FERRAMENTA DE CONSTRUÇÃO DE DATA WAREHOUSE

Maurício Capobianco Lopes[1], Percio Alexandre de Oliveira[2]

[1]Departamento de Sistemas e Computação - Universidade Regional de Blumenau (FURB)
Rua Braz Wanka, 238 – 89.035-260 – Blumenau, SC – Brasil
mclopes@furb.br, percio@inf.furb.br

RESUMO: *O Data Warehouse é uma solução que procura de maneira flexível e eficiente tratar grandes volumes de dados e obter informações que auxiliem no processo para tomada de decisão. Assim, este artigo apresenta uma ferramenta com foco a usuários e projetistas de data warehouse, visando reduzir custos no processo de construção deste ambiente. A ferramenta foi desenvolvida em Java garantindo a portabilidade de seu sistema que implementa as principais fases de um projeto de data warehouse: extração, transformação e carga dos dados; visualização, análise e tratamento das informações.*

Introdução

A crescente competição em mercados cada vez mais dinâmicos está levando as empresas a tomarem decisões mais rapidamente. Sendo assim a informação tornou-se o bem mais valioso dentro das instituições. Os administradores tomam suas decisões com base na análise de dados objetivos, sintetizados e confiáveis acima de tudo, sempre com o intuito maior de melhorar e aperfeiçoar processos internos. É dentro deste cenário que hoje se torna imprescindível a utilização de recursos computacionais para levantar e analisar as informações necessárias. Uma das principais ferramentas que constitui a nova geração de Sistemas de Apoio a Decisão (SAD) é o Data Warehouse (DW), um banco de dados específico para propósitos gerenciais e estratégicos (DW BRASIL, 2005).

Para Kimball (1998 apud COME, 2001, p. 2), DW é o lugar onde as pessoas podem acessar seus dados. A abordagem de Ralph Kimball veio com um estilo mais simples e incremental, baseado na metodologia estrela que aponta para *Data Marts* (DM) separados, que deverão ser integrados na medida da sua evolução (BARBIERI, 2001). Já Wang (1998 apud COME, 2001, p. 2) tem uma definição um pouco mais elaborada quando diz que DW é o processo pelo qual os dados relacionados de vários sistemas operacionais são fundidos para proporcionar uma única e integrada visão de informação de negócios que abrange todas as divisões da empresa.

Assim, este trabalho apresenta uma ferramenta de DW para auxiliar seus usuários no processo de transformação de dados operacionais em informações gerenciais, viabilizando consultas em diversos níveis de detalhe. A ferramenta é totalmente executável em ambiente web, sendo, desta forma, acessível através dos principais navegadores hoje disponíveis no mercado. Trata-se, portanto, de uma ferramenta genérica para a construção e implantação de um DW, sem perder de vista sua

usabilidade, dando suporte tanto ao projetista do DW quanto ao seu usuário. Todos os detalhes de especificação e desenvolvimento desta ferramenta estão disponíveis no trabalho de Oliveira (2007).

DATA WAREHOUSE

O termo Data Warehouse significa armazém de dados. É definido como um ambiente que provê informações de suporte à decisão que, no ambiente operacional, se tornariam difíceis de serem obtidas. Em outras palavras, pode ser comparado como um banco de dados especial, estruturado de forma a facilitar o processamento para análise dos dados.

O conceito de DW surgiu da necessidade de integrar dados corporativos espalhados em diferentes máquinas e sistemas operacionais, para torná-los acessíveis a todos os usuários dos níveis decisórios (NAVARRO, 1996). Entretanto, essa integração deve ser feita com uma seleção cuidadosa e otimizada dos dados já que a prioridade na utilização do ambiente do DW é o processamento de consultas e não o processamento de transações. A Figura 1 ilustra toda a estrutura interna que o ambiente de DW representa.

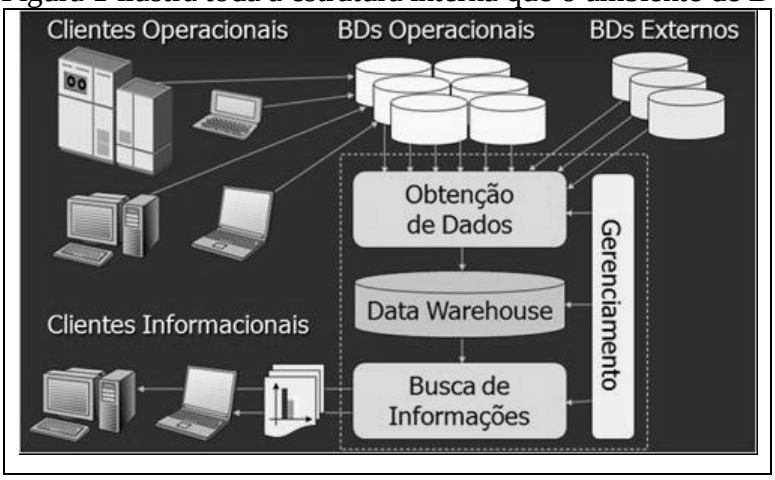


Figura 1 – Estrutura interna do DW

Um DW exige a criação de metadados que define as principais informações de um projeto de DW, além de sua documentação (VIEIRA, 2000). De acordo com Vieira (2000) algumas informações que o metadados deve conter são: a estrutura dos dados segundo a visão do programador e dos analistas de SAD; a origem das fontes de dados que alimentam o DW; a transformação dos dados ocorrida no processo de migração para o DW; o modelo de dados e seu relacionamento com o DW; o histórico das extrações de dados; as informações sobre as consultas e relatórios; acesso e segurança e os indicadores de qualidade de dados.

Uma das técnicas utilizadas para a criação do projeto lógico de um DW é a da modelagem dimensional. Esta técnica é caracterizada pela criação do esquema estrela a partir do esquema conceitual criado na fase de análise do DW. Para Kimball (1997 apud COME, 2001, p. 51) modelagem dimensional é uma técnica utilizada para a definição do projeto lógico de um DW. Três conceitos básicos são importantes nesta modelagem: tabelas fatos ou cubos de decisão que representam as transações de negócios, as dimensões que são os diferentes tipos de visões que os usuários irão utilizar para analisar as métricas e os indicadores ou métricas que podem ser definidos como os

atributos numéricos de um fato representando o comportamento de um negócio para as dimensões.

Outro conceito importante em um projeto de DW é o processo de extração, transformação e carga (ETC) que é o mais trabalhoso na construção de um DW. Durante essa etapa é importante ter uma eficiente integração de dados já que os mesmos podem vir de múltiplas fontes. Sua transformação deve ser feita de forma a gerar informações consistentes e de qualidade. Essa etapa é caracterizada por ser uma das mais críticas já que uma informação carregada erroneamente trará conseqüências imprevisíveis nas fases posteriores (SILVA, 2005, p. 19).

A FERRAMENTA

A ferramenta para a construção de um DW conta com dois atores: o usuário de consultas e o administrador projetista. Neste trabalho, a ênfase principal é com as funções disponibilizadas ao administrador, uma vez que ao usuário caberá apenas a tarefa de efetuar as consultas.

As operações realizadas pelo administrador podem ser divididas nos seguintes processos: montagem do projeto de DW, ETC, consultas, metadados e recursos adicionais da ferramenta.

Para as operações de montagem de projeto destacam-se os seguintes casos de uso:

a) **cadastrar Data Warehouse:** cria um novo projeto de DW baseado no modelo dimensional estrela;
b) **cadastrar dimensão:** grava as definições referentes a uma dimensão bem como seus atributos e chave primária;
c) **cadastrar cubo:** grava as definições referentes a um cubo de decisão bem como seus indicadores e dimensões relacionadas.

Para o processo de ETC destacam-se:

a) **cadastrar conexão:** cria uma nova conexão com um banco de dados que será disponibilizado para extração de dados para as dimensões e cubos do DW;
b) **cadastrar fonte de dados:** cria uma ou múltiplas fonte de dados através das quais irá se fazer a extração, transformação e carga dos dados para as dimensões ou cubos de um DW.

Para as operações de consultas destaca-se:

a) **cadastrar consulta:** define consultas gerenciais baseadas na modelagem dimensional do cubo de decisão;
b) **visualizar e configurar consultas**: acesso e configuração sobre as consultas cadastradas.

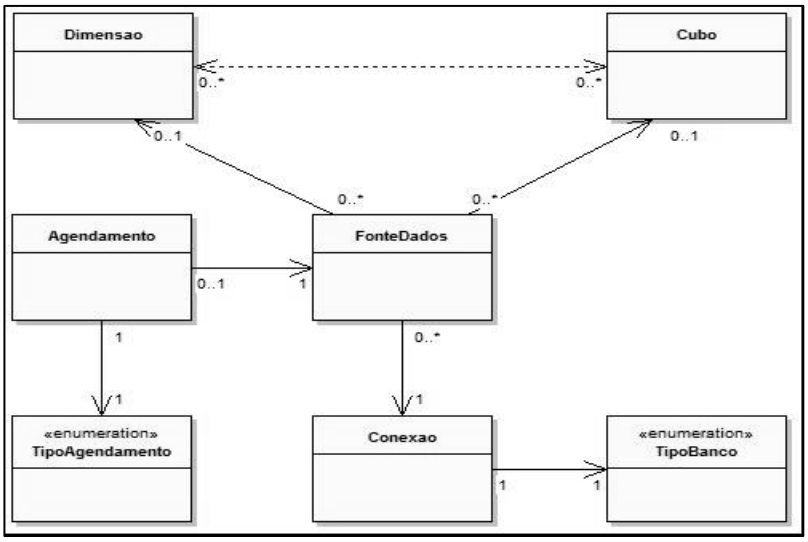
Outros recursos adicionais da ferramenta são:

a) **exportar metadados: e**xporta todas as definições referentes a um projeto de DW em padrão XML,
b) **importar metadados**: importa para o sistema um novo projeto de DW gerado em XML;
c) **visualizar agendamento:** apresenta ao administrador todos os agendamentos de fontes do dia corrente que ainda estão em aberto para processamento;
d) **visualizar log de mensagens:** mostra as principais ocorrências dentro do sistema como informações de importação, erro e tratamento de exceções;

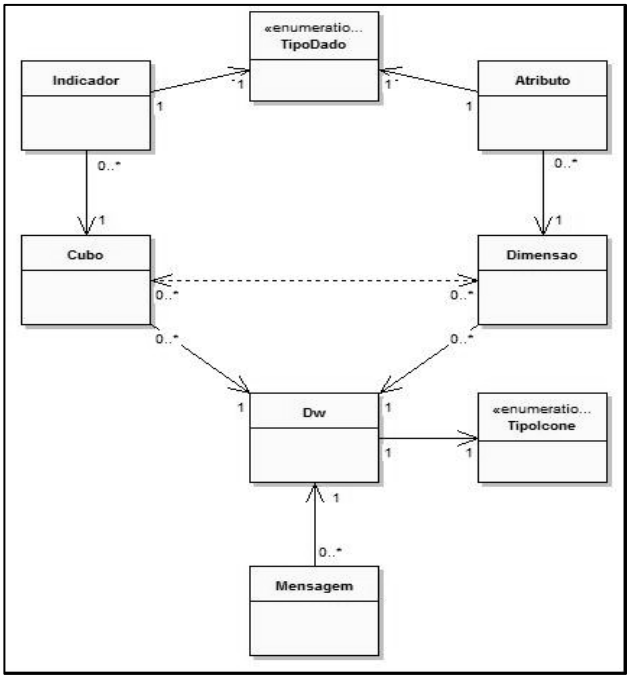
e) **limpar Data Warehouse:** processa limpeza de dados e do conteúdo dos projetos de DW do sistema;
f) **cadastrar usuário:** cria novos usuários para acesso ao sistema.

O diagrama de classes da ferramenta apresentado na Figura 2, foi dividido em três partes: (a) ETC que apresenta o modelo necessário para o processo de extração, transformação e carga dos dados dentro do DW; (b) PROJETO que apresenta o modelo necessário para o processo de construção de um projeto de DW; (c) CONSULTA que apresenta o modelo necessário para a visualização e configuração das consultas do DW. Em cada processo existem classes comuns que são utilizadas e que possuem funcionalidades específicas.

ETC



CONSULTA



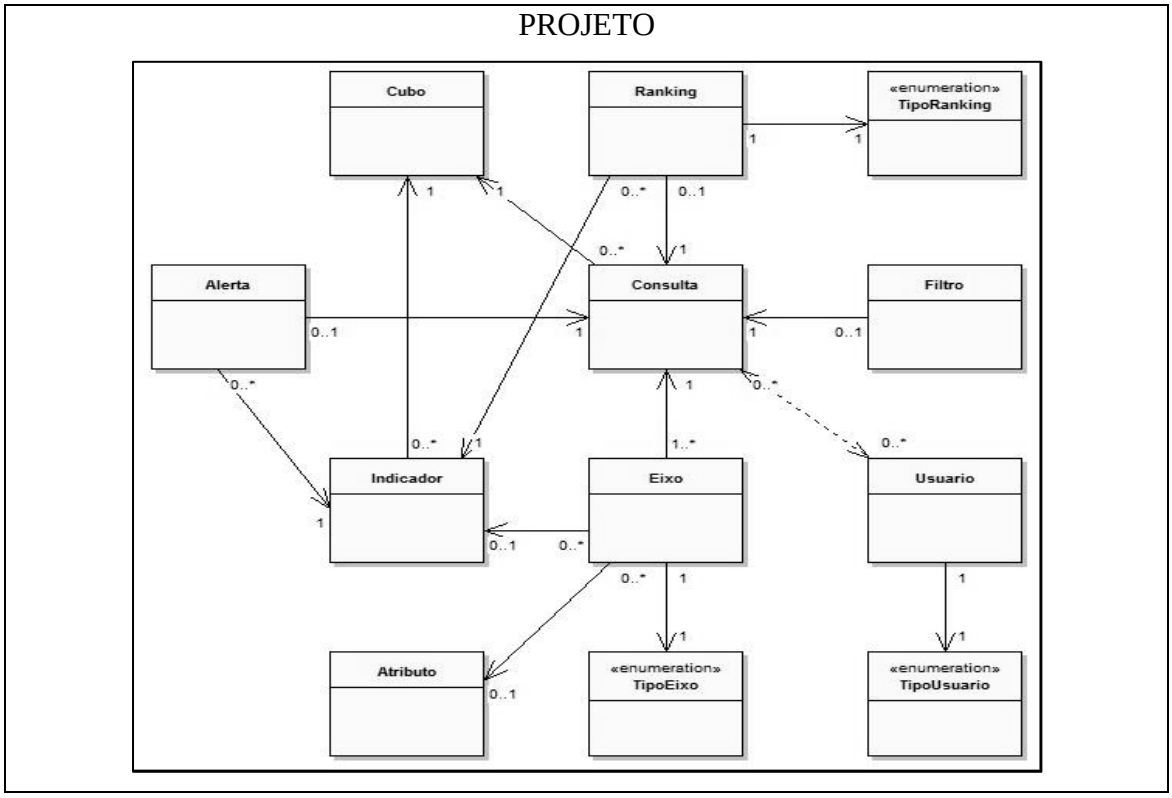


Figura 2 - Diagrama de classes dos pacotes principais do sistema

O sistema ainda apresenta outros pacotes menores que apresentam diagramas mais simples e que possuem finalidades específicas (OLIVEIRA, 2007).

A ferramenta foi desenvolvida na plataforma Java. O sistema foi compilado utilizando o J2SE 1.5 e roda em um servidor que implementa a especificação J2EE 1.4 ou superior. Para o desenvolvimento de aplicações web foi utilizado o *Integrated Development Environment* (IDE) Eclipse 3.2.1 acrescido do *plugin* MyEclipse 5.1.1 que utiliza *servlets* com interfaces JSP. O servidor de aplicações utilizado foi o Apache Tomcat 5.5.23. Para a implantação do AJAX foram utilizadas implementações javascripts com *grids*, para os quais utilizou-se a biblioteca de *scripts* da Zapatec, que possui diversas modelagens para tabela de dados. O banco de dados utilizado foi o MySQL 5.0 com interfaces de conexão JDBC. As tabelas do modelo objeto relacional estão descrita no trabalho de Oliveira (2007).

A seguir será apresentada a operacionalidade do sistema assim como suas principais interfaces e operações. Para ilustração do funcionamento de todo o processo de construção de um DW tomou-se como estudo de caso o faturamento de uma empresa fictícia com as seguintes definições de projeto:

a) manter o histórico de vendas da empresa;
b) consultar a soma das vendas em valor total da nota fiscal por data (ano e mês) e clientes (estado);
c) consultar a média de vendas em valor total da nota fiscal por representante (nome) e clientes (nome), alertando o usuário onde a média das vendas foram abaixo de duzentos reais na cor vermelha e acima na cor verde;
d) consultar os vinte produtos mais vendidos em quantidade no ano de 2007.

Para apresentar este estudo de caso foi criado dentro da ferramenta um projeto de DW utilizando o usuário padrão DWADMIN. A Figura 3 ilustra a tela de login de usuário e a tela onde o administrador pode cadastrar um novo projeto de DW ou já selecionar projetos existentes.

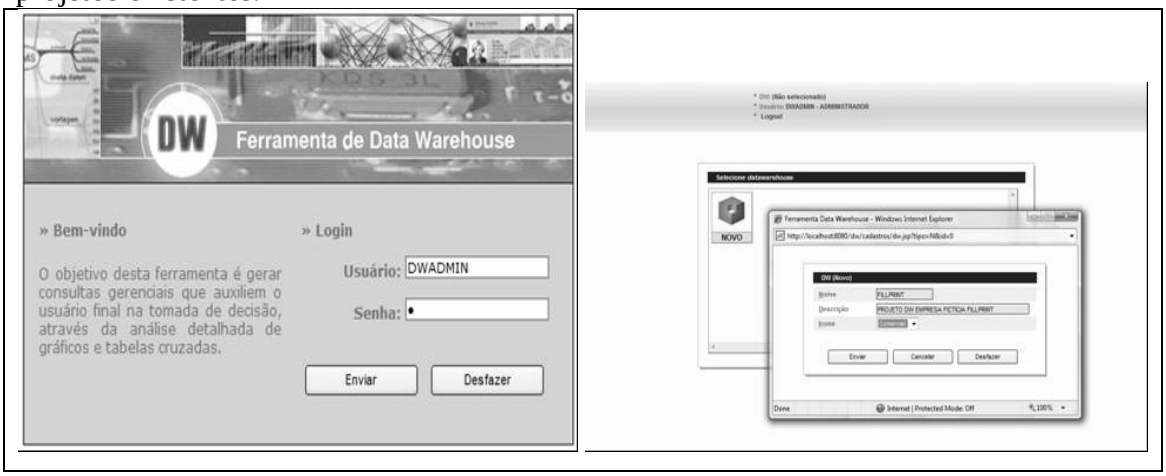


Figura 3 – Tela de login

Com o projeto criado, o administrador pode dar seqüência à montagem do DW. Assim, inicialmente é necessário cadastrar uma conexão com o banco de origem dos dados de vendas da empresa. Tendo uma conexão estabelecida com uma base de dados, o projeto de DW pode começar a ser definido através do cadastro das dimensões e atributos, cubos e indicadores e suas fontes de dados respectivamente. Para o estudo de caso foram feitas as seguintes definições: (a) dimensão cliente: atributos nome e estado; (b) dimensão data: atributos ano e mês; (c) dimensão representante: atributo nome; (d) dimensão produto: atributos nome e tipo; (e) cubo venda: indicadores valor total e quantidade.

A Figura 4 ilustra o cadastro de uma dimensão e de um atributo. Após todos os atributos estarem definidos para a dimensão, é realizada a definição da chave primária da dimensão ilustrada pela Figura 5.

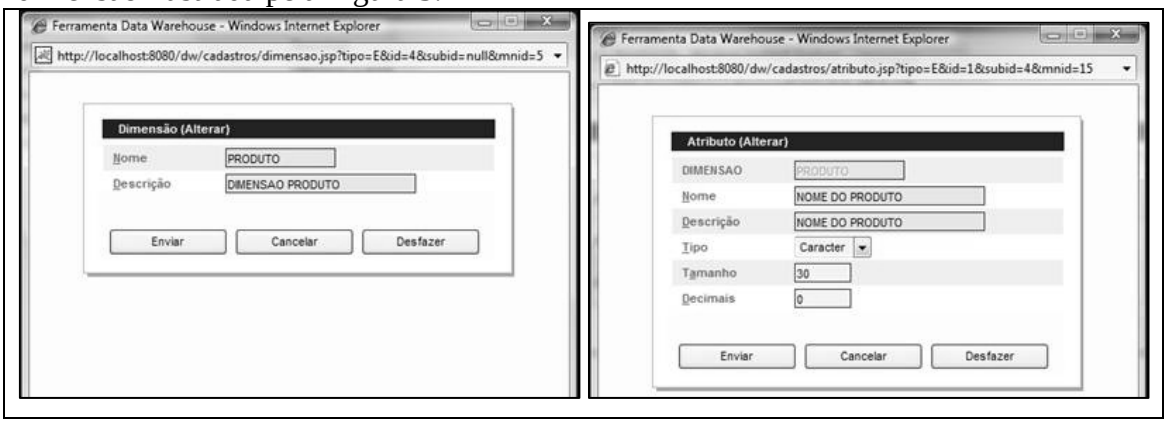


Figura 4 – Tela de cadastro de dimensão

Figura 5 – Tela de definição de chave primária da dimensão

Após o administrador cadastrar todas as dimensões a serem utilizadas no cubo de decisão, foi feito o cadastro do mesmo e definidos seus indicadores. Para o cubo também é necessário a definição das dimensões utilizadas.

O cadastro do cubo de decisão determina que o modelo projetado para o DW esteja pronto para ser carregado com os dados. Desta forma, o processo de ETC é realizado através das fontes de dados que cada dimensão e cubo possuem, pelo roteiro desta fonte de dados que realiza o mapeamento da origem do dado com o projeto definido no DW e, por último, pela importação dos dados que podem ainda ser agendados e processados periodicamente. A Figura 6 ilustra o cadastro de uma fonte de dados, neste caso para o cubo de decisão VENDA e o roteiro desta mesma fonte de dados. Neste caso, por ser um cubo, é necessário mapear tanto os indicadores como as chaves primárias de cada dimensão relacionada. Para as dimensões apenas os atributos necessitam ser roteirizados.

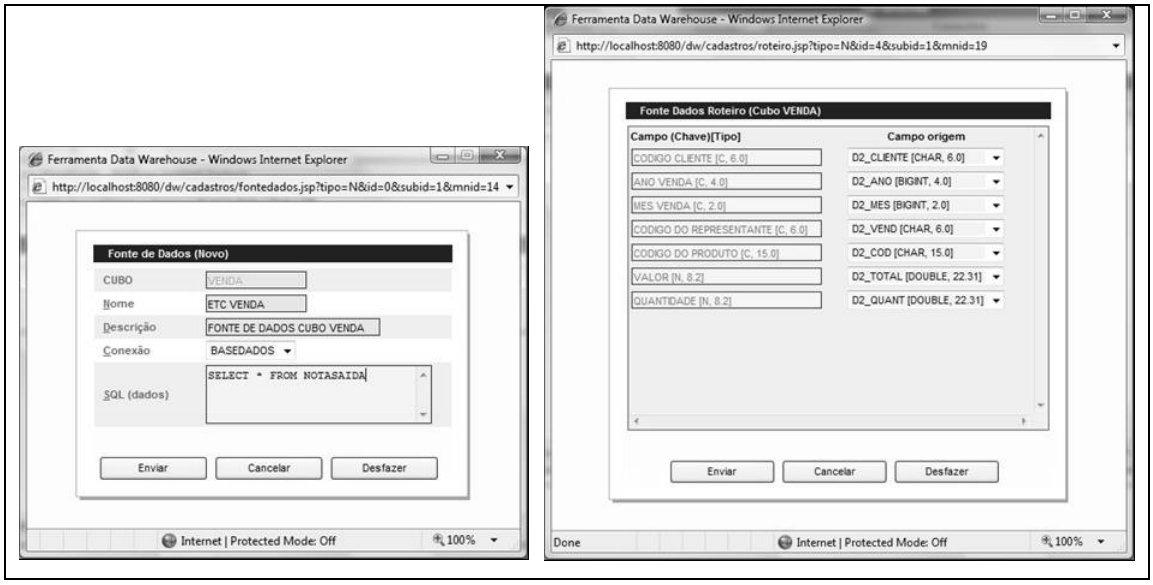


Figura 6 – Tela de cadastro de fonte de dados e tela de roteiro de uma fonte de dados

O próximo passo é a montagem das consultas. O administrador deve cadastrar a consulta e definir os seus eixos. A Figura 7 ilustra o cadastro de uma consulta e a

definição de um novo eixo para esta consulta. O eixo de indicadores possui a definição das funções de agregação representada na Figura 8, que também apresenta a consulta de vendas por estado, onde o administrador utiliza o recurso de drill-down.

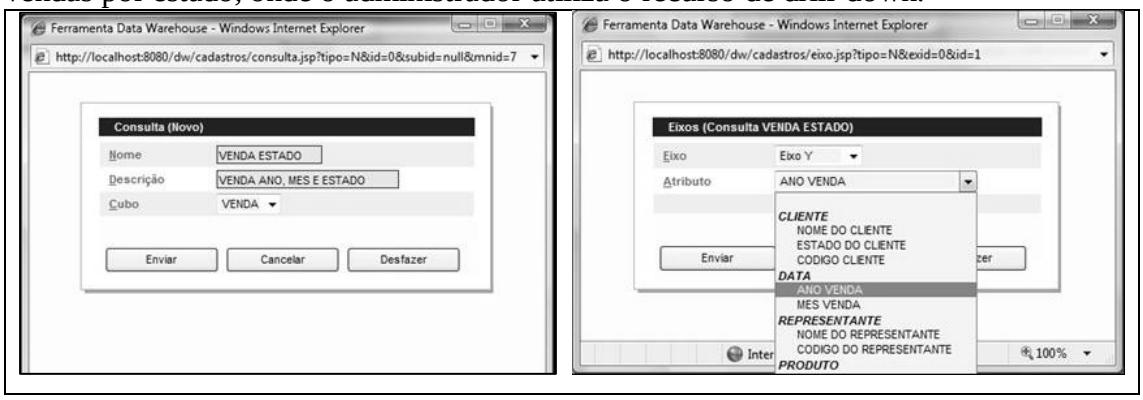


Figura 7 – Tela de cadastro de consultas e Tela de cadastro de eixos da consulta

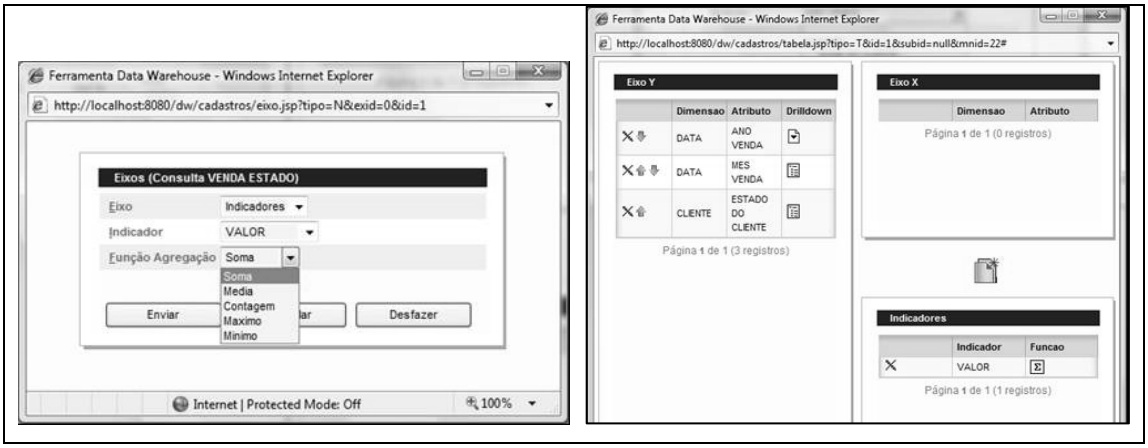


Figura 8 – Tela de cadastro de eixos de indicadores da consulta e Tela de definição da consulta com drill-down

Para visualizar o resultado das consultas é utilizado um perfil de usuário onde apenas a visualização das consultas, a edição de filtro, alerta e ranking estão disponíveis.

As Figura 9, Figura 10 e Figura 11 mostram um exemplo do resultado da consulta de vendas por ano, mês e estado do cliente.

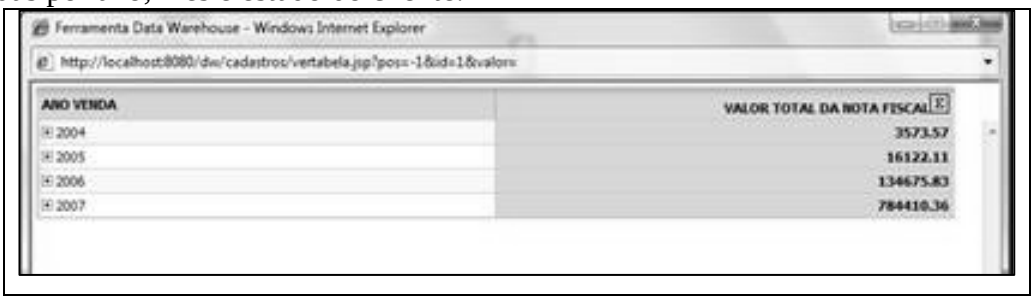


Figura 9 – Tela de consulta de vendas por ano com drill-down

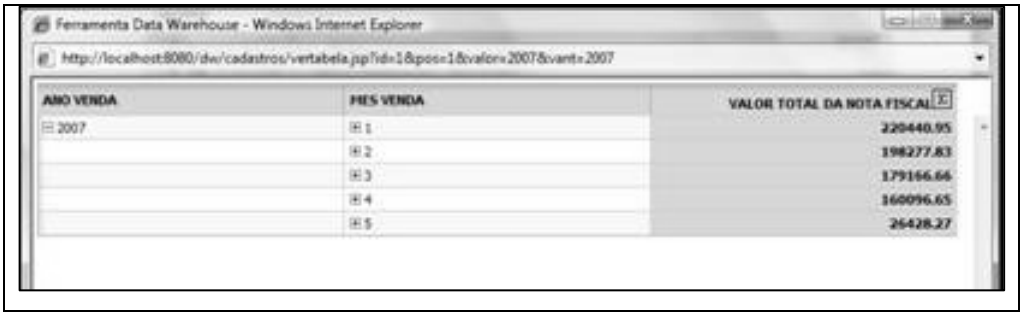


Figura 10 – Tela de consulta de vendas por ano e mês com drill-down

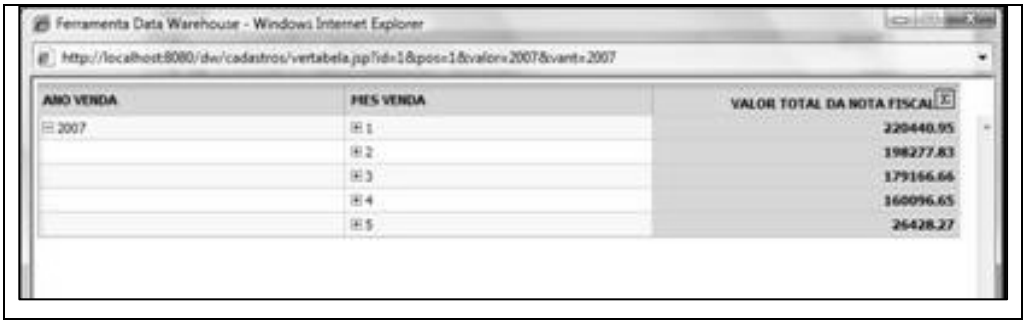


Figura 11 – Tela de consulta de vendas por ano, mês e estado com drill-down

Para montar a consulta de vendas por representantes e clientes é utilizado o recurso de alerta para destacar os resultados obtidos.

A Figura 12 mostra exemplos do resultado da consulta de vendas por representante e por clientes.

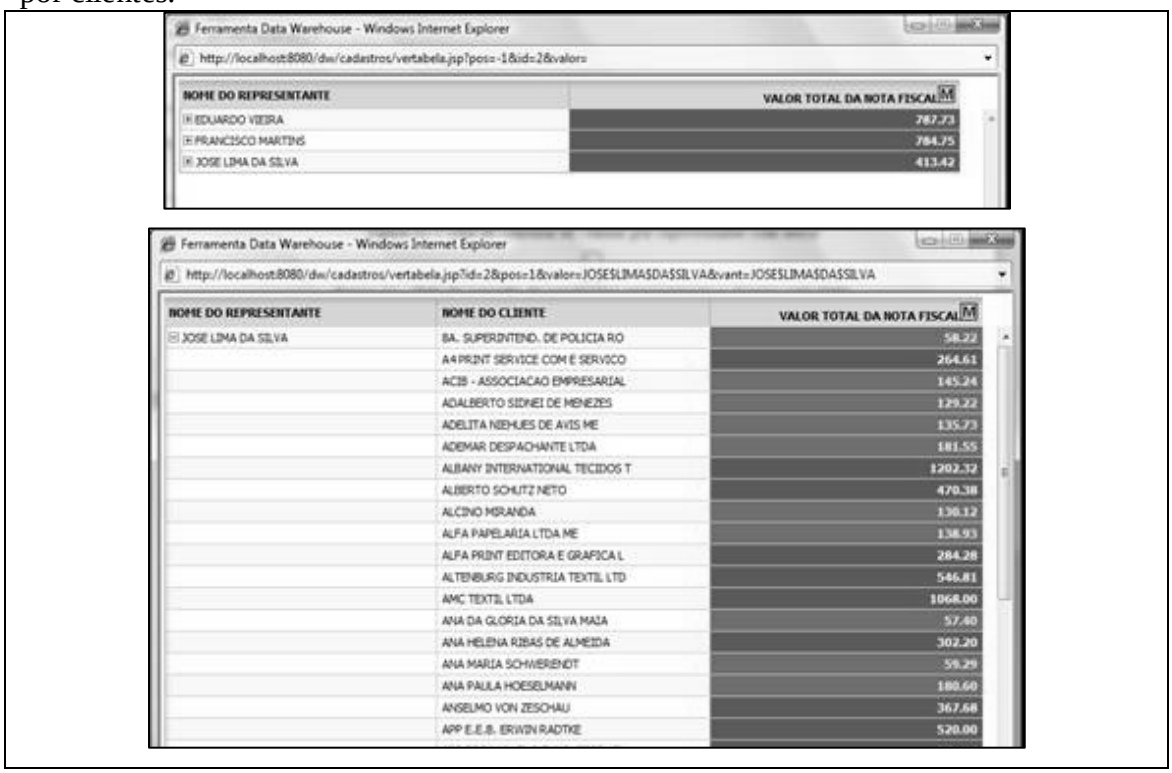
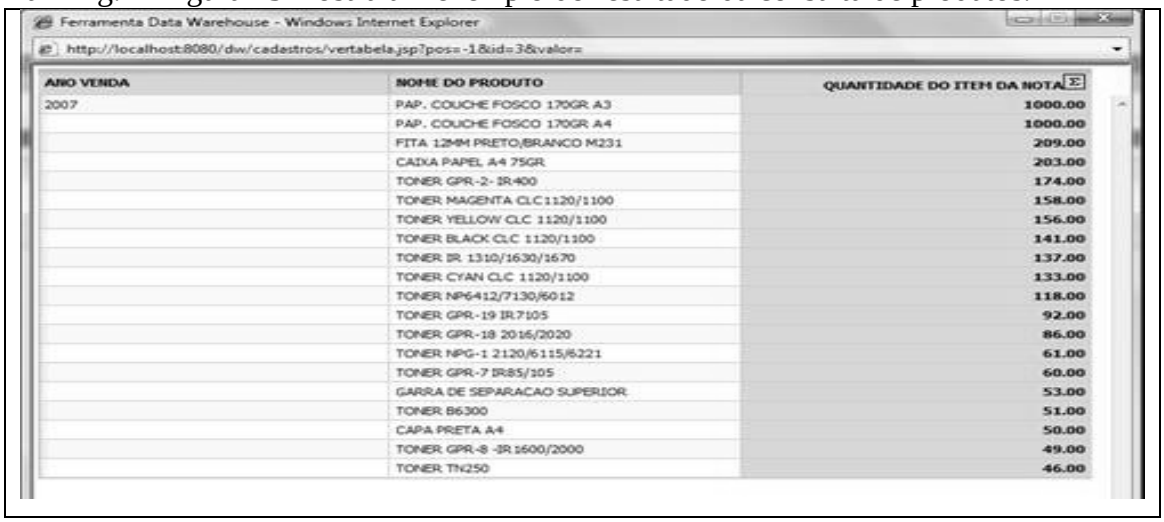


Figura 12 – Tela de consulta de vendas por representante com alerta e Tela de consulta de vendas por representante e clientes com alerta

Para montar a consulta de vendas por produto são utilizados os recursos de filtro e ranking. A Figura 13 mostra um exemplo do resultado da consulta de produtos.



ANO VENDA	NOME DO PRODUTO	QUANTIDADE DO ITEM DA NOTA
2007	PAP. COUCHE FOSCO 170GR A3	1000.00
	PAP. COUCHE FOSCO 170GR A4	1000.00
	FITA 12MM PRETO/BRANCO M231	209.00
	CAIXA PAPEL A4 75GR	203.00
	TONER GPR-2- IR400	174.00
	TONER MAGENTA CLC1120/1100	158.00
	TONER YELLOW CLC 1120/1100	156.00
	TONER BLACK CLC 1120/1100	141.00
	TONER IR 1310/1630/1670	137.00
	TONER CYAN CLC 1120/1100	133.00
	TONER NP6412/7130/6012	118.00
	TONER GPR-19 IR7105	92.00
	TONER GPR-18 2016/2020	86.00
	TONER NPG-1 2120/6115/6221	61.00
	TONER GPR-7 IR85/105	60.00
	GARRA DE SEPARACAO SUPERIOR	53.00
	TONER B6300	51.00
	CAPA PRETA A4	50.00
	TONER GPR-8 -IR1600/2000	49.00
	TONER TN250	46.00

Figura 13 – Tela de consulta de vendas por produto com filtro e ranking

RESULTADOS E DISCUSSÃO

A ferramenta de DW apresentada neste artigo foi desenvolvida com o propósito final de garantir desempenho e usabilidade funcional nas consultas gerenciais que se utilizam de grande volume de dados gravados em bases históricas e banco de dados transacionais.
Para avaliar o resultado obtido foi utilizado o estudo de caso citado neste trabalho sendo feitas comparações entre as consultas processadas diretamente no banco transacional e um banco de dados gerado pela ferramenta de DW. Os pontos avaliados são: tempo de processamento de um consulta, total de tabelas consultadas e quantidade de registros processados. A Tabela 1 mostra os resultados obtidos para a consulta de vendas por ano, mês e estado do cliente em cada nível de detalhamento dos dados. A Tabela 2 mostra os resultados obtidos para a consulta de vendas por representante e clientes e a

Tabela 3 mostra os resultados obtidos para a consulta de venda por ano e produto.

Tabela 1 – Resultados da consulta venda por data e estado

Venda ano, mês e estado						
Nível	Tempo processamento (s)		Total de tabelas		Total de registros	
	BD	DW	BD	DW	BD	DW
1	0,0185	0,0064	1	2	3518	2323 x 1
2	0,0192	0,0066	1	2	3518	2323 x 1
3	1,2249	0,0164	2	3	982 x 3518	27 x 145 x 1

Tabela 2 – Resultados da consulta venda por representante e cliente

Venda representante e cliente						
Nível	Tempo processamento (s)		Total de tabelas		Total de registros	
	BD	DW	BD	DW	BD	DW
1	0,0414	0,0119	2	2	3 x 3518	4 x 232
2	4,3536	0,0145	3	3	3 x 3518 x 982	4 x 232 x 1

Tabela 3 – Resultados da consulta venda por produto

Venda produto						
Nível	Tempo processamento (s)		Total de tabelas		Total de registros	
	BD	DW	BD	DW	BD	DW
1	18,1847	0,0195	2	3	3518 x 15969	27 x 145 x 1

O processo de ETC das principais fontes de dados envolvidas neste projeto é o responsável por este desempenho favorável ao DW, tornando as consultas mais eficientes e rápidas.

A Tabela 4 mostra o tempo exigido pelo processo de ETC. A otimização deste processo através de consultas SQL, bancos de dados indexados e relacionamentos corretos é que garantem um processo de ETC mais eficiente.

Tabela 4 – Resultados do processo de ETC

	Tempo processamento (min.)	Total de registros
Dimensão Cliente	1m 51s	970
Dimensão Produto	26min 12s	11594
Dimensão Data	5min 23s	3496
Dimensão Representante	< 1s	3
Cubo Venda	6min 45s	4359

CONSIDERAÇÕES FINAIS

Com o propósito principal de obter informações gerenciais detalhadas e resumidas provenientes de banco de dados históricos e transacionais, a ferramenta apresentada atinge os objetivos a que se propõe demonstrando ser bastante eficiente, uma vez que aplica os conceitos e técnicas de um sistema de apoio à decisão, neste caso o DW, auxiliando no processo de extração de dados, transformando-os em informações e apresentando-os de forma a obter indicações da evolução e histórico dos dados.

Uma das principais vantagens de migrar os dados transacionais para um banco de dados DW é a organização dos dados garantindo a integridade e qualidade com que os dados são gravados. É no processo de ETC que as informações passam a ser distribuídas e modeladas seguindo as definições do projeto criado na ferramenta. Para obter qualidade dos dados é implementado o conceito de chave primária dentro das dimensões e chave estrangeira dos cubos em relação às dimensões. As chaves primárias definidas garantem a unicidade dos registros, não permitindo ocorrências duplicadas, servindo ainda como referência na montagem das chaves primárias dos cubos de decisão. A ferramenta utiliza-se destas referências para executar a limpeza dos dados garantindo que registros que não possuem integridade referencial válidas sejam descartados.

A ferramenta apresenta técnicas de modelagem de dados que mostram o quanto é importante organizar, referenciar e garantir a integridade dos dados, transformando-os em informações de grande valor para as organizações. Através dos cubos de decisão a ferramenta orienta as informações por assunto, permitindo montar consultas para cada característica em comum que os dados possam apresentar. Ainda, de forma integrada, preocupa-se em trazer dados que possuam informações idênticas, porém de diferentes fontes unificando o estado do dado.

Desenvolvida em ambiente web a ferramenta é acessível através de qualquer navegador de internet, tendo assim um grande diferencial de usabilidade.

Ainda por estar em sua primeira versão, a ferramenta pode evoluir em diversas extensões, tais como, novas funções de agregação e aplicação de algoritmos de mineração de dados, entre outros, permitindo ganhos de qualidade e aproveitamento no gerenciamento estratégico, tático e operacional de uma organização.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBIERI, Carlos. **BI – Business Intelligence – Modelagem & Tecnologia.** Rio de Janeiro: Editora Axel Books, 2001.

COME, Gilberto. **Contribuição ao estudo da implementação de data warehousing**: um caso no setor de telecomunicações. 2001. 132 f. Dissertação (Mestrado em Administração) – Curso de Pós-graduação em Administração, Universidade de São Paulo, São Paulo.

DW BRASIL. **Características de um data warehouse**. Brasília, 2005. Disponível em: <http://www.dwbrasil.com.br/html/artdw_carac.html>. Acesso em: 11 set. 2006.

INMON, William H. **Como construir o data warehouse**. Tradução Ana Maria Netto Guz. Rio de Janeiro: Campus, 1997.

NAVARRO, Maria C. **O que é Data Warehouse?** Brasília, 1996. Disponível em: <http://www.serpro.gov.br/publicacao/tematec/1996/ttec27>. Acesso em: 13 mai. 2007.

OLIVEIRA, Pércio A. Ferramenta de construção de Data Warehouse**.** 2007. 89 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Departamento de Sistemas e Computação, Universidade Regional de Blumenau, Blumenau.

SILVA, Diogo. **SITC:** uma ferramenta de transformação e carga para um data warehouse**.** 2005. 31 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Instituto de Matemática, Universidade da Bahia, Salvador.

VIEIRA, Fernando. **Alguns conceitos sobre DW**. São Paulo, 2000. Disponível em: <http://www.datawarehouse.inf.br/>. Acesso em: 19 set. 2006.