



Pós-Graduação em Ciência da Computação

**“METODOLOGIA E USO DE TÉCNICAS DE
EXPLORAÇÃO E ANÁLISE DE DADOS NA
CONSTRUÇÃO DE DATA WAREHOUSE”**

Por

ROBERTO ÂNGELO FERNANDES SANTOS

Dissertação de Mestrado



Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, SETEMBRO/2002



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ROBERTO ÂNGELO FERNANDES SANTOS

**METODOLOGIA E USO DE TÉCNICAS DE EXPLORAÇÃO E ANÁLISE DE
DADOS NA CONSTRUÇÃO DE DATA WAREHOUSE**

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco.

Orientador: Prof. Décio Fonseca

Banca: Prof. Dr. Paulo Jorge Leitão Adeodato-UFPE
Prof. Dr. Marcus Costa Sampaio – UFPB
Prof. Dr. Décio Fonseca – UFPE

Apoio: CNPq.

RECIFE, SETEMBRO 2002

EPIGRAFE

A mente, como o lar, é mobiliada pelo proprietário, portanto, se sua vida for fria e árida, a culpa será somente dele.

Louis L' Amor

Dedico esta dissertação a todos que de alguma forma participaram e contribuíram da elaboração.

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus, pela minha existência.
A toda minha família, em especial meu falecido pai Augusto Ângelo e a minha mãe Sônia Santos, que me ensinaram o valor dos estudos em minha vida.

A minha noiva por sua paciência e pelo seu apoio.

A minha tia que me acolheu em seu lar e me trata como se fosse seu filho.
Um muito obrigado a meu orientador Doutor Décio Fonseca, que me apoiou e acreditou em mim, mesmo antes do início da jornada. Sem ele tudo seria muito mais difícil ou talvez nem acontecesse.

Agradeço ao meu colega de trabalho, professor e amigo Paulo Adeodato pela orientação informal e pelas longas discussões sobre o assunto aqui abordado.

Meu agradecimento a todos os professores que sempre me deram incentivo. Em especial aos meus mestres Prof. João Gualberto e Prof^a. Simone Branco. A empresa Neurotech que me deu oportunidade de ter acesso à tecnologias de manipulação e tratamento de dados, bem como acesso à aplicações práticas e para estudos de casos reais.

Ao Núcleo de Tecnologia da Informação da UFPE (NTI-UFPE) pela disponibilidade dos dados dos COVEST e pelo espaço cedido em seu laboratório.

Agradecimento a Vitor Guedes por me ajudar em algumas implementações da dissertação, Ana Soraya por me ajudar com a formatação e correção, a Hélio por discutir assuntos relacionados à dissertação, tornando meu trabalho menos solitário.

E por fim agradeço a todos, não citados aqui, mas que me apoiaram e me ajudaram de alguma forma.

SUMÁRIO

1 – INTRODUÇÃO.....	12
1.1 – CONTEXTO.....	12
1.2 – MOTIVAÇÃO.....	13
1.3 – OBJETIVOS.....	15
1.4 – ORGANIZAÇÃO RESUMIDA DA DISSERTAÇÃO.....	16
2 – O AMBIENTE REDIRIS.....	17
2.1 – INTRODUÇÃO.....	17
2.2 – CONCEITOS BÁSICOS.....	18
2.3 – A DINÂMICA REDIRIS.....	19
2.4 – O COMPONENTE PRÉ-PROCESSADOR NO REDIRIS.....	21
3 – REVISÃO BIBLIOGRÁFICA.....	23
3.1 – CONSTRUÇÃO DE DATA WAREHOUSE.....	23
3.1.1 – Metodologia Baseada em Dados.....	23
3.1.2 – Banco de Dados Amostra Viva de INMON.....	27
3.1.3 – Granularidade.....	29
3.1.4 – Integração de Dados EM DW.....	30
3.2 – MINERAÇÃO DE DADOS.....	31
3.3 – TÉCNICAS E ALGORITMOS PARA PRÉ-PROCESSAMENTO DE DADOS.....	34
3.4 – QUALIDADE DOS DADOS.....	36
3.4.1 – Qualidade dos Dados no DW.....	37
3.4.1 – Processo de Qualidade de Dados.....	38
3.5 – EXPLORAÇÃO E ANÁLISE DE DADOS NO PRÉ-PROCESSAMENTO.....	44
3.5.1 – Visualização dos Dados no Pré-Processamento.....	45
3.5.2 – Distribuição de Frequência.....	47
3.5.3 – Dist. de Frequência na Exploração e Análise de Dados.....	50
4 – METODOLOGIA FASTCUBE.....	53
4.1 – INTRODUÇÃO.....	53
4.2 – VISÃO GERAL DA METODOLOGIA.....	54
4.3 – MAPEAMENTO E INTEGRAÇÃO DE DADOS.....	55
4.4 – FRAGMENTAÇÃO DOS DADOS.....	57
4.5 – ANÁLISE DOS DADOS.....	58
4.6 – TRATAMENTO DOS DADOS.....	61
4.7 – PROTOTIPAÇÃO RÁPIDA.....	63
5 – UMA IMPLEMENTAÇÃO DO FASTCUBE.....	71
5.1 – UM CICLO DO FASTCUBE NA ARQUITETURA DO REDIRIS.....	71
5.3 – ARQUITETURA.....	72
5.4 – O MODELO DE DADOS E METADADOS.....	75
5.3 – QUALIDADE DOS DADOS.....	79
5.3 – TÉCNICAS DO PRÉ-PROCESSAMENTO - UM MODELO EXTENSÍVEL.....	81
5.3 – TÉCNICAS IMPLEMENTADAS NO REPOSITÓRIO DE TÉCNICAS.....	84
5.4 – A INTERFACE E NAVEGAÇÃO DO PROTÓTIPO.....	87
5.5 – POSSIBILIDADES COMPLEMENTARES DO MODELO.....	91

6 – ESTUDO DE CASO – COVEST	93
6.1 – OBJETIVOS E REQUISITOS DO DATA MART - COVEST	93
6.2 – FERRAMENTAS UTILIZADAS.....	93
6.3 – DESCRIÇÃO DOS DADOS	94
6.4 – GRANULARIDADE	96
6.5 – SELEÇÃO E INTEGRAÇÃO DOS DADOS	97
6.6 – MANIPULAÇÃO E TRATAMENTO DOS DADOS	100
6.7 – GERAÇÃO DO MODELO E POVOAMENTO	101
6.8 – EXTRAÇÃO E ANÁLISE DOS RESULTADOS	102
7 – CONSIDERAÇÕES FINAIS	109
7.1 – RESULTADOS E CONCLUSÕES	109
7.2 – TRABALHOS CORRELATOS	111
7.3 – TRABALHOS FUTUROS	112
ANEXO I - ALGORITMO DE SUGESTÃO DE DIMENSÕES	114
ANEXO II - ALGORITMO DE MATCHING MELHORADO	117
REFERÊNCIAS BIBLIOGRÁFICAS	120

LISTA DE ILUSTRAÇÕES

Figura 2.1 – Arquitetura do REDIRIS.....	20
Figura 3.1 – Etapas da fase METODO2	24
Figura 3.2 – Banco de Dados Amostra Viva.....	28
Figura 3.3 – Problemas Típicos de Integração de Dados.....	31
Figura 3.4 – Etapas do Processo KDD [FAY96].....	32
Figura 3.5 – Taxonomia das Técnicas de Mineração de dados [AUR97].....	33
Figura 3.6 – Fórmula para Normalização segundo a Distribuição.....	35
Figura 3.7 – Fórmula para Normalização segundo a Amplitude.....	35
Figura 3.8 – Medição da Qualidade na Construção de um DW [CAM01].	39
Figura 3.9 – Critérios de Qualidade	40
Figura 3.10 – Critérios de qualidades e as tarefas realizadas em um DW [JAR00].....	41
Figura 3.11 – Dimensões de qualidade propostas por [JAR97a]	42
Figura 3.12 – Dados representados na Forma Gráfica [ALM01].....	47
Figura 3.13 – Distribuição de Frequência do campo INSCRICAO (chave primária).....	51
Figura 4.1 –Um ciclo na metodologia FastCube.....	54
Figura 4.2 – Fragmentação aplicada a uma tabela de várias colunas.	58
Figura 4.3 – Tela de análise e tratamento de dados do NeuralScorer.	61
Figura 4.4 – Um exemplo de um modelo de classes de metadados para tratamento de dados.	63
Figura 4.5 – Seleção dos fragmentos (atributos) relevantes para o Data Mart.	65
Figura 4.6 – Uma tabela de fato e dimensões com apenas um atributo	66
Figura 4.8 – Modelo com <i>surrogate key</i> e dimensões montadas.	69
Figura 4.9 – SQL Simplificado de carga de dimensões e fatos.....	70
Figura 5.1 – Um ciclo no REDIRIS.	71
Figura5.2 – Arquitetura básica da implementação do FastCube	72
Figura 5.3 – Principais pacotes do FastCube.....	73
Figura 5.4 – Diagrama de classes do pacote dataManipulation.....	74
Figura 5.5 – Diagrama de classes do pacote Storage.	75

Figura 5.6 – Principais pacotes de dados e metadados	76
Figura 5.8 –Principais cenários para o processo de qualidade dos dados.....	81
Figura 5.9 – Fluxo da aplicação de uma técnica no módulo Pré-Processador do REDIRIS.	82
Figura 5.10 – Aplicação de várias técnicas em cascata.	84
Figura 5.11 – Diagrama de classes de implementação das técnicas.....	86
Figura 5.12 – Tela inicial do protótipo FastCube.....	87
Figura 5.13 – Tela de criação aplicação.	88
Figura 5.14 – Tela de abertura de aplicação.....	88
Figura 5.15 – Tela de fragmentação da <i>TabelaAmostra</i>	89
Figura 5.16 – Tela de pré-processamento de dados.	89
Figura 5.17 – Tela de montagem de modelo dimensional.....	90
Figura 5.18 – Tela de sugestão de dimensão.....	91
Figura 6.1 – Modelo Lógico Relacional dos dados de entrada.....	96
Figura 6.2 – Modelo estrela do protótipo final do COVEST.	101
Figura6.3– Gráficos considerando a variável MOTIVO_CURSO	102
Figura 6.4 – Gráficos considerando a variável Renda Familiar	102
Figura 6.5 – Gráficos considerando a variável Renda Familiar	103
Figura 6.6 – Gráficos considerando a variável Renda Familiar	103
Figura 6.7 – Tela do SAGENT com gráfico de Aprovação X Cursos X Rede.	104

LISTA DE TABELAS

Tabela 3.1: Algumas tarefas de KDD e suas técnicas de mineração de dados [AUR97]	32
Tabela 3.2: Características e métricas geralmente usado na qualidade dos dados [HUF96]......	43
Tabela 3.3 – Dados apresentados na forma tabular [ALM01].....	46
Tabela 3.4 – Distribuição de frequência 1.....	48
Tabela 3.5 – Distribuição de frequência 2.....	48
Tabela 3.6 – Distribuição de frequência 3.....	48
Tabela 3.7 – Distribuição de frequência de estados.	49
Tabela 3.8 – Sugestão de tratamento 1.....	49
Tabela 3.9 – Sugestão de tratamento 2.....	49
Tabela 4.1 – Quadro comparativo de aspectos de construção de DW dos principais autores.	70
Tabela 6.1 – Atributos da tabela desnormalizada.....	100
Tabela 6.2 - Estatística geral de aprovação	105
Tabela 6.3 - Estatística de aprovação segundo sexo e opção	105
Tabela 6.4 - Estatística de aprovação segundo Ar-Condicionado, Computador e Internet	105
Tabela 6.5 - Estatística de aprovação segundo posse de computador	105
Tabela 6.6 - Estatística de aprovação segundo sexo.....	106
Tabela 6.7 - Estatística de aprovação segundo domínio de língua estrangeira	106
Tabela 6.8 - Estatística de aprovação segundo quantidade de vestibulares prestados	106
Tabela 6.9 - Conhecimento descoberto pela equipe diretamente no Data Mart.....	106
Tabela 6.10 - Estatística de aprovação segundo sexo e participação da renda familiar	107
Tabela 6.11 – Relação entre número de familiares e participação em ensino público e privado....	108

LISTA DE ABREVIATURAS E SIGLAS

DCBD	- Descoberta de Conhecimento em Base de Dados
DW	- Data Warehouse
DWQ	- Data Warehouse Quality
ETL	- Extraction, Transformation and Loading
KDD	- Knowledge Discovery in Databases
KNN	- K Nearest Neighbours
OLAP	- On-Line Analytical Process
REDIRIS	- Reuse Environment on Data Integration, Reuse and Quality in Information Systems
SAD	- Sistemas de Apoio à Decisão
SGBD	- Sistema Gerenciador de Banco de Dados
SQL	- Structured Query Language
UML	- Unified Modeling Language
XML	- eXtensible Markup Language

RESUMO

O volume de informações a ser trabalhado na tomada das decisões gerenciais supera largamente a capacidade do processamento humano, mecânico e dos sistemas transacionais atuais, exigindo ferramentas de apoio à decisão mais adequadas aos novos desafios gerenciais. Mesmo aplicando-se modelos de decisão tidos como adequados, uma grande parte das implementações de Sistemas de Informação não atingem os resultados esperados, o que levam muitos deles ao fracasso total ou parcial. Acredita-se que com obtenção de resultados rápidos se possa conseguir um maior envolvimento do usuário final, o que segundo os especialistas diminui bastante a possibilidade de fracasso.

Esse trabalho visa a utilizar técnicas de análise e exploração de dados na construção de soluções de Sistemas de Apoio à Decisão, em especial na construção de Data Warehouse(DW). Aproveita-se o conhecimento adquirido com a aplicação dessas técnicas, mostrando a sua importância nas diversas fases de sua construção de um DW. Propõe-se e implementa-se uma metodologia chamada FASTCUBE, que é baseada em um modelo de pré-processamento de dados. Ela incorpora de maneira rápida os metadados extraídos diretamente da massa de dados. Acelerar e sedimentar a compreensão do problema, sempre levando-se em consideração a qualidade dos dados, durante todas as suas fases é um dos pontos forte dessa metodologia. O seu objetivo final é acelerar o processo de visualização do modelo de decisão, através de um protótipo de modelo dimensional, com dados operacionais amostrados no início do processo e tratados durante o mesmo.

ABSTRACT

This effort sights to use analysis techniques and data exploration in the construction of solutions of Support Decision Systems, especially in the construction of Data Warehouse (DW). It is Utilized the knowledge acquired with these techniques application, showing its importance in the several stages of its construction of a DW. It is suggested and implemented a methodology called FASTCUBE, which is based on a pre-processing data model. It incorporates in fast way metadata extracted directly of the data mass. Accelerating and sediment the comprehension of the problem, always, considering the data quality, this is one of the strong points of this methodology during all the phases. Its final objective is accelerate the process of visualization of the decision model, through a dimensional model prototype, with operational data showed in the beginning of the process and treated during itself.