Uma Metodologia para Desenvolvimento de Data Warehouse

Sergio Luis Dill[1], Aline França de Abreu[2], Edson Luis Padoin[1], Martinho Luis Kelm[1]

[1] UNIJUI – Universidade Regional do Noroeste do Estado do Rio Grande do Sul
Ijuí –RS – Brasil.

[2] UFSC – Universidade Federal de Santa Catarina – Florianópolis – SC – Brasil.
{dill,padoin,martinho}@unijui.tche.br, aline@deps.ufsc.br

Resumo. *O projeto de data warehouse é uma tarefa complexa e abrangente cujo sucesso está estreitamente ligado ao entendimento das várias etapas que compõem o processo de construção de tais ambientes. Neste artigo apresenta-se uma metodologia para desenvolvimento de data warehouse cujo objetivo principal é a proposição de diretrizes que permitam guiar o projetista ao longo do processo. As principais vantagens desta proposta em relação as existentes na literatura são a sua abrangência, sua aplicabilidade prática e a possibilidade da utilização de uma ferramenta de desenvolvimento para dar suporte ao processo de desenvolvimento. Utiliza-se um estudo de caso para contextualizar o trabalho e facilitar o entendimento da metodologia proposta.*

Abstract. *A data warehouse project is a great and complex task whose success is closely related to the understanding of the several steps that compose the development process of such environments. In this job we show a methodology for data warehouse design whose main objective is a proposition of lines of direction that guide the designer during the whole process. The main advantages of this proposal in relation of others available in literature are its applicability, completeness and the ability to use a development tool during the whole process. A case study is used to show the methodology and to facilitate its understanding.*

1. Introdução

O ambiente de *Data Warehouse* (DW) surgiu como uma evolução dos ambientes de suporte a decisão, integrando dados de uma ou várias fontes. Sua crescente popularidade reflete a necessidade das empresas em obter informações analíticas derivadas dos seus sistemas transacionais. O ambiente de DW tem características diferentes do ambiente tradicional e é construído com o objetivo de suprir as necessidades de processamento analítico das organizações.

A criação de um ambiente de DW surgiu como uma alternativa viável cujo princípio está na criação de um banco de dados especializado capaz de manipular grande volume de informações com bom desempenho, melhorando a gerência, o controle e o acesso aos dados. A função do DW é tornar as informações corporativas, obtidas a partir de bancos de dados operacionais e de fontes de dados externas à

organização, acessíveis para entendimento e uso das áreas estratégicas de uma organização.

O projeto de DW é uma tarefa complexa envolvendo um conjunto de conceitos e tecnologias. O sucesso de um projeto de DW está estreitamente relacionado com o entendimento e domínio destes conceitos e tecnologias. A causa principal que resulta em falha e insucesso de um projeto de DW está relacionada à ausência de uma metodologia abrangente capaz de fornecer uma visão geral do processo envolvendo conceitos e tecnologias [Kelly 1997] e o propósito de uso do DW. Os projetos de DW têm mais chances de sucesso quando desenvolvidos através de uma metodologia consistente que identifique e guie o projetista durante as várias fases do projeto.

Conforme abordado em [Dill 2002], as propostas de metodologias existentes na literatura apresentam deficiências em aspectos importantes destacando:

- A metodologia descrita em [Golfarelli and Rizzi 1998] não é suportada por uma ferramenta de desenvolvimento. Isto torna o processo de construção extremamente trabalhoso elevando o tempo e custo do projeto. Adicionalmente, o autor propõem a utilização do modelo DFM (Dimensional Fact Model) o qual adiciona complexidade ao projeto.
- O trabalho descrito em [Herdem 2000] limita-se na apresentação de um esquema genérico (*framework*) apresentando os tópicos gerais que envolvem a construção de um DW. Sendo o projeto de DW uma tarefa complexa, uma metodologia deve descrever e detalhar cada uma das etapas.
- A proposta apresentada em [Moody and Kortnik 2000] preocupa-se em derivar esquemas dimensionais a partir da existência de um (único) modelo entidade/relacionamento (E/R) normalizado. Entretanto, a realidade das empresas nem sempre contempla este requisito e muitas vezes, vários bancos de dados são usados como fonte de dados para o DW. Além disso, a metodologia é incompleta, pois não considera: 1) os requisitos dos usuários; 2) os metadados; 3) a granularidade do DW; e 4) o projeto físico do DW.

O objetivo principal deste trabalho é elaborar uma metodologia consistente caracterizada pela sua aplicabilidade prática, suprindo as deficiências das metodologias avaliadas em [Dill 2002]. Em especial, concentramo-nos na clara identificação e descrição das várias fases do projeto aliada a possibilidade do processo todo ser suportado por uma ferramenta de desenvolvimento.

O artigo está dividido em quatro seções incluindo esta introdução. Na seção dois descrevemos o ambiente. Em seguida, na seção 3, esse ambiente será utilizado para apresentar e validando a metodologia e as suas etapas através de um estudo de caso. Por fim apresenta-se a conclusão do trabalho.

2. O Ambiente de DW

O aspecto fundamental na criação de um DW reside na separação dos dados do ambiente operacional para o ambiente de DW. A Figura 1 apresenta um ambiente típico de DW. Basicamente, pode-se dividir este ambiente em três componentes principais:

1) As fontes de dados: Os dados carregados para o DW são extraídos dos bancos de dados operacionais e fontes externas.

2) O DW: Os dados carregados para o DW passam pelo processo de extração, transformação e então armazenados em um formato apropriado (esquema estrela) que facilite o processamento analítico.

3) Os usuários: Os dados do DW são acessados pelos usuários através de ferramentas analíticas que possuem um conjunto de funcionalidades que facilitam o processo de exploração dos dados.

Um aspecto importante na construção de um ambiente de DW é a escolha da abordagem de desenvolvimento. A decisão de usar a estratégia *botton up* ou *top down* deve ser tomada com cuidado. Nesta proposta de metodologia adota-se a abordagem *top down* para a fase de definição do projeto (planejar o todo) e *botton up* para as demais fases que conjuntamente representam a fase de desenvolvimento do DW (implementar em partes). A vantagem desta abordagem é que o planejamento global resulta em um projeto mais consistente e integrado permitindo que o mesmo seja implementado aos poucos e assim, partes do DW sejam liberadas e utilizadas com maior rapidez.

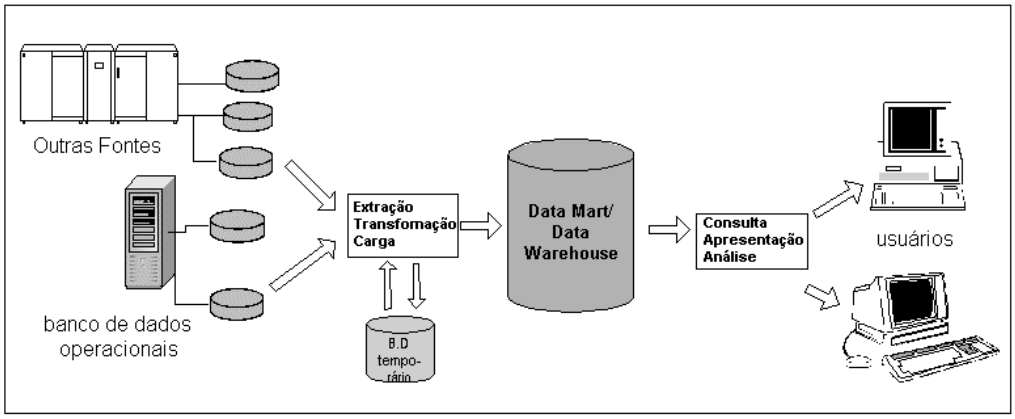

Figura 1: Um ambiente típico de DW (Fonte: Dill, 2002)

Neste trabalho, com o objetivo de apresentar e aplicar a metodologia para construir o ambiente mostrado na Figura 1, utiliza-se ferramentas de desenvolvimento de DW as quais auxiliam o projetista nas várias fases do projeto. Estas ferramentas estão divididas nas seguintes categorias:

- Servidor de banco de dados: IBM DB2 V7.2
- Ferramenta ETC: IBM DB2 *Warehouse Manager* V7.2
- Ferramenta OLAP: IBM DB2 OLAP *STARTER KIT* V7.2

Para viabilizar o estudo de caso, usa-se o banco de dados operacional da Universidade Regional do Noroeste do Estado do Rio Grande do Sul (Unijuí) limitando-se ao sistema de concurso vestibular.

3. Metodologia para desenvolvimento de DW

A metodologia que será desenvolvida neste trabalho estende e complementa o trabalho apresentado em [Herdem 2000] cujas etapas são muito familiares e utilizadas para o desenvolvimento dos tradicionais sistemas transacionais. Embora o desenvolvimento de um DW possua aspectos diferenciados com relação aos sistemas tradicionais muitas das lições aprendidas no desenvolvimento de sistemas SPT são de grande valia e devem ser utilizadas no projeto de DW. Esta metodologia foi escolhida pelo fato de considerar a experiência da equipe responsável pelo ambiente SPT considerado requisito fundamental para o sucesso de um projeto de DW.

A Figura 2 mostra as fases da metodologia que compõem o ciclo de desenvolvimento do DW. O principal aspecto a ser considerado é a natureza iterativa do desenvolvimento do DW característica que distingue o ciclo de vida de um projeto de DW de outros projetos de desenvolvimento e que permite rapidamente liberar partes do DW para o usuário enquanto outra parte pode estar sendo desenvolvida[Ballard et al. 2000]. A seguir descreve-se individualmente cada uma das fases da metodologia.

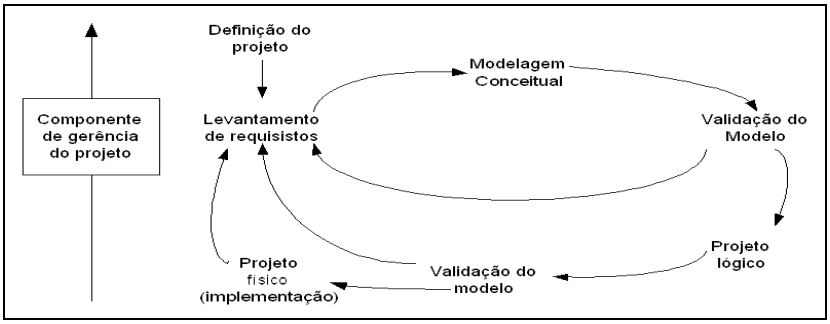

Figura 2: Fases da metodologia de desenvolvimento de DW [Fonte: Ballard et al. 2000]

3.1 – Gerência e definição do projeto

O componente de gerência do projeto tem a responsabilidade de estabelecer o plano geral do projeto. Este plano deve ser conhecido por todos os membros que farão parte da equipe de desenvolvimento do projeto. O plano deve estabelecer o prazo do projeto, os recursos disponíveis e principalmente a expectativa dos usuários com relação ao projeto. O gerente de projeto tem a responsabilidade de estabelecer as principais variáveis do projeto incluindo: 1) As funções que o DW irá disponibilizar; 2) Alocação de recursos (máquinas, ferramentas, pessoas); 3) Qualidade (definição de prazos não realísticos pode levar a equipe a seguir atalhos e comprometer a qualidade do DW*).*

Na definição do projeto estabelecem-se os objetivos maiores prevenindo assim as constantes mudanças que podem ocorrem durante as fases do ciclo de desenvolvimento à medida que novos requisitos são identificados. Contudo, deve-se ter como desafio a construção de um DW flexível e que tenha a habilidade de absorver as futuras expansões. Esta fase inclui também o entendimento dos conceitos e tecnologias relacionados ao ambiente de inserção do DW, sendo recomendado um planejamento prévio para determinar a escolha da arquitetura e infra-estrutura necessária para possibilitar o pleno desenvolvimento do DW.

3.2 – Modelagem Conceitual

Na modelagem conceitual de DW não basta apenas realizar o levantamento de requisitos dos usuários. Adicionalmente, as estruturas dos bancos de dados operacionais devem ser consideradas. Os requisitos dos usuários e as estruturas dos bancos de dados possuem influência estática e dinâmica, caracterizadas pelas possíveis alterações nos requisitos dos usuários e pela mudança na estrutura do banco de dados em questão [Bohnlein and Ende 1999]. Existem basicamente duas abordagens para obter os requisitos do DW. A primeira alternativa concentra-se mais diretamente no usuário. A segunda alternativa dá maior ênfase aos dados existentes nos sistemas da organização. Neste trabalho, a fase de levantamento de requisitos teve como base às informações do usuário e os dados existentes nos bancos de dados operacionais.

Em [Sapia 1998] encontra-se uma extensão ao modelo E/R para o paradigma multidimensional. Novos elementos gráficos são introduzidos para estender o modelo E/R conforme mostra a Figura 3. Neste trabalho, para a elaboração do modelo conceitual, adotou-se a referida representação gráfica.

A Figura 3 mostra o modelo conceitual elaborado a partir dos requisitos levantados. No centro do esquema encontra-se o fato vestibular que é composto por seis medidas: 1) **Vagas -** Corresponde ao número de vagas oferecidas; 2) **Inscritos -** Número de candidatos inscritos; 3) **Classificados -** Número de candidatos classificados no vestibular; 4) **Aprovados -** Número de candidatos aprovados no vestibular; 5) **Não aprovados -** Número de candidatos não aprovados no vestibular; 6) **Suplentes**: Número de candidatos suplentes.

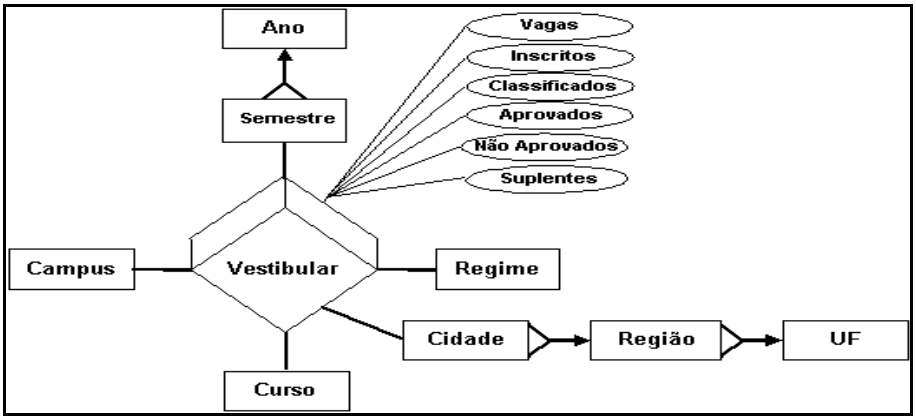

Figura 3: Modelo conceitual do sistema de vestibular

Além da tabela de fatos, o esquema conceitual possui cincos dimensões: 1) A dimensão **CAMPUS**: A Universidade oferece curso de graduação em vários campus; 2 A dimensão **REGIME**: Um curso pode pertencer ao Regime Regular (normal) ou Especial (período de férias, meses de janeiro, fevereiro e julho); 3) A dimensão **CURSO**: Os vários cursos oferecidos a cada vestibular; 4) A dimensão **TEMPO**: A cada ano são realizados dois vestibulares (Semestral); 5) A dimensão **CIDADE**: Tem a finalidade de realizar a estatística da origem dos candidatos do vestibular.

3.3 – Projeto lógico

Neste trabalho foi usado a técnica de modelagem dimensional de [Kimball 1998] para a criação do projeto lógico do DW. Esta técnica é caracterizada pela criação do esquema estrela a partir do esquema conceitual criado na fase anterior. Esta fase é inteiramente desenvolvida através da utilização de uma ferramenta que suporta a construção do esquema estrela. O Centro de DW do IBM DB2 guia o projetista através das várias etapas do projeto lógico conforme mostra a Figura 4.

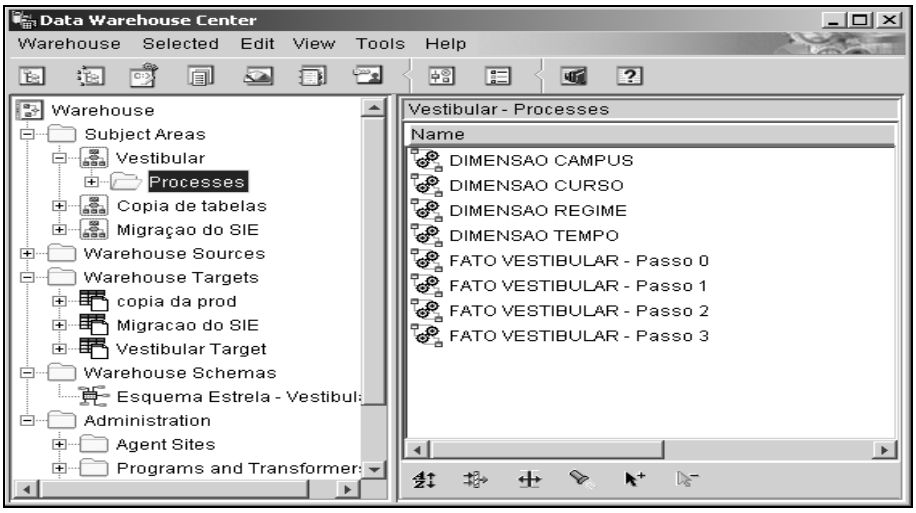

Figura 4: Centro de DW do IBM DB2.

A primeira etapa do projeto lógico do DW é a definição de um assunto(Subject Áreas). Um assunto compreende um conjunto de processos relacionados á uma área especifica do negócio. No presente estudo de caso, foi definido o assunto Vestibular. O objetivo principal de um assunto é a elaboração de um esquema de DW (Cubo de dados). Este esquema é construído gradativamente através dos processos que estão relacionados ao assunto. Um processo tem a finalidade de transformar os dados que armazenados nos sistemas fonte cuja origem dos dados pode derivar de várias bases de dados e podem estar armazenadas em sistemas diferentes.

Um exemplo de processo de transformação de dados é apresentado na Figura 5a. o qual transforma dados de um arquivo texto para uma tabela e será armazenada no banco de dados do DW e compreende a dimensão Tempo do esquema estrela resultante.

Através do uso do centro de DW, podem-se modelar complexos processos de transformação de dados. Este trabalho não tem o objetivo de mostrar as potencialidades (e limitações) da ferramenta. Apenas serão apresentados aqueles recursos utilizados na fase de criação do estudo de caso.

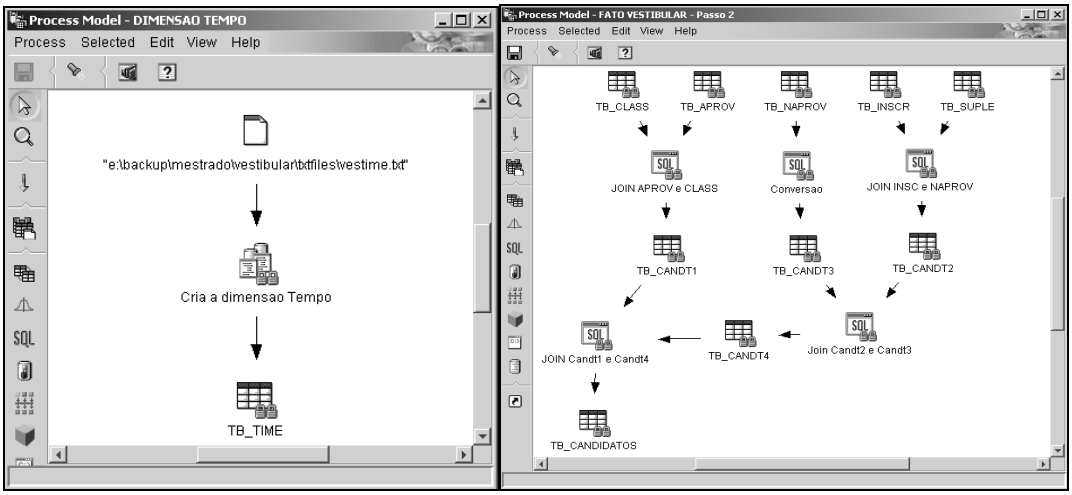

Figura 5a: Dimensão Tempo **Figura 5b: Fato Vestibular**

Ao final da execução de todos os processos de transformação dos dados fontes, obtêm-se então o esquema estrela resultante, conforme ilustra a Figura 6. O esquema compreende as seguintes tabelas: Uma tabela de fatos ao centro (DW.TB_FATVES) e quatro tabelas dimensionais correspondentes as dimensões Campus (DW.TB_CAM), Regime (DW.TB_REG), Curso (DW.TB_CUR), e Tempo (DW.TB_TIME).

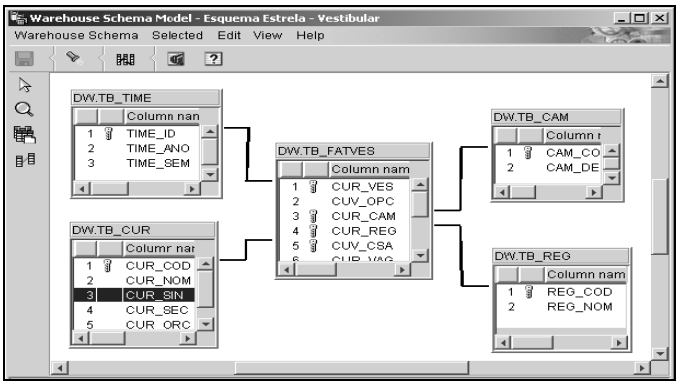

Figura.6: Esquema estrela do sistema de vestibular

3.4 – Projeto Físico

Os principais aspectos a serem considerados no projeto físico do DW são [13]: 1) Indexação; 2) Materialização de visões; 3) Particionamento, paralelismo; 4) Nível de redundância dos dados; 5) Sintonia dos parâmetros do banco de dados.

A sintonia do banco de dados é fundamental no ambiente de DW dado que a natureza da carga é diferente do ambiente transacional. Os ambientes OLTP são configurados para realizar as transações dos vários usuários simultâneos no menor tempo possível. Os parâmetros de configuração ajustados são responsáveis pela melhora do desempenho em 20 a 25% [Hayes and Gunning 2002]. Os 75% restantes derivam de ajustes nas instruções de consulta. Isso envolve alterações no projeto físico do banco de dados, disponibilidade e características dos índices, replicação e particionamento de tabelas.

3.5 – Outros aspectos importantes da metodologia

Outros aspectos importantes que devem ser considerados no projeto de DW são: 1) Metadados; 2) Granularidade do DW e 3) Atualização do DW. Os metadados mantêm informações sobre o conteúdo que está armazenado no DW e são elaborados gradativamente ao longo de todo o processo de desenvolvimento do DW. Através da exploração dos metadados, os usuários podem encontrar as tabelas que originaram os dados do DW. A granularidade do DW registra que nível de detalhe os dados estarão disponíveis para a análise do usuário, isto é, determina a sua dimensionalidade possuindo influência direta no tamanho e desempenho do banco de dados. A escolha de um nível de granularidade inadequada pode comprometer e até inviabilizar o uso do DW. Por último, a etapa de atualização de dados deve ser suportada pela ferramenta de desenvolvimento DW a qual deve suportar as seguintes atividades:

- Automação do processo de extração, conversão e carga dos dados;
- Definição da periodicidade da atualização;
- Possibilidade de integração com outras ferramentas.

4. Conclusão

O objetivo deste trabalho foi a proposição de uma metodologia de desenvolvimento de DW e a avaliação da sua aplicabilidade através de um estudo de caso. A metodologia é dividida em quatro etapas principais: Gerência e definição do projeto, modelagem conceitual, projeto lógico e projeto físico. Conclui-se que a proposta elaborada constitui um avanço em relação as anteriores, pois apresenta uma sistemática mais apropriada a qual adere à realidade dos sistemas existentes nas empresas. Também, valoriza a experiência que a equipe possui no desenvolvimento de sistemas transacionais, pois as fases que compõem a metodologia já são largamente utilizadas no desenvolvimento de sistemas OLTP reduzindo significativamente a complexidade do processo. Outro ponto positivo da metodologia é a possibilidade do processo todo ser suportado por uma ferramenta de desenvolvimento a qual aumenta a produtividade, simplificando e automatizando tarefas complexas e trabalhosas comuns em sistemas de DW.

Referências

Kelly, S., The Data Warehouse Toolkit, John Wiley & Sons Inc., 1997.

Dill, Sergio Luis. Uma Metodologia para Desenvolvimento de DW e Estudo de Caso, Dissertação de Mestrado, UFSC, Florianópolis, 2002.

Golfarelli, M. and Rizzi, S. “A methodological framework for DW Design”. DOLAP 98 Washington, D.C., USA.

Herdem, O. (2000) “A Design Methodology for DWs”. Oldemburg Research and Development Institute for Computer Science Tools and Systems (OFFIS). Oldenburg, Germany.

Moody, D. L. and Kortnik, M. A.R. “From Enterprise Models to Dimensional Models: A Methodology for DW and Data mart Design”. (DMDW´2000).

Ballard, C.; Herreman D. and Schau D.;et al. Data Modeling Techniques for Data Warehousing. IBM – ITSO redbooks, 1998.

Bohnlein, M. and Ende A. Deriving Initial DW Structures from the Conceptual Data Models of the Underlying Operational Information Systems. Ulbrich-vom. Kansas City Mo USA, 1999.

Sapia, C., Blaskhka, M., Hofling, G. and Dinter, B. Extending the E/R Model for the Multidimensional Paradigm. Proc of International Workshop on DW and Data Mining, November 1998.

Kimball, R., Data Warehouse Toolkit, Makron Books, 1998.

Hayes, S and Gunning, P. “Tunning Up for OLTP and data warehousing. DB2 Magazine.Vol 7 Num 3”, 2002.