo atualizações ou alterações. Ele concede apenas a carga inicial e consulta (acessos) aos dados. Ao contrário, a volatilidade é uma propriedade bastante observada em ambientes operacionais tradicionais, pois os registros dos dados são atualizados constantemente.

---

[2] Conhecido também por "Ambiente Operacional". O termo mais utilizado nesta dissertação é "Ambiente Transacional".
[3] Neste caso não estão sendo mencionados os ambientes que utilizam Banco de Dados Temporais (BDT), apenas os que utilizam Banco de Dados Relacionais.

2.2.3 Granularidade

A questão da granularidade é um dos mais importantes aspectos no projeto de *Data Warehouse*. Corresponde ao nível no qual os dados estão sumarizados no *Data Warehouse*, ou seja, refere ao nível de detalhamento das informações armazenadas. Quanto mais detalhados os dados, menor a granularidade do DW (granularidade fina ou baixa). Quanto maior o nível de granularidade, menor será os detalhes dos dados (granularidade grossa ou alta).

Segundo (PONNIAH, 2001 p. 23), a granularidade está diretamente ligada ao volume de informações armazenadas e aos tipos de consultas que podem ser realizadas pelo usuário de um DW. Ao definir um nível muito detalhado, o usuário poderá ver a informação em qualquer nível de agregação e maior será o detalhamento das consultas. Contudo, a escolha de um nível baixo demais poderá ocasionar em um aumento do volume de dados armazenado e, consequentemente, afetará a performance do sistema. Por outro lado, ao definir um nível pouco detalhado, o usuário ficará impossibilitado de realizar consultas mais detalhadas, visto que o volume de informações armazenadas é menor, porém, permite maior desempenho e rapidez nas respostas das consultas.

Portanto, quanto mais alto o nível de granularidade, menor o volume de dados e o número de índices e, indiretamente, menor o processamento necessário. O problema existente é que o nível de granularidade é também inversamente proporcional ao número de consultas que podem ser atendidas.

A utilização de apenas um nível de granularidade em projetos de *Data Warehouse* não é recomendada como solução eficiente. Afinal, o nível de granularidade é inversamente proporcional à quantidade de consultas atendidas e/ou desempenho do processamento. O modelo dimensional (ver item 2.3) é o mais utilizado nas aplicações de DW, e este utiliza técnicas de níveis duais de granularidade.

O desenvolvimento de um ambiente com níveis duais de granularidade consiste em ter dados de um mesmo assunto em granularidades diferentes. A opção pelo uso de níveis duais tem como finalidade baixos tempos de resposta nas consultas de granularidade alta e análise dos dados em maior detalhe nas consultas com níveis de granularidade baixa.

A razão pela qual a granularidade é a principal questão de projetos de *Data Warehouses* consiste no fato de que ela afeta profundamente o volume de dados, ao mesmo

tempo afeta no tipo de consulta que pode ser atendida. O volume de dados residentes no DW deve ser balanceado de acordo com o nível de detalhe de uma consulta.

2.2.4 Arquitetura do Data Warehouse

Em um ambiente projetado de *Data Warehouse* há duas espécies de dados: Dados Primitivos (operacionais ou atômicos) e Dados Derivados (de apoio à decisão ou sumarizados). Os dados primitivos consistem em valores referentes ao momento presente, e são baseados em aplicações, podem ser atualizados, são detalhados, e processados repetitivamente. Enquanto que os dados derivados são geralmente valores históricos, baseados em assuntos ou negócios, são resumidos, ou refinados, não são atualizados, representam valores de momentos já decorridos ou instantâneos e são processados de forma heurística (INMON, 2005).

A escolha de dados primitivos para o armazenamento em um DW proporciona vários benefícios, porém gera algumas desvantagens. O maior benefício está na possibilidade de se pesquisar em base de dados mais rica, proporcionando uma análise mais aprofundada e cuidadosa nos dados, o que permite a verificação do histórico, de tendências, de previsões e de elaboração de cenários. A principal desvantagem é a necessidade de um espaço muito maior nos dispositivos de armazenamento, assim como uma maior capacidade de processamento para que não haja baixa performance nas consultas e análises dos dados.

A escolha de dados derivados para o armazenamento em DW também traz benefícios e desvantagens. O maior benefício é que os dados já estão sumarizados, ou seja, já estão resumidos e armazenados em um formato no qual são mais consultados. Ocupam menos espaço nos dispositivos de armazenamento e a performance das consultas e das análises dos dados é mais rápida. A desvantagem é que o armazenamento dos dados sumarizados limita bastante a capacidade de pesquisa e de análise. A maioria das empresas opta pelas duas formas de armazenamento simultaneamente. Desta forma, somam-se as vantagens e reduzem-se as desvantagens de ambas.

Segundo (INMON, 2005), com estas diferenças nos dados, tem-se a projeção de quatro níveis do ambiente arquitetural de um DW, são eles: Nível Operacional (ou Transacional), Nível Atômico (ou *Data Warehouse*), Nível Departamental (ou *Data Mart*) e Nível Individual, como mostra a Figura 2.2.

Figura 2.2 - os quatro níveis de dados do ambiente arquitetural de um *data warehouse*

Fonte: Adaptação de (INMON, 2005)

O nível Operacional de dados detém apenas a aplicação orientada a dados primitivos e atende à comunidade de processamento de transações de alta performance. O nível de *Data Warehouse* contém dados primitivos que não são atualizados, além de alguns dados derivados. O nível Departamento contém quase que exclusivamente dados derivados. Este nível é moldado pelas necessidades dos usuários finais adaptadas às necessidades do departamento. E o nível individual de dados é onde muitas das análises heurísticas são realizadas. Segue a Figura 2.3 com exemplos dos quatro níveis de dados.

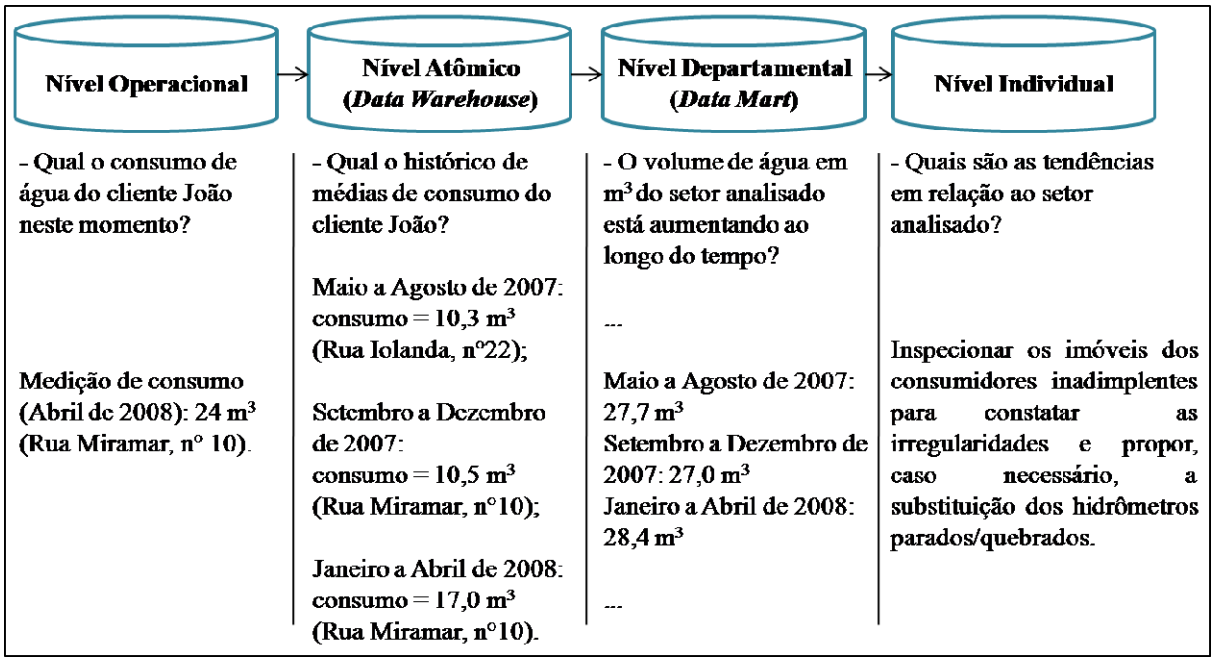


Figura 2.3 - exemplos de consultas referentes aos quatro níveis de dados

O Nível Operacional retornará a média na medição de consumo de água do cliente João (nome e endereço fictício) na última medição efetuada, ou seja, em Abril de 2008 e que corresponde a 24 m^3de água. O registro neste nível contém os valores recentes do cliente, onde para se conhecer a situação atual dele, é acessado o registro existente neste nível. Para alteração dos dados de João, o registro do nível operacional será alterado, com o objetivo de refletir os novos dados atualizados.

O segundo nível, nível de *Data Warehouse*, resulta no histórico de consumo do consumidor João, isto é: 10,3 m^3 entre Maio e Agosto de 2007, 10,5 m^3 entre Setembro e Dezembro de 2007 e média de volume de 17,0 m^3 entre Janeiro e Abril de 2008. Neste nível existem vários registros do João, apresentando o histórico das informações sobre ele. Não há sobreposição nos registros existentes no ambiente de DW. Quando houve mudança de endereço do consumidor (da Rua Iolanda para Rua Miramar), foi gerado um novo registro no DW, refletindo as datas do período que João residiu naquele local.

O terceiro nível, nível de *Data Mart*, permitirá ao executor extrair informações de maior complexidade e específico do negócio, facilitando as tomadas de decisões. Um exemplo seria uma lista com todos os clientes por categoria, sendo o consumidor João incluído nesse resumo de cada quadrimestre. Como consulta do nível 3 tem-se: "O volume de água em m^3 do setor analisado está aumentando ao longo do tempo (relatório quadrimestral)?". O retorno desta consulta são as médias de consumo agrupadas por quadrimestre (Maio a Agosto de 2007; Setembro a Dezembro de 2007 e Janeiro a Abril de 2008).

Por fim tem-se o nível Individual, que possibilita a previsão de informações, fornecendo visões futuras por meio das análises heurísticas. Os dados neste nível são, geralmente, temporários e de pequenas proporções.

No exemplo apresentado na Figura 2.3, ao analisar o setor observou-se que a maioria dos consumidores inadimplentes possui hidrômetros instalados a mais de 10 anos e com capacidade de vazão de até 3 m^3. Ainda no nível Individual, verificou-se que aproximadamente metade dos consumidores está com consumo de água igual a zero, o que representa hidrômetro parado. Estes resultados indicam casos onde uma inspeção técnica poderia ser realizada, afinal os equipamentos de medição podem estar defasados e/ou

danificados, gerando perdas aparentes no sistema. Na seção 3.3.3 é proposto um modelo de Mineração de Dados aplicado à inspeção e troca de hidrômetros.

2.3 MODELAGEM DIMENSIONAL

A modelagem dimensional[4] é uma metodologia que possibilita que os dados sejam modelados visando aperfeiçoar o desempenho de consultas e oferecer facilidades de utilização a partir de um grupo de eventos simples de medição. A visão dimensional facilita o entendimento e visualização de problemas típicos de sistemas de apoio à decisão, é mais intuitiva e eficaz para o processamento analítico e é utilizada pelas tecnologias OLAP (discutidas na seção 2.4).

Três conceitos estão envolvidos com a modelagem dimensional, são eles: *fatos*, *dimensões* e *métricas* (medidas ou atributos). De acordo com (BALLARD, et al., 1998), um *fato* é uma coleção de itens de dados que consiste de métricas e do contexto do negócio. A *dimensão* é uma coleção de itens do mesmo tipo que representa as visões do negócio. A *métrica* é definida como um atributo numérico de um fato, e representa o comportamento do negócio para as dimensões.

Os fatos são reunidos na tabela de fatos. Segundo (KIMBALL, 1997), as tabelas de fatos normalmente contém dados numéricos e somatórios. Como os Data Warehouses geralmente recuperam muitos registros em uma única consulta, é uma tendência agrupar os dados para análise, pois esta compactação proporciona ganhos de performance. Cada dimensão possui uma tabela de dimensão associada que armazena as descrições textuais das dimensões do negócio. Cada tabela de dimensão tem uma chave primária que corresponde exatamente a um dos componentes da chave composta da tabela de fatos.

A Tabela 2.2 a seguir apresenta o modelo dimensional implementado em SGBD Multidimensional e SGBD Relacional. Os dados da tabela correspondem às médias de consumo em m3 das quadras 010, 015, 020 e 025, agrupadas por categoria de consumo durante o período de 2007 a 2008.

[4] Os termos “modelagem dimensional” e “modelagem multidimensional” são utilizados na literatura para expressar o mesmo conceito. Não há uma definição padrão que indique uma diferença precisa entre os dois termos.

Tabela 2.2 - exemplo da modelagem dimensional em SGBDS

Quadra	Categoria		
	Comercial	Industrial	Residencial
Quadra_010	190.0	-	-
Quadra_015	34.3	23.5	114.0
Quadra_020	38.2	-	88.8
Quadra_025	-	-	19.8

Modelagem Dimensional em SGBD Multidimensional

Painel de saída

Saída de Dados | Explain | Mensagens | Histórico

	quadra text	categoria text	media_consu numeric
1	Quadra_010	COMERCIAL	190.0
2	Quadra_015	COMERCIAL	34.3
3	Quadra_015	INDUSTRIAL	23.5
4	Quadra_015	RESIDENCIAL	87.9
5	Quadra_020	COMERCIAL	25.5
6	Quadra_020	RESIDENCIAL	88.8
7	Quadra_025	RESIDENCIAL	19.8

Modelagem Dimensional em SGBD Relacional (PostgreSQL)

A principal vantagem na utilização de SGBDs Multidimensionais é que eles implementam fisicamente o modelo dimensional. Contudo, uma das desvantagens é a esparsidade, ou seja, células que ocupam espaços em disco, mas não contêm dados cadastrados, como é caso das quadras 010, 020 e 025. Outra desvantagem é considerada quando o modelo dimensional possui um grande número de dimensões, pois traz como consequências, problemas de desempenho e tempo maior de processamento das consultas. Os SGBDs Relacionais possuem uma maior aceitação e utilização, entretanto, exigem adaptações, visto que eles não implementam fisicamente o modelo dimensional.

Existem três esquemas utilizados para modelagem dimensional dos dados, são eles: Esquema Estrela (S*tar Schema*), Esquema Floco de Neve (S*nowflake* S*chema*) e Esquema Constelação de Fatos (*Facts Constallation Schema*).

2.3.1 Esquema Estrela

Idealizado e criado por Ralph Kimball, o Esquema Estrela é uma forma de dispor as tabelas do modelo relacional para o modelo dimensional, podendo ser implementado em BD relacionais e principalmente, em BD multidimensional (KIMBALL, et al., 2002).

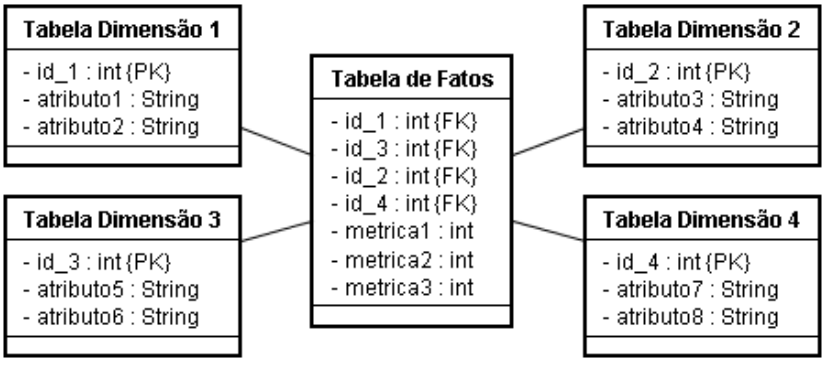


Figura 2.4 - exemplo geral do esquema estrela

Conforme ilustra a Figura 2.4, o Esquema Estrela é uma estrutura com tabelas e ligações bem definidas, baseado no formato de uma estrela. É formado por uma tabela central, denominada *tabela de fatos*, a qual possui os dados principais da visão da análise, ou seja, o assunto que está sendo analisado, por exemplo, o consumo, as quantidades de inadimplentes, as quantidades de consumidores, etc. Nela ficam ligadas as tabelas de dimensão, que possuem os aspectos pelos quais se deseja observar as medidas relativas ao processo que se está analisando.

De acordo com (HAN, et al., 2006), as tabelas dimensionais são desnormalizadas para aumentar o desempenho das consultas. A consulta ocorre inicialmente nas tabelas de dimensão e em seguida na tabela de fatos, assegurando a precisão dos dados através de uma estrutura completa de chaves onde não é preciso percorrer todas as tabelas. Isso garante um acesso mais eficiente e um melhor desempenho.

Ao contrário das tabelas de dimensão, a tabela de fatos armazena grandes quantidades de dados históricos, normalmente numéricos, obtidos a partir da interseção de todas as dimensões do Esquema Estrela. Ela também armazena os indicadores de desempenho (medidas) do negócio. Para cada dimensão há uma chave primária que corresponde a um dos campos, chave estrangeira, da chave da tabela de fatos.

A Tabela 2.3 apresenta um comparativo entre os dois tipos de tabelas do Esquema Estrela, mostrando as diferenças entre elas.

Tabela 2.3 - comparativo entre as tabelas de fatos e dimensão

Tabela de Fatos	Tabela de Dimensão
Grande volume de dados	Volume comparativamente menor
Chave composta	Chave simples
Referencia cada tabela de dimensão	Descrevem os fatos
Histórica	Atributos usados como filtro nas consultas
Agiliza consultas, pois os fatos (variáveis) são usualmente numéricos e tipicamente aditivos	Desnormalizada (redundâncias)

Fonte: (KIMBALL, et al., 2002)

Apesar do Esquema Estrela apresentar desvantagens em termos de espaço de armazenamento devido à redundância dos dados e, principalmente, fazer com que o desempenho diminua nas operações de atualização dos dados, no qual o custo para manter a

integridade é muito alto, esta característica não possui importância em um *Data Mart* por se tratar de uma estrutura de dados que sofre pouca ou nenhuma atualização.

2.3.2 Esquema Floco de Neve

O Esquema Floco de Neve é uma extensão do Esquema Estrela e consiste na decomposição de uma ou mais dimensões, formando hierarquias nas dimensões, isto é, normalizando-as. Esse tipo de esquema é utilizado quando se tem dimensões grandes que são estáticas ou semi-estáticas. A Figura 2.5 ilustra um exemplo geral deste tipo de esquema, nele as dimensões 2 e 4 foram normalizadas.

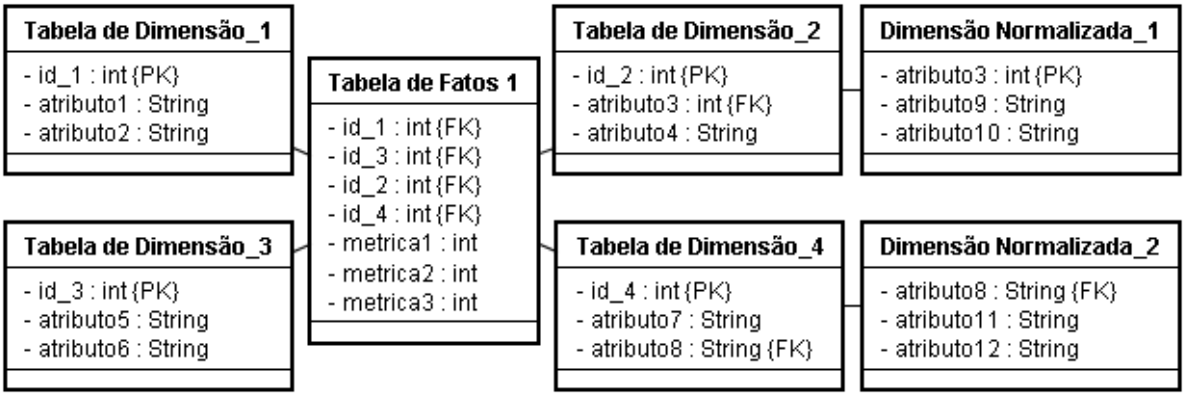


Figura 2.5 - exemplo geral do esquema floco de neve

A vantagem do seu uso está na diminuição do volume de dados trazido para a memória, além dos *inner join* com a tabela normalizada ser mais facilmente resolvido. No Esquema Floco de Neve o número de relacionamentos entre as tabelas é maior, fazendo com que o tempo de execução das consultas aumente devido à necessidade de operações de junção. Durante a especificação das tabelas do *Data Mart* é importante levar em consideração estas características de forma a normalizar as tabelas somente nos casos em que não haja uma grande perda de desempenho. Em geral, recomenda-se utilizar o Esquema Estrela ou o Esquema Constelação de Fatos, pois ambos possuem dimensões desnormalizadas.

2.3.3 Esquema Constelação de Fatos

O Esquema Constelação de Fatos é constituído de duas ou mais tabelas de fatos que compartilham uma ou mais dimensões. Esse tipo de esquema pode ser visto como uma coleção de esquemas estrelas, conforme ilustra a Figura 2.6, na qual a tabela Dimensão 2 e Dimensão 4 são compartilhadas pela Tabela de Fatos 1 e 2.

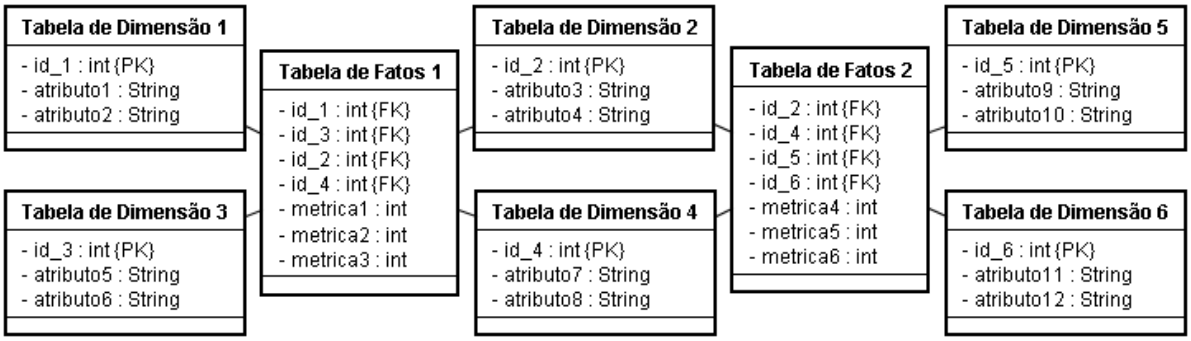


Figura 2.6 - exemplo geral do esquema constelação de fatos

Para *Data Warehouses* (ou *Data Marts*), o esquema de Constelação de Fatos é mais comumente utilizado, visto que ele pode modelar assuntos múltiplos e inter-relacionados. Desta forma, o Esquema Constelação de Fatos foi o que apresentou mais adequação para a modelagem dimensional do *Data Mart* implementado para este trabalho, visto que algumas tabelas de dimensão precisaram ser compartilhadas entre as tabelas de fatos. O capítulo 1.3, item 3.2.4, apresenta um exemplo de consulta SQL ao Esquema Constelação de Fatos modelado para o estudo de caso proposto por este trabalho, e apresenta também a tabela resultante com os valores obtidos da consulta.

Na Figura A.1 do APÊNDICE A encontra-se a modelagem completa do Esquema Constelação de Fatos para o Perfil do Setor e das Perdas Aparentes da Companhia de Abastecimento de Água e Esgoto da Paraíba. A Figura A.1 representa a tabela de fatos “Perfil do Setor” e suas 11 dimensões, juntamente com a tabela de fatos “Perdas Aparentes” associada a suas 12 dimensões. Quatro dimensões (Quadra, Matrícula, Inadimplência e Referência de Consumo) são compartilhadas pelas duas tabelas de fatos.

2.4 TECNOLOGIAS OLAP

Inicialmente, surgiram as tecnologias conhecidas como *On-Line Transaction Processing* (OLTP) que atendem às necessidades de operações transacionais. Elas denotam as movimentações tradicionais que acessam registros pequenos e individuais. As principais operações neste tipo de processo são alteração, inclusão, exclusão e consultas. Estas operações ocorrem muitas vezes em um mesmo dia e podem ser requisitadas ao sistema simultaneamente por muitos usuários, o que demanda uma resposta quase imediata do sistema. (AURÉLIO, et al., 2000)

As tecnologias *On-Line Analytical Processing* (OLAP), por sua vez, são projetadas para apoiar análises e consultas, além de auxiliar seus usuários a sintetizar informações através de comparações, visões personalizadas e análises históricas. As tecnologias OLAP têm como característica principal permitir uma visão mais fácil e intuitiva dos dados multidimensionais, por meio de análises em diferentes perspectivas (INMON, 2005).

De acordo com (HAN, et al., 2006), OLAP faz parte do processo que habilita usuários a explorar os dados do *Data Warehouse*, fornecendo funcionalidades para análise interativa de dados em diferentes dimensões e granularidades.

Alguns tipos de informações podem ser interessantes ao gerente de uma companhia de abastecimento, como por exemplo: “Qual a quantidade de consumidores, pontos de utilização e quantidade de inadimplências da subcategoria FAVELA, agrupados pelas categorias de consumo (Comercial, Industrial, Público e Residencial), situações da ligação de água (Cortada, Ligada, Suprimida parcial e Suprimida total) e estado de inadimplência (Inadimplência e Adimplência) dos consumidores?”, ou ainda, “Qual a média de faturamento das quadras agrupadas pela categoria de consumo comercial e semestres de referência (primeiros seis meses e últimos seis meses de medição)?”. Estas e outras consultas utilizando tecnologias OLAP são apresentadas em detalhes a partir da seção 3.2.6, página 93.

O processamento analítico é necessário em diversas situações no qual se deseja obter informações referentes à evolução histórica. Tecnologias OLAP permitem esses tipos de consultas e melhoram o desempenho de tempo em relação àquelas feitas em BD convencionais, ou seja, BD relacionais.

O *On-line Analytical Processing* (OLAP), ou Processamento Analítico On-Line, surgiu pela necessidade de minerar conhecimento e padrões em diferentes níveis de abstração através de análises multidimensionais dos dados, ou seja, uma visão lógica dos dados. É uma análise interativa dos dados, através de agregações em todas as interseções de dimensões necessárias. Permite obter informações sumarizadas e mostrá-las em tabelas 1-D (planilhas), 2-D (dimensões em xy), 3-D (dimensões em xyz), mapas e gráficos, com suporte para modificações dos eixos. Além disso, compõe análises estatísticas (razões, médias, somatórios, mínimos, máximos, contagens, variâncias, etc.) envolvendo quaisquer medidas ou dados numéricos entre muitas dimensões. A Tabela 2.4 mostra as diferenças entre as duas abordagens, OLTP *versus* OLAP.

Tabela 2.4 - diferenças entre OLAP e OLTP

OLAP	OLTP
- Relevância para dados históricos;	- Mantém usualmente a situação corrente;
- Necessidade de ver o dado sob diferentes perspectivas: aplicações dinâmicas;	- Voltado para velocidade e automação de funções repetitivas;
- Atualizações quase inexistentes, apenas novas inserções;	- Atualizações em grande número;
- Baseado em dados históricos, consolidados e frequentemente totalizados;	- Baseado em transações;
- Operações de agregação e cruzamentos.	- Alto nível de detalhe.

Fonte: (COLAÇO, 2004)

De acordo com (GONZALES, 2003), o termo OLAP também é usado para descrever a estrutura de armazenamento dos dados e os métodos utilizados para acessá-los. OLAP representa diversos tipos de tecnologias que variam no método de acesso. Há três adaptações de métodos de acesso OLAP, que são: OLAP Multidimensional (MOLAP); OLAP Relacional (ROLAP); OLAP Híbrido (HOLAP).

Os métodos de acesso do tipo MOLAP utilizam a estrutura de dados multidimensional e permitem a navegação pelos níveis de detalhamento em tempo real. Utiliza SGBDs Multidimensionais otimizados ao máximo para as consultas OLAP e com tratamento dimensional nativo. Requer migração dos dados do SGBD Relacional para o armazenamento multidimensional e a sua constante atualização. Teoricamente, é a melhor arquitetura de acesso a ambientes multidimensionais, mas na prática deixa a desejar pela falta de SGBDs Multidimensionais mais consolidados, dificultando sua aplicação.

Os métodos de acesso do tipo ROLAP é a solução mais utilizada hoje e surgiram em decorrência do uso consagrado dos SGBDs Relacionais nos BDs operacionais (transacionais), com todas as vantagens da tecnologia aberta e padronizada da linguagem SQL. Os dados obtidos dos bancos fontes são armazenados em SGBDs Relacionais, formando o *Data Warehouse* com tabelas implementadas em estruturas relacionais clássicas. O método de acesso ROLAP foi a solução adotada neste trabalho.

É uma tendência dos SGBDs Relacionais modernos adicionarem uma arquitetura multidimensional para prover facilidades à ambientes de suporte a decisão. Tal conceito fez surgir os métodos de acesso do tipo HOLAP, isto é, mistura do ROLAP com o MOLAP, que proporciona o desempenho e flexibilidade de um BD Multidimensional e mantém a gerenciabilidade, escalabilidade, confiabilidade e acessibilidade conquistadas pelos BDs

Relacionais. A idéia é armazenar dados de maior granularidade do DW em estruturas relacionais normalizadas e os dados agregados de granularidade inferior em estruturas dimensionais nativas.

A visualização multidimensional dos dados através das tecnologias OLAP favorece a análise de várias dimensões em única tela, em virtude da estrutura conceitual conhecida por cubos de dados. A visualização se dá através de configurações tridimensionais de linhas, colunas, operações *Slice and Dice* e gráficos, como mostra a Figura 2.7. Os cubos de dados e operações *Slice and Dice* serão discutidos nas seções 2.4.1 e 2.4.2, respectivamente.

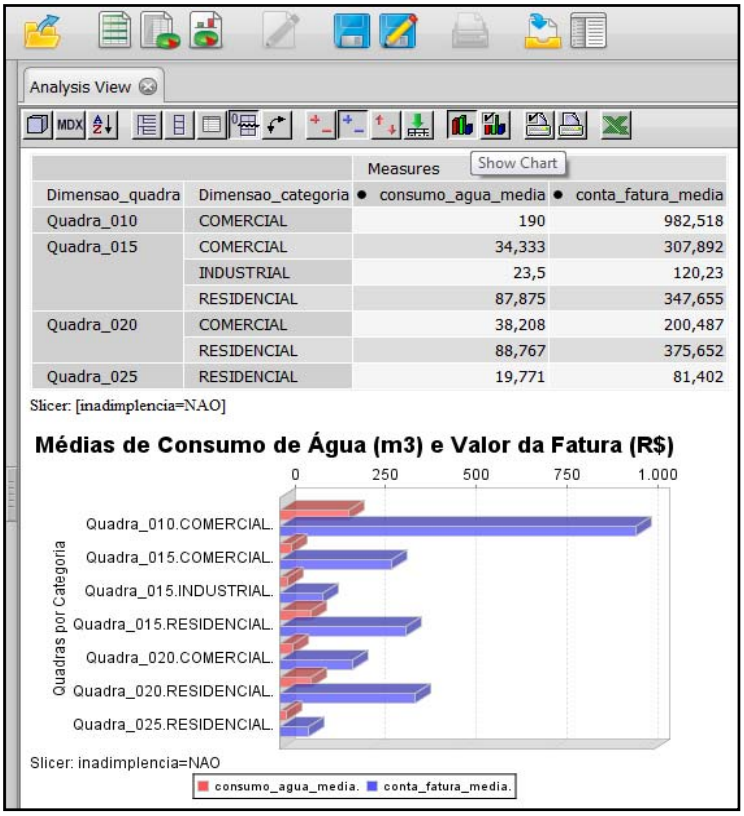


Figura 2.7 - visualização dos dados através de ferramenta OLAP pentaho analysis view[5]

Fonte: Dados do setor de saneamento de João Pessoa.

Os dados da Figura 2.7 foram obtidos através de uma consulta ao "Esquema Constelação de Fatos" implementado para o estudo de caso deste trabalho. O retorno desta consulta corresponde às médias de consumo de água em m^3 e médias da fatura dos

[5] A ferramenta *OLAP Pentaho Analysis View* será discutida com mais detalhes na seção 3.2.6 (página 119).

consumidores adimplentes[6] agrupadas por quadra (010, 015, 020 e 025) e por categoria de consumo durante o período de 2007 a 2008.

A Figura 2.8 ilustra a mesma consulta executada acima, contudo, utilizando o software *pgAdmin III* (desenvolvido para dar suporte ao SGBD *PostgreSQL*).

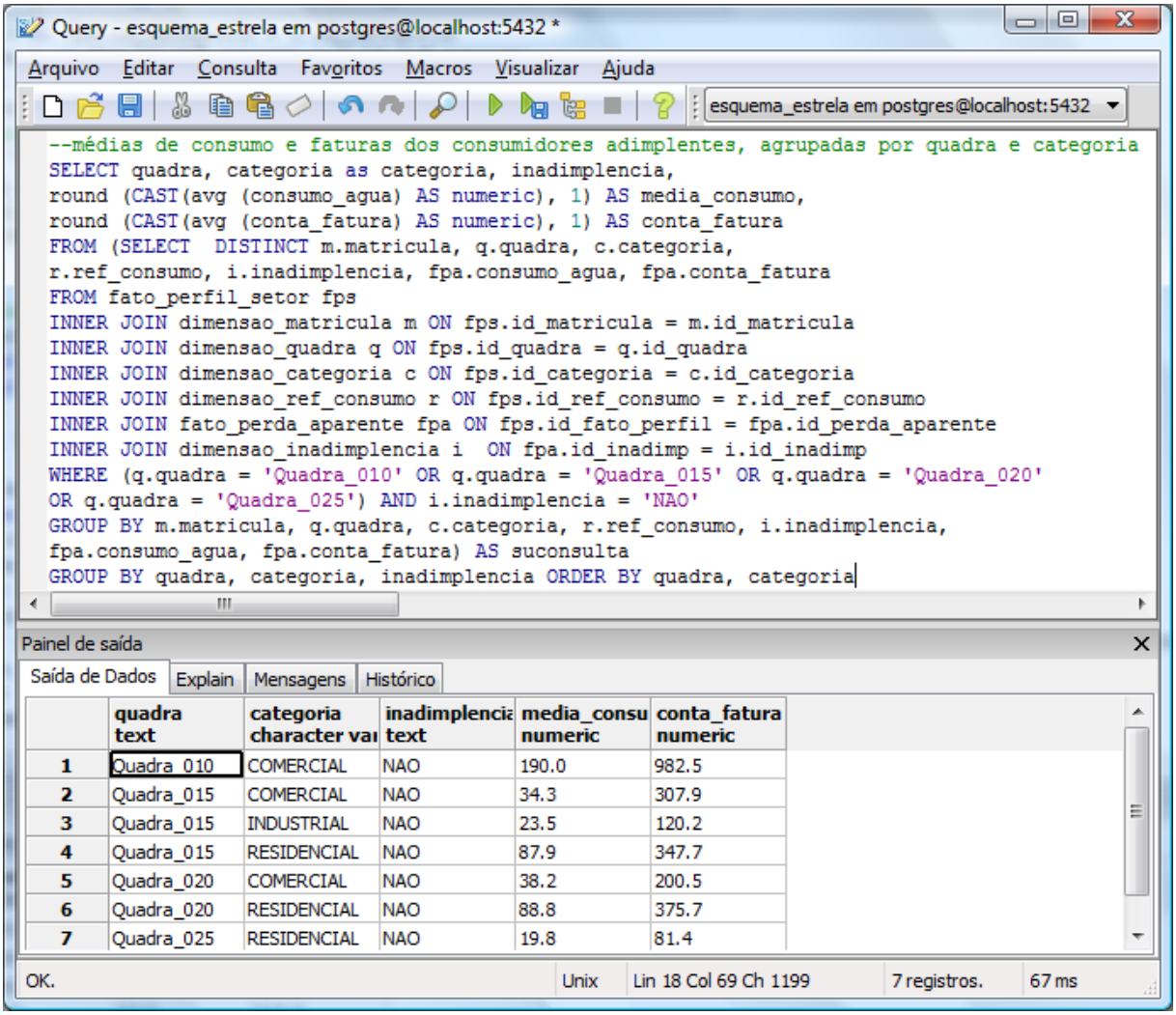


Figura 2.8 - visualização dos dados através do software PgAdmin

A principal vantagem em utilizar uma ferramenta OLAP ao invés de uma ferramenta puramente de Banco de Dados, é a facilidade proporcionada pela ferramenta OLAP quanto à visualização e manipulação do modelo dimensional (tabelas de fatos e dimensões). Outra vantagem é que o analista não precisa escrever as *queries* SQL, como ocorre em ambientes puramente de BD, pois a ferramenta OLAP dispõe de *interface* gráfica para dá o suporte a

[6] Inadimplência igual a "NAO" significa que a conta de água foi quitada pelo consumidor junto à companhia de distribuição de água.

realização das consultas. Neste trabalho optou-se por utilizar a ferramenta *OLAP Pentaho Analysis View*, que é apresentada no Capítulo 3, item 3.2.6.

2.4.1 Estrutura Multidimensional: Cubo de Dados

A principal característica das tecnologias OLAP é permitir uma visão conceitual multidimensional dos dados de uma empresa. Um cubo de dados é uma estrutura que armazena os dados em formato dimensional. Uma dimensão é uma unidade de análise com dados agrupados.

Por exemplo, a dimensão tempo tem os dados agregados por meses, quadrimestres e semestres. A dimensão categoria tem os dados agregados em comercial, industrial, público e residencial, etc. A Figura 2.9 apresenta os dados modelados numa estrutura conhecida por Cubo, onde cada Dimensão (D1, D2 e D3) representa um tema importante da companhia para realização de análises e comparações. O cubo da Figura 2.9 é "Fato Perfil do Setor" e suas dimensões são Categoria, Status da Água e Status do Esgoto.

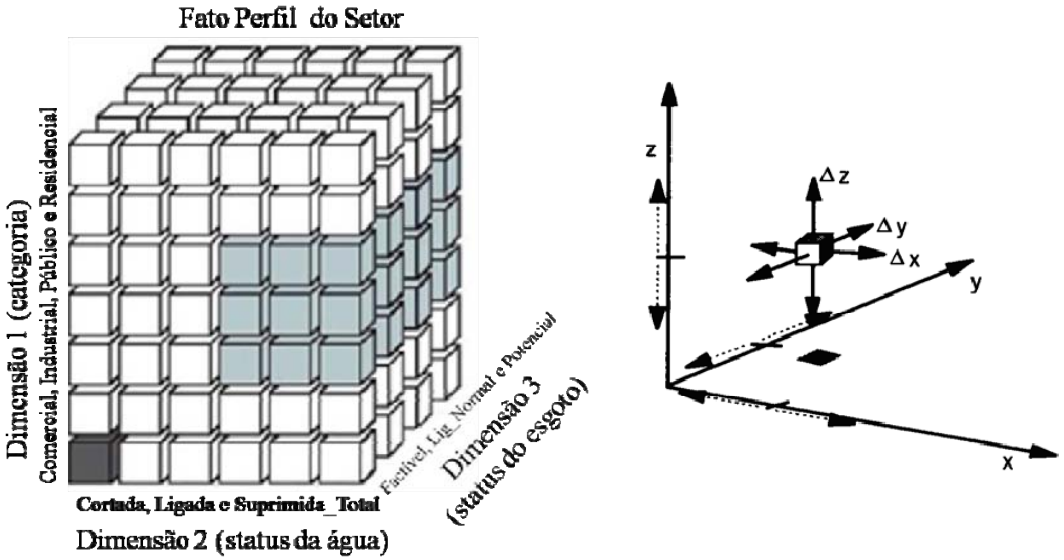


Figura 2.9 - (a) um cubo de dados com três dimensões. (b) busca tridimensional de células no cubo

Fonte: Adaptação de (RAINARDI, 2008).

A partir da modelagem do Esquema Estrela, Floco de Neve ou Constelação de Fatos pode-se construir os cubos de dados e realizar buscas nesse espaço multidimensional. Os cubos de dados são visões lógicas multidimensionais dos dados com referência hierárquica. As tecnologias OLAP fornecem funcionalidades para análise interativa de dados em diferentes visões e granularidades, permitindo visualizar as hierarquias e navegar pelas dimensões (THOMSEN, 2002).

As operações sobre os cubos de dados foram introduzidas por (GRAY, et al., 1996) visando suportar múltiplas agregações em sistemas de Banco de Dados com suporte a OLAP. O operador Cubo é uma generalização n-dimensional da operação *group-by*, sendo capaz de executar diversos *group-by* correspondentes a diversas combinações.

Na Figura 2.10 é apresentada a idéia envolvendo os operadores de cubo de dados, para isto utilizaram-se as dimensões categoria, situação da água e situação do esgoto, ambas associadas à tabela de fatos "Fato Perfil do Setor" do esquema Constelação de Fatos[7].

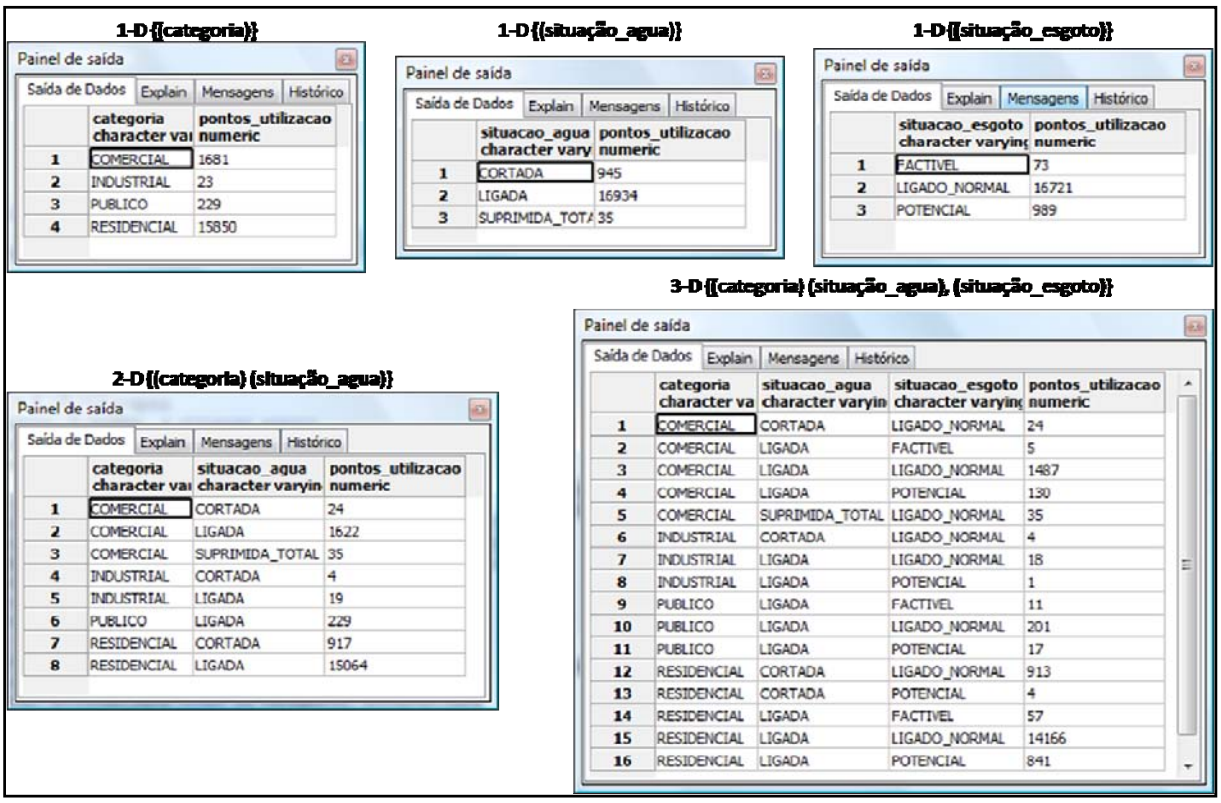


Figura 2.10 - exemplo de *cuboids* (1-D), (2-D) e (3-D) para o esquema constelação de fatos

Os agrupamentos das dimensões do esquema constelação de fatos para o perfil do setor geram a computação da ordem 2^3 agregações, ou seja, 8 *group-by* (*cuboids*) formado pelas combinações 3-D {(categoria) (situação_agua) (situação_esgoto)}; 2-D {(categoria) (situação_agua)}, {(categoria) (situação_esgoto)}, {(situação_agua) (situação_esgoto)}; 1-D {(situação_agua)}, {(situação_esgoto)}, {(categoria)}; e (vazio)[8].

[7] Esquema Constelação de Fatos encontra-se ilustrado na Figura A.1 do APÊNDICE A.
[8] (vazio) representa um *group-by* vazio

No exemplo da Figura 2.10 a dimensão categoria foi associada à dimensão situação da água, o que resultou no *cuboids* de duas dimensões (2-D). A Figura 2.11 ilustra a rede de cubóides completa formada pelas três dimensões agrupadas em cuboids de uma, duas e três dimensões.

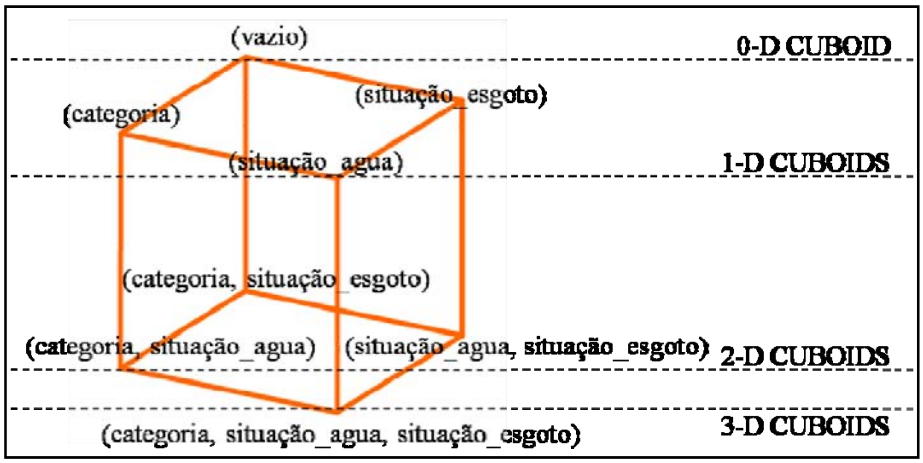


Figura 2.11 - Rede de *cuboids* para um cubo de três dimensões

Estudos voltados para manipulação eficiente da estrutura dimensional dos cubos de dados, bem como seus mecanismos de acesso, estão em constante evolução. As pesquisas nesta área buscam otimizar cada vez mais as consultas e operações OLAP, visando o melhor desempenho dos sistemas de apoio à decisão.

2.4.2 Conjunto de Operações OLAP

Ao iniciar uma consulta a um *Data Warehouse* é necessário traduzi-la de forma inteligível ao ambiente computacional. Assim, devem ser oferecidos aos analistas meios para realizar eficientemente uma consulta, a fim de obter resultados coerentes. Como solução, os desenvolvedores de ferramentas OLAP fornecem suporte para as operações de derivação de dados complexos, que recebem o nome de *Slice and Dice*.

Segundo (WREMBEL, et al., 2007), o suporte às operações *Slice and Dice* é uma das principais características de uma ferramenta OLAP. A operação *Slice*, suportada pelas ferramentas OLAP, faz restrição de um valor ao longo de uma dimensão. Já a operação *Dice* é mais complexa, pois faz restrições de valores em várias dimensões.

O *Slice and Dice* compreende quatro operações, que são o *Ranging, o Drilling, o Rotation/Pivotting e o Ranking*. A Figura 2.12 ilustra de forma genérica a operação de *Slice, Dice, Rotate, Drill-down e Drill-up/Roll-up.*

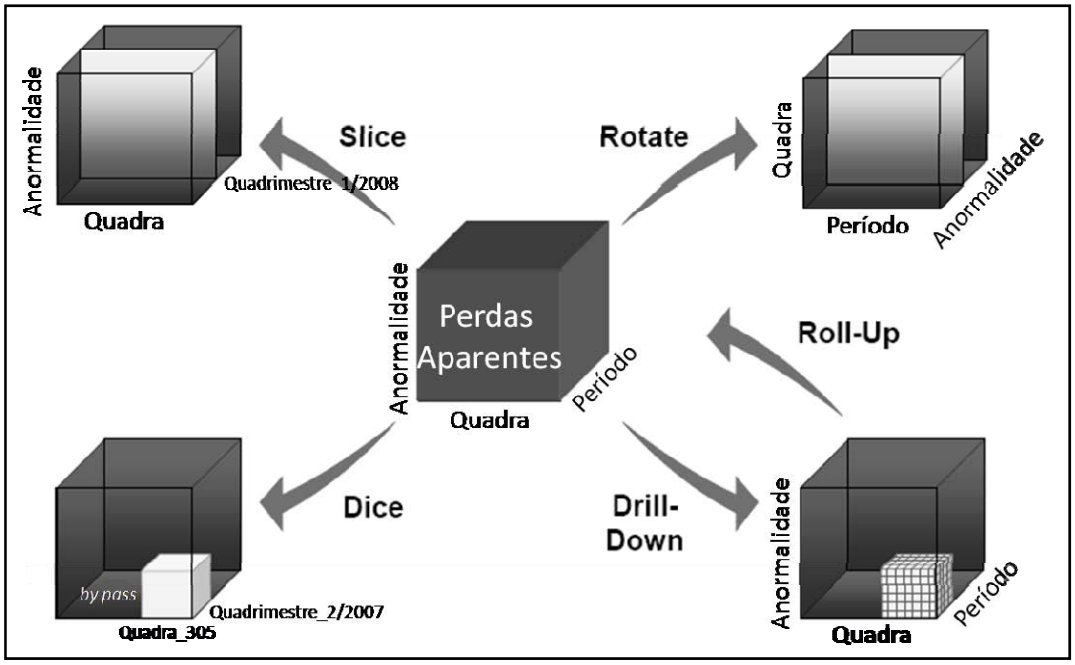


Figura 2.12 - exemplo da operação slice, dice, drill-down, drill-up e rotate.

A operação *Ranging* é responsável por, a qualquer momento, alterar o resultado das consultas, inserindo novas posições ou removendo as que estão em foco. Por exemplo, a inserção de uma nova quadra de consumidores em uma consulta representa uma operação de *Ranging*. O resultado dessa operação será considerado para todas as demais, e assim, pode-se encarar o resultado como um novo cubo gerado a partir do cubo original.

A operação *Drilling* consiste em escolher o que deseja analisar, o analista ainda pode mudar o escopo do que está analisando, porém os dados podem encontrar-se agregadas em diversos níveis. A Figura 2.12 ilustra a operação *Drilling.* O *Drilling* permite navegação por entre os níveis. Existem três operações OLAP que permitem mudar o escopo dos dados, são elas: *Drill-Down*, *Drill-Up* e *Drill-Across*.

A operação *Drill Down* navega verticalmente na hierarquia, no sentido em que os dados são mais atômicos (valores simples, indivisíveis e monovalorados). Consiste em desagregar dimensões. Exemplo: Semestre → Quadrimestre. A operação *Drill-Across* permite navegar transversalmente no eixo da árvore hierárquica. O *Drill-Across* é uma operação de grande utilidade, pois permite inserir e retirar posições do corrente cenário. O *Drill Up* ou *Roll Up* faz parte da operação *Drilling* e realiza a função inversa do *Drill-Down*. Ela permite ao usuário uma visão mais agregada das informações. Exemplo: Quadrimestre → Semestre.

Com esta técnica pode-se navegar nos diversos níveis de maiores detalhes para os níveis mais sumarizados.

A operação *Rotation ou Pivot* além de permitir ao analista mudar as posições das dimensões em foco, e tem também a flexibilidade de alterar o eixo de visualização dos dados, alterando linhas por colunas, com intuito de facilitar a compreensão. Vale salientar que *Rotation* não adiciona nem retira posições do cenário, mas permite ao analista alterar a visão que se tem dos dados. Um exemplo desta operação seria alterar a dimensão quadra da horizontal para a vertical, e desta forma, o cubo rotacionaria no sentido horário.

Com a operação *Ranking* o analista pode filtrar as informações que se deseja obter. É possível fazer uma classificação dos dados adquiridos e operar diretamente sobre os valores das células. Todas as operações anteriores atuavam apenas sobre as posições ou dimensões dos dados, entretanto, através do *Ranking*, o analista pode executar diversos tipos de filtros, eliminando assim os dados desnecessários e inconsistentes.

De acordo com o estudo de caso desenvolvido neste trabalho, as operações OLAP foram aplicadas nas tabelas de fatos e dimensões do *Data Warehouse* Comercial implementado para um setor do saneamento da cidade de João Pessoa. Em todos os casos e exemplos utilizou-se a ferramenta de código aberto *OLAP Pentaho Analysis View*[9].

2.5 DATA MINING

As técnicas de *Data Mining* podem ser aplicadas em diversas áreas do conhecimento, dentre elas na Engenharia Hidráulica, que por sua vez, é o objeto do estudo de caso deste trabalho. A sua principal característica é a aplicação dos algoritmos aos dados pré-processados, com o objetivo de auxiliar as companhias, que no caso deste trabalho é a companhia de abastecimento de água e esgoto, a gerar indicadores numéricos, indicadores gráficos e relatórios *ad hoc*, i.e., relatórios onde o analista define o que deseja obter no momento da consulta, através de aplicações que possam servir de apoio à tomada de decisão nos diferentes níveis, sejam eles estratégicos, táticos ou operacionais.

De acordo com (BATISTA, 2003), as etapas de *Data Mining* são:

---

[9] A ferramenta *Pentaho Analysis* faz parte do software livre *Pentaho BI Suite Enterprise Edition* , que se encontra na versão 3.0 disponível em <http://www.pentaho.com/products/analysis/>.

- Escolha da tarefa de *Data Mining*: uma combinação de tarefas deve ser escolhida dentre os vários tipos de tarefas possíveis como: classificação, regressão, associação, *clustering* (Ver Item 2.5.4);
- Escolha do algoritmo de *Data Mining*: de acordo com a tarefa selecionada, um determinado algoritmo, também denominado de técnica, será aplicado nos dados, utilizando-se os modelos e parâmetros mais apropriados (Ver Item 2.5.5);
- Aplicação de *Data Mining*: busca por padrões de interesse particular em uma forma representacional particular ou em um conjunto de aplicações.

2.5.1 Metas do Data Mining

Existem duas metas primárias que podem ser alcançadas através de *Data Mining* (FAYYAD, et al., 1996):

- PREVISÃO: antecipar os valores de variáveis desconhecidas ou analisar um possível valor para uma variável com o passar do tempo, utilizando algumas variáveis, como atributos da base de dados. Logo, indica as chances de uma ação ocorrer.
- DESCRIÇÃO: procurar por padrões que descrevem os dados e que sejam de entendimento dos usuários.

A Mineração Preditiva consiste na generalização de exemplos ou experiências passadas com respostas conhecidas ou regras de negócio estabelecidas por especialistas. A Mineração Descritiva consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada.

2.5.2 Aprendizado Indutivo

A indução é um meio de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. É caracterizada como o raciocínio que parte do específico para o geral, do particular para o universal, da parte para o todo.

De acordo com (BATISTA, 2003), um argumento indutivo e correto pode, perfeitamente, admitir uma conclusão falsa, ainda que suas premissas sejam verdadeiras. Se as premissas de um argumento indutivo são verdadeiras, o melhor que pode ser dito é que a sua conclusão é *provavelmente* verdadeira. Desta forma, esse recurso deve ser utilizado com os devidos cuidados, dado que se o número de observações for insuficiente ou se os dados relevantes forem mal escolhidos, as hipóteses induzidas poderão produzir conclusões

errôneas. Apesar disso, a inferência indutiva é um dos principais meios de criar novos conhecimentos e prever eventos futuros.

O *Data Mining* compreende dois tipos de aprendizado indutivo: Supervisionado e Não-Supervisionado. O aprendizado Supervisionado é direcionado a tomada de decisão e é através dele onde se realiza inferências nos dados com o intuito de realizar predições, envolvendo o uso dos atributos para prever o valor futuro. Enquanto que no Aprendizado Não-Supervisionado as atividades são descritivas, o que permite a descoberta de padrões e novos conhecimentos.

2.5.2.1 Aprendizado Supervisionado

O aprendizado supervisionado serve para identificar a classe a que pertence uma nova amostra de dados. Neste tipo de aprendizado é sempre conhecida a classe dos dados que são usados para treino e há um histórico de dados que permite prever sobre dados futuros.

Inicialmente é fornecido ao sistema de aprendizado um conjunto de exemplos E = {E_1, E_2,..., E_N}, onde cada exemplo $E_i \in E$ possui um rótulo associado. Esse rótulo define a classe a qual o exemplo pertence. Formalmente, cada exemplo $E_i \in E$ corresponde a uma tupla $E_i = (\overrightarrow{x_i}, y_i)$. Sendo $\overrightarrow{x_i}$ um vetor de valores que representam as características (atributos) do exemplo E_i, e y_i o valor da classe desse exemplo. O objetivo do aprendizado supervisionado é induzir um mapeamento geral dos vetores $\overrightarrow{x_i}$ para valores y. Portanto, o sistema de aprendizado deve construir um modelo, tal que y = f ($\overrightarrow{x_i}$), onde f é uma função desconhecida (função conceito) que permite predizer valores y.

2.5.2.2 Aprendizado Não-Supervisionado

Neste tipo de aprendizado o rótulo da classe de cada amostra de treino não é conhecido e o número de classes a ser treinada pode não ser conhecido a priori. É fornecido ao sistema de aprendizado um conjunto de exemplos E, no qual cada exemplo consiste somente de vetores $\overrightarrow{x_i}$, não incluindo a informação sobre a classe y. O objetivo é construir um modelo que procura por regularidades nos exemplos, formando agrupamentos ou clusters de exemplos com características similares.

O aprendizado não-supervisionado utiliza-se de algoritmos descritivos. As atividades descritivas trabalham com conjuntos de dados que não possuem uma classe determinada e têm

o objetivo de identificar padrões de comportamento semelhantes nestes dados. As tarefas descritivas podem ser divididas em: Associação, Segmentação e Generalização.

A Figura 2.13 apresenta a divisão dos algoritmos de *Data Mining* de acordo com a tarefa (preditiva ou descritiva) da qual fazem parte. Todas as tarefas apresentadas nesta Figura 2.13 serão detalhadas na seção 2.5.4, com ênfase para as tarefas de Classificação e Associação, visto que elas foram utilizadas nos modelos de *Data Mining* aplicados ao estudo de caso.

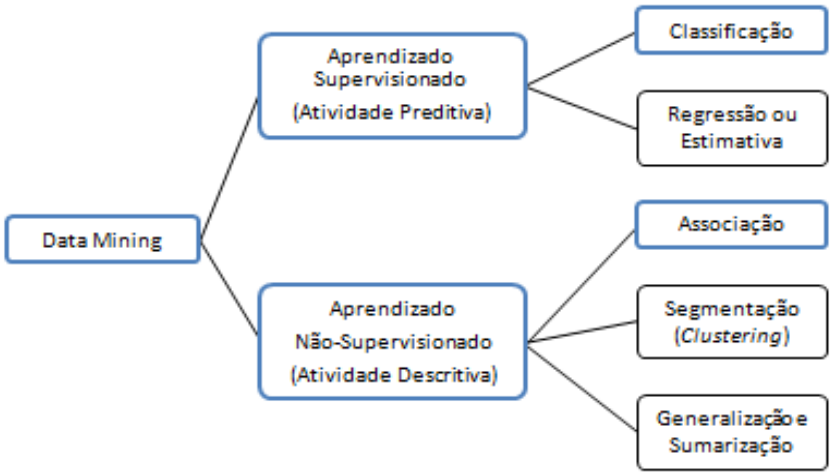


Figura 2.13 - taxonomia do d*ata mining*

Adaptação (REZENDE, et al., 2003)

2.5.3 O Processo Iterativo do Data Mining

O primeiro passo no processo de *Data Mining* é a identificação da fonte de dado. A tarefa de identificar os dados começa com a decisão sobre que dados serão necessários para resolver o problema. O próximo passo é o *cleaning*, ou seja, a preparação dos dados. O principal desafio do *cleaning* é formatar os dados de forma compatível com a representação do *software* que será utilizado para mineração.

O terceiro passo é construção do modelo de mineração, e este é constituído das regras que descrevem os dados analisados no banco. Isso é feito automaticamente através de dados analíticos e de algoritmos de *Data Mining*. O quarto passo no processo iterativo de mineração é a avaliação do modelo criado, que consiste em estimar a precisão do modelo e refinar sua compreensão e sua utilidade. Por último surge o desdobramento do modelo, que serve para aplicar o modelo a novos dados, a fim de dar lugar ao surgimento de novas perguntas, trazendo um refinamento adicional às descobertas (SANCHES, 2003).

Ao final do processo de mineração, espera-se, como principal objetivo, o uso das descobertas úteis e suas representações. Desta forma, seguem as ações para a etapa de pós-processamento:

- Interpretação dos Padrões: avaliação e interpretação dos padrões encontrados, a fim de determinar aqueles que terão alguma utilidade e gerarão algum conhecimento. Nesta etapa, pode ocorrer a necessidade de retorno a umas das etapas anteriores;
- Consolidação do Conhecimento: verificação e utilização do novo conhecimento sobre os dados através das ferramentas de visualização. E produção da documentação para auxiliar a compreensão do usuário.

2.5.4 Principais Tarefas do Data Mining

A tarefa de *Data Mining* precisa ser definida no início do processo de KDD, no momento em que for decidido o domínio da aplicação (FAYYAD, et al., 1996). Existem diversas tarefas para alcançar as metas de previsão e descrição, discutidas na seção 2.5.1, dentre elas: Classificação; Regressão ou Estimativa; Associação; Segmentação (*clustering*) e Generalização ou Sumarização. Nas seções seguintes serão descritas as duas tarefas de mineração utilizadas no trabalho

2.5.4.1 Classificação

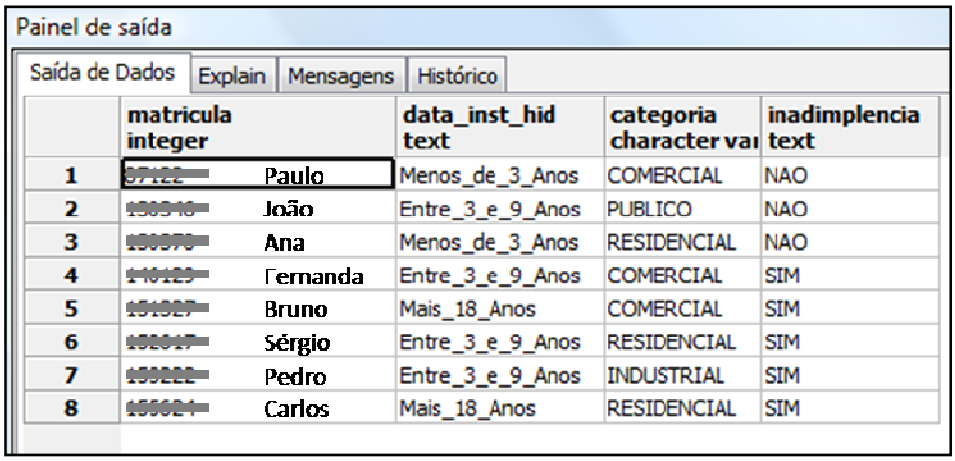
A tarefa de classificação consiste em encontrar propriedades comuns em um determinado conjunto de objetos de um banco de dados e classificá-los em diferentes classes. Os passos para classificação são: definição de um conjunto de exemplos conhecidos (treinamento); treinamento sobre esse conjunto; e geração de regras de classificação ou descrição.

Conforme (BARROSO, et al., 2006), o princípio desta tarefa é descobrir algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo a descobrir um conhecimento que possa ser utilizado para prever a classe de uma tupla desconhecida, ou seja, que ainda não possui uma classe definida.

O conhecimento descoberto é frequentemente representado na forma de regras SE→ENTÃO. Essas regras são interpretadas da seguinte maneira: “SE os atributos preditivos

de uma tupla satisfazem as condições no antecedente da regra, ENTÃO a tupla tem a classe indicada no consequente da regra".

Como exemplo da tarefa de classificação utilizou-se os dados da Figura 2.14. Ao aplicar a Classificação sobre esses dados, são geradas as regras de classificação, conforme apresenta a Tabela 2.5


Painel de saída

Saída de Dados | Explain | Mensagens | Histórico

	matricula integer		data_inst_hid text	categoria character var	inadimplencia text
1	[redacted]	Paulo	Menos_de_3_Anos	COMERCIAL	NAO
2	[redacted]	João	Entre_3_e_9_Anos	PUBLICO	NAO
3	[redacted]	Ana	Menos_de_3_Anos	RESIDENCIAL	NAO
4	[redacted]	Fernanda	Entre_3_e_9_Anos	COMERCIAL	SIM
5	[redacted]	Bruno	Mais_18_Anos	COMERCIAL	SIM
6	[redacted]	Sérgio	Entre_3_e_9_Anos	RESIDENCIAL	SIM
7	[redacted]	Pedro	Entre_3_e_9_Anos	INDUSTRIAL	SIM
8	[redacted]	Carlos	Mais_18_Anos	RESIDENCIAL	SIM

Figura 2.14 - exemplo de dados utilizados na tarefa de classificação

Foram geradas quatro regras de classificação. Por exemplo, a primeira regra determina que todos os hidrômetros com mais de 18 anos de funcionamento são de consumidores inadimplentes. Enquanto que a última regra determina que os hidrômetros entre 3 e 9 anos de funcionamento e que pertencem aos consumidores que não sejam da categoria Público, estão inadimplentes.

Tabela 2.5 - regras de classificação geradas (descobertas) com os dados da Figura 2.14

SE (Instalação_Hid = Mais_18_Anos) ENTÃO Inadimplência = Sim
SE (Instalação_Hid = Menos_de_3_Anos) ENTÃO Inadimplência = Não
SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria = Público) ENTÃO Inadimplência = Não
SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria != Público) ENTÃO Inadimplência = Sim

No contexto deste trabalho, seguem alguns exemplos onde a tarefa de classificação poderia ser aplicada: classificação do consumidor quanto ao risco (baixo, médio ou alto risco) de inadimplência; classificação do consumidor potencialmente fraudador a julgar pelo seu perfil; classificação das categorias quanto às anormalidades; classificação do tipo de ligações de água quanto legal, clandestina ou suspensa por irregularidade etc.

2.5.4.2 Associação

A tarefa de associação foi introduzida por (AGRAWAL, et al., 1993) e tem a finalidade de determinar os grupos de itens que tendem a ocorrer ao mesmo tempo, em uma mesma transação, gerando-se as regras de associação. Elas podem ser vistas como regras do tipo SE-ENTÃO integrada a duas medidas de interesse: *confiança* e *suporte*. A primeira medida corresponde à probabilidade condicional e a segunda corresponde à fração que sustenta a regra. Ambas serão definidas mais adiante.

A regra de associação é um relacionamento X (antecedente) ⇒ Y (consequente), onde X e Y são conjuntos de itens da transação e a interseção X ∩ Y é o conjunto vazio. Cada regra está associada a um Fator de Suporte Superior "Fs" (medida de interesse *suporte*) e a um Fato de Confiança "Fc" (medida de interesse *confiança*). Seguem as fórmulas dos dois fatores:

$$Fs = \frac{|X \cup Y|}{N}, \text{onde N é o número total de tuplas} \qquad Fc = \frac{|X \cup Y|}{X}$$

O fator de suporte[10] pode ser descrito como a probabilidade de uma transação qualquer satisfazer tanto X como Y, ao passo que o fator de confiança[11] é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaça X. A tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras com "Fs" e "Fc" maiores ou iguais a um "Fs" e "Fc" especificado pelo analista.

A descoberta de regras de associação segue normalmente em dois passos. Primeiramente, o algoritmo determina todos os conjuntos de itens que têm "Fs" maior ou igual ao "Fs" especificado pelo analista. Estes conjuntos são chamados conjuntos de {Itens Frequentes}. Em seguida, todas as possíveis regras candidatas são geradas e testadas para cada conjunto de {Itens Frequentes} com relação ao "Fc". Apenas as regras candidatas com "Fc" maior ou igual ao "Fc" especificado pelo analista são dadas como saída do algoritmo.

Segue na Tabela 2.6 abaixo, um exemplo do processo de descoberta de regras de associação. A primeira coluna da Tabela mostra o identificador da transação, e as demais

[10] O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.
[11] O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador se refere à quantidade de transações em que o item X ocorre.

colunas indicam se um determinado item foi ou não localizado na transação correspondente. Suponha que o analista especificou os parâmetros Fs = 0.3 e Fc = 0.8.

Tabela 2.6 - exemplo de dados para descoberta de regra de associação

ID	Consumidor Comercial	Fraude	Corte Ligação	Multa	Parcelamento Fatura	Pagamento Fatura
1	N	S	S	S	N	N
2	S	S	N	S	S	N
3	N	S	S	S	N	N
4	S	S	S	S	N	N
5	N	N	N	N	S	N
6	N	N	N	S	N	N
7	N	S	N	N	N	N
8	N	N	N	N	N	S
9	N	N	N	N	N	S
10	N	N	N	N	N	N

Tabela 2.7 - descoberta de regras de associação com fs = 0.3 e fc = 0.8

Conjunto de Itens Frequentes: Corte_Ligação, Fraude. Fs = 3/10 = 0.3 Regra: SE (Corte_Ligação) ENTÃO (Fraude). Fc = 3/3 = 1.
Conjunto de Itens Frequentes: Corte_Ligação, Multa. Fs = 3/10 = 0.3 Regra: SE (Corte_Ligação) ENTÃO (Multa). Fc = 3/3 = 1.
Conjunto de Itens Frequentes: Fraude, Multa. Fs = 4/10 = 0.4 Regra: SE (Fraude) Então (Multa). Fc = 4/5 = 0.8. Regra: SE (Multa) ENTÃO (Fraude). Fc = 4/5 = 0.8
Conjunto de Itens Frequentes: Corte_Ligação, Fraude, Multa. Fs = 3/10 = 0.3 Regra: SE (Corte_Ligação E Fraude) ENTÃO (Multa). Fc = 3/3 = 1. Regra: SE (Corte_Ligação E Multa) ENTÃO (Fraude). Fc = 3/3 = 1 Regra: SE (Corte_Ligação) ENTÃO (Fraude E Multa). Fc = 3/3 = 1

Os atributos “Consumidor Comercial”, “Parcelamento Fatura” e “Pagamento Fatura” possuem Fs = 0.2 e não pertencem ao Conjunto de Itens Frequentes, visto que o Fs deve ser maior ou igual a 0.3. Já os atributos “Fraude”, “Corte Ligação” e “Multa” possuem Fs igual a 0.5, 0.3 e 0.5 respectivamente e desta forma, pertencem ao Conjunto de Itens Frequentes.

A Tabela 2.7 demonstra as regras de associação que são descobertas dos dados oriundos da Tabela 2.6 utilizando-se os valores de Fs e Fc maiores ou iguais aos especificados pelo analista, que foram respectivamente 0.3 e 0.8. Na Tabela 2.7 as regras de associação são agrupadas por três conjuntos de {Itens Frequentes}, sendo dois deles formados pelos itens

frequentes {Corte_Ligação e Fraude} e {Corte_Ligação e Multa}, e o terceiro conjunto formado pelos itens frequentes {Corte_Ligação, Fraude e Multa}.

2.5.5 Técnicas de Data Mining

Segundo afirma (BALLARD, et al., 1998), não há uma técnica que resolva todos os problemas de *Data Mining*. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e desvantagens, por isso, é importante conhecer bem o ambiente de aplicação e as técnicas disponíveis para que se possa escolher a mais adequada.

Dentre as técnicas de DM normalmente utilizadas tem-se: Árvores de Decisão; Regras de Associação; Redes Neurais Artificiais; Algoritmos Genéticos e Classificação Bayesiana. O item 3.3 discute a aplicação do algoritmo de mineração que se mostrou mais adequado para escopo do problema abordado neste trabalho.

A Tabela 2.8 apresenta a descrição das principais técnicas e tarefas de DM, e cita alguns algoritmos relacionados com as respectivas técnicas. Nas seções seguintes serão discutidas as técnicas utilizadas neste trabalho, que foram: Árvore de Decisão, Classificação Bayesiana e Regras de Associação.

Tabela 2.8 - técnicas, tarefas e algoritmos de *data mining*

Técnica	Descrição	Tarefas	Algoritmos
Árvore de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	Classificação Regressão	J4.8, One-R, ID-3, CART, CHAID, C4.5, C5.0, SPRINT, etc.
Classificação Bayesiana	Métodos estatísticos que podem predizer a probabilidade de um registro pertencer a uma determinada classe.	Classificação	NaïveBayes
Regras de Associação	Estabelece uma correlação estatística entre os atributos de dados e conjunto de dados.	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP, etc.
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.	Classificação Segmentação Regressão	Perceptron, Rede MLP, Redes ART, Rede IAC, Rede BSB, etc.
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter “descendentes”.	Classificação Segmentação	Algoritmo Genético Simples, Genitor, CHC, Algoritmo de Hillis, GANuggets, etc.

Fonte: (FAYYAD, et al., 1996)

2.5.5.1 Árvores de Decisão

As árvores de decisão são uma maneira de representar uma série de regras que conduzem a uma classe ou a um valor. De acordo com (SYMEONIDIS, et al., 2005), o objetivo principal de uma árvore de decisão é separar as classes, onde as tuplas de classes diferentes tendem a ser alocadas em subconjuntos diferentes, cada um descrito por regra simples em um ou mais itens de dados. Essas regras podem ser expressas como declarações lógicas, em uma linguagem como SQL, de modo que possam ser aplicadas diretamente a novas tuplas.

Uma das principais vantagens das árvores de decisão é o fato de que o modelo é bem explicável, uma vez que tem a forma de regras explícitas, podendo ser representada como um conjunto de regras (galhos), onde cada nó não terminal representa um teste ou decisão sobre o item considerado.

Na árvore de decisão cada nó não terminal representa um teste ou decisão sobre o item de dado. Assim, os nós representam os atributos, as ligações entre os nós representam os valores dos atributos e as folhas representam as classes. Cada caminho da árvore pode ser convertido numa regra. O nó interno e os valores das setas são convertidos no antecedente da regra (parte SE); o nó folha é convertido no consequente da regra (parte ENTÃO).

Um exemplo poderia ser a classificação de consumidores na categoria “Aceitável” ou “Risco”, onde a primeira indica confiança na adimplência do consumidor e a segunda indica risco de inadimplência perante a companhia de abastecimento de água. A Figura 2.15 ilustra uma árvore simplificada de decisão que resolve esta situação.

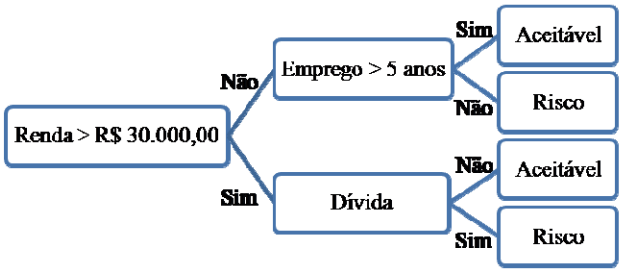


Figura 2.15 - exemplo de árvore de decisão

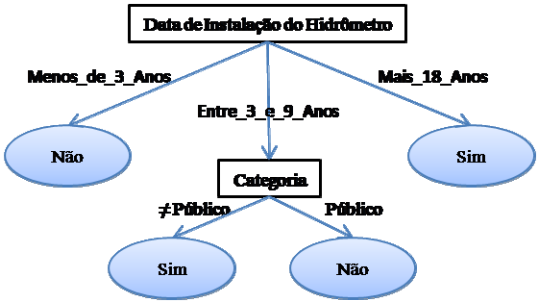


Figura 2.16 - arvore de decisão gerada com os dados da Figura 2.14

Na Figura 2.16 tem-se uma árvore de decisão para o exemplo da Figura 2.14. Cada caminho da árvore pode ser convertido numa regra. A tupla <“João”, “Entre_3_e_9_Anos”,

"Público", ?> identifica um tipo de consumidor. A interrogação (?) representa o valor do atributo objetivo (Estado de Inadimplência), e este é responsável por informar se o consumidor está inadimplente (Sim) ou adimplente (Não) perante a companhia de abastecimento de água.

O nó raiz da árvore de decisão da Figura 2.16 representa o atributo "Data de Instalação do Hidrômetro". Na tupla dada como exemplo, o nó raiz direciona a regra para data de instalação do hidrômetro "Entre_3_e_9_Anos". Seguindo a hierarquia da árvore, a regra passa pelo seu segundo nó, que é o atributo "Categoria", e o seu valor na tupla é "Público". Por fim o algoritmo direciona a regra para o atributo objetivo, que por sua vez, está no nó folha rotulado pelo valor "Não", que indica que João está na classe dos consumidores adimplente.

Geralmente uma árvore de decisão classifica uma nova tupla de maneira *top-down*, utilizando um algoritmo baseado na aproximação "dividir para conquistar" (WITTEN, et al., 2005). Inicialmente todas as tuplas que estão sendo mineradas são associadas ao nó raiz da árvore. Então o algoritmo seleciona uma partição de atributos e divide o conjunto de tuplas no nó raiz de acordo com o valor do atributo selecionado. O objetivo desse processo é separar as classes para que tuplas de classes distintas tendam a ser associadas a diferentes partições. Esse processo é recursivamente aplicado a subconjuntos de tuplas criados pelas partições, produzindo subconjuntos de dados cada vez menores, até que um critério de parada seja satisfeito. A fim de minimizar o tamanho da árvore, sem prejudicar a qualidade da solução, aplica-se algoritmo de poda de árvore de decisão.

As principais vantagens de algoritmos baseados em árvores de decisão são sua eficiência computacional e simplicidade. Devido ao uso da aproximação "dividir para conquistar", entretanto essa aproximação também possui desvantagem. Por exemplo, uma condição envolvendo um atributo que será incluído em todas as regras descobertas. Essa situação possivelmente produz regras com informações irrelevantes, além de desperdício de processamento. Três algoritmos de árvore de decisão serão analisados, são eles: Algoritmo de Indução de Regras, Algoritmo ID-3 e Algoritmo J4.8.

Outro exemplo simples de como funciona um algoritmo de classificação, que apresenta seu resultado sob a forma de árvore de decisão, está ilustrado na Figura 2.17. Neste exemplo, os consumidores podem ser classificados em confiáveis ou não confiáveis junto à

companhia de abastecimento, baseando-se na quantidade de pontos de utilização e o valor da conta (fatura a ser paga).

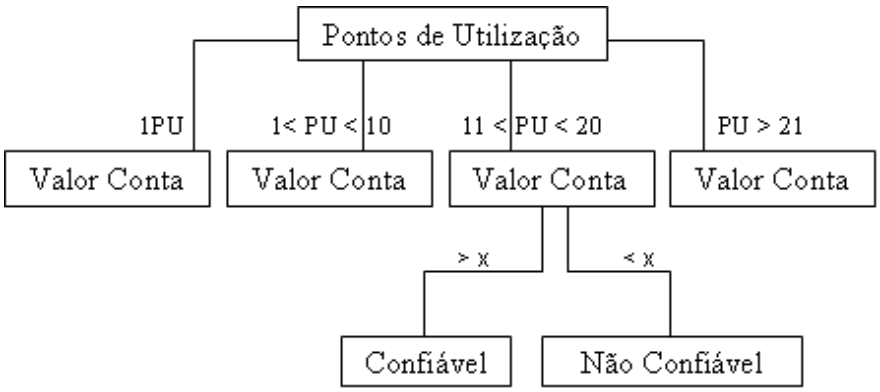


Figura 2.17 - classificação por árvore de decisão (pontos de utilização *versus* fatura)

Cada regra tem seu início na raiz da árvore e caminha até suas folhas. A interpretação de um dos galhos é que uma pessoa que possua entre 11 a 20 Pontos de Utilização (PU) de água e o Valor da Conta (VC) maior que "x" (valor em real), onde x representa um intervalo de consumos e contas aceitáveis para uma determinada categoria do sistema. Por exemplo: um consumidor que possui 15 pontos de utilização, pertence à categoria residencial cujo "x" está definido entre [R$ 50,00 e R$ 200,00], e pagou pela conta um valor menor do que R$ 50,00, então ele passará a ser classificado com não confiável. Tal regra extraída da base de dados permite ao gerente tomar a decisão de realizar uma intervenção/vistoria no abastecimento dos consumidores com baixo grau de confiança.

Existem alguns algoritmos de classificação que, ao invés de montarem uma árvore de decisão, expressam o conhecimento extraído através de regras do tipo "SE condição ENTÃO classe" ou X, Y → Z, chamadas simplesmente de Regras de Classificação. Cada galho de uma árvore de classificação representa uma regra. A seguir, são apresentadas as regras da Figura 2.17, sob a forma de regras de classificação:

Se (Pontos de Utilização corresponde ao intervalo [11..20]), (VC < x)) → Não Confiável

Se (Pontos de Utilização corresponde ao intervalo [11..20]), (VC > x)) → Confiável

Os ramos da árvore podem crescer de maneiras diferentes. Por exemplo, caso não exista consumidor com apenas 1 ponto de utilização de água (1° regra mais a esquerda) em todo o setor da rede de distribuição de água, então a regra nunca será aplicada e o ramo ficará estático. Além disso, todas as regras geradas a partir de uma AD terão que conter o atributo

raiz em seu antecedente. No exemplo da Figura 2.17, como "Pontos de Utilização" é o atributo raiz escolhido, não há como se ter uma regra do tipo: Se (" ", (x > 300)) → Confiável.

a) Algoritmo de Indução de Regras

Este tipo de algoritmo é baseado em duas idéias chaves: Estado e Operador. Um Estado é a descrição da situação de um problema num dado instante e um Operador é um procedimento que transforma um estado em outro. Resolver um problema utilizando esse algoritmo consiste em encontrar uma sequência de operadores dos quais transformam um estado inicial num estado objetivo, ou estado meta. Um estado corresponde a uma regra candidata e os operadores correspondem a operações de generalização e/ou especialização que transformam uma regra candidata em outra (LAROSE, 2005).

A principal vantagem desse algoritmo é que geralmente ele produz conhecimento compreensível e o conhecimento descoberto está na forma de regras "SE→ENTÃO", desse modo as regras podem ser facilmente entendidas e validadas pelo usuário.

Um exemplo da utilização do algoritmo de indução de regras se encontra na Tabela 2.9. Este exemplo remete-se aos dados da Figura 2.14, e que também foram utilizados na tarefa de classificação.

Tabela 2.9 - operações de especialização e generalização por indução de regras

Especializando uma regra pela adição da conjunção em seu antecedente
Regra Original: SE (Instalação_Hid = Menos_de_3_Anos) ENTÃO Inadimplência = Não Regra Especializada: SE (Instalação_Hid = Menos_de_3_Anos e Categoria = Comercial) ENTÃO Inadimplência = Não
Generalizando uma regra relaxando uma condição no antecedente
Regra original: SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria = Comercial) ENTÃO Inadimplência = Sim Regra Generalizada: SE (Instalação_Hid = Entre_3_e_9_Anos e Categoria != Público) ENTÃO Inadimplência = Sim

A operação de especialização mostra que a regra pode ser especializada pela adição de novas condições ao antecedente. Note que a nova regra é uma especialização da original porque o antecedente da regra é satisfeito por um número menor de tuplas no banco de dados. A regra original atende 2 registros (Bruno e Carlos), enquanto a regra especializada, com a conjunção "Categoria = Público", atende apenas 1 registro (Bruno).

Na generalização a idéia é estender o intervalo de valores cobertos pelo atributo "Categoria", relaxando-o de modo que o antecedente da regra satisfaça um número maior de tuplas na base de dados. No exemplo dado, os registros atendidos passou de 1 (Fernanda) para 3 (Fernanda, Sérgio e Pedro).

b) Algoritmo ID-3

O *Iterative Dichotomiser* (ID-3) proposto por J. Ross Quinlan é um dos mais conhecidos algoritmos destinados a construção de árvore de decisão para a tarefa de classificação. O algoritmo cria a árvore de decisão a partir dos exemplos de treinamento utilizando o método de indução *top-down induction of decision trees* (TDIDT). A evolução do ID-3 são os algoritmos ID-4, ID-6, C4.5 e C5.0 (QUINLAN, 1993). No tópico seguinte se dará uma maior atenção ao algoritmo C4.5[12], visto que é um dos algoritmos utilizados como técnica de *Data Mining* proposta por este trabalho.

Ele recebe como entrada um conjunto de tuplas para treinamento, chamado exemplos; um atributo objetivo, chamado meta; e um conjunto de atributos preditivos, chamado atributos. Não é considerado um algoritmo incremental, pois todos os exemplos de treinamento devem estar disponíveis no início do processo.

Para a geração da árvore, são utilizados exemplos de treinamento rotulados que possuem os atributos e a classe a que pertence. Aplica-se uma função de avaliação para cada atributo verificando aquele que discrimina melhor os conceitos positivos dos negativos, deixando na raiz da árvore de decisão o atributo mais informativo.

O processo é recursivo, gerando sub-árvores até que se atenda um critério de parada, que no caso ideal seria obter nós contendo apenas exemplos de uma mesma classe (KANASHIRO, 2007). De uma maneira geral, os passos do algoritmo ID-3 são apresentados na Tabela 2.10.

Tabela 2.10 - passos para construção da árvore de decisão através do ID-3

1. Dado um NÓ na árvore e todas as tuplas do conjunto de treinamento S; 2. Selecione o melhor atributo A para esse nó; 3. Para cada valor v_i de A, cresça uma subárvore, ou uma folha, sob o nó.

[12] É similar ao algoritmo J4.8. A diferença é que o C4.5 foi desenvolvido na linguagem C e o J4.8 em Java.