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

EPÍGRAFE

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 Profª.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

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

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 Analitical 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.