UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE INFORMÁTICA

DATA WAREHOUSING – UMA EXPERIÊNCIA NA CONSTRUÇÃO DE UM DATA MART

Autor : Simone Maria Straiotto
Orientador : Prof. Dr. Martin A Musicante

CURITIBA
2004

SIMONE MARIA STRAIOTTO

DATA WAREHOUSING – UMA EXPERIÊNCIA NA CONSTRUÇÃO DE UM DATA MART

Monografia apresentada para obtenção do título de Especialista em Informática no Curso de Especialização em Informática com Ênfase em Tecnologia da Informação, Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientador : Prof. Dr. Martin A Musicante

CURITIBA

2004

Parecer de Aprovação
Monografia de Especialização
Programa de Pós-Graduação em Informática/UFPR

Declaramos que a aluna **Simone Maria Straiotto** entregou a versão final da sua Monografia de Especialização em Informática da UFPR intitulada *Data Warehousing: Uma experiência na construção de um Data Mart.*

Curitiba, 11 de agosto de 2004.

Martin A. Musicante (Orientador)
Professor Adjunto III
Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Informática
Caixa Postal 19018
CEP 81531-990 Curitiba PR

Elias Procópio Duarte Jr.
Professor Adjunto III
Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Informática
Caixa Postal 19018
CEP 81531-990 Curitiba PR

Agradeço à Cláudia, pela amizade e pela generosidade em dividir o seu conhecimento, ao João Luiz por permitir que este estudo fosse realizado, e a todos os amigos e familiares que pacientemente me suportaram durante este período.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

Resumo

A apresentação deste trabalho consiste de uma exposição teórica sobre a tecnologia de Data Warehouse e a experiência na construção de um Data Mart. Na revisão bibliográfica é dada uma visão sobre o assunto, desde o seu conceito, mostrando os benefícios e componentes, passando pela Arquitetura até a Modelagem. Após finalizada esta primeira parte é mostrado um caso prático de construção de um Data Mart para uma Instituição Financeira e são comentadas as diversas partes do projeto.

1 Introdução

Cada vez mais as empresas precisam ter informações que auxiliem no processo de tomada de decisão e assim atingir maiores e melhores resultados, mas o que acontece é que estas informações estão espalhadas pelos sistemas operacionais das empresas e raramente servem como recurso estratégico no seu estado original. Então, possuir uma ferramenta que permita reunir rapidamente estas informações e que ainda as apresente conforme a visão de quem precisa delas, é um enorme diferencial.

Para alguém que já trabalhou criando sistemas gerenciais a partir de grandes bancos de dados no mainframe e telas estáticas com informações pré formatadas estar conhecendo a tecnologia do Data Warehouse é abrir uma nova perspectiva de atuação. Ela permite gerar um produto que oferece uma grande flexibilidade ao usuário final, combinando com rapidez as informações dos aplicativos operacionais com uma visão gerencial e garante que os administradores obtenham respostas instantâneas e pessoais e não mais informações padronizadas.

Este trabalho tem por objetivo consolidar o conhecimento teórico sobre Data Warehouse com uma experiência prática e está organizado como segue: O Capítulo 2 apresenta uma Revisão Bibliográfica onde se procura dar uma visão sobre o assunto, desde o seu conceito, mostrando os Benefícios e Componentes, passando pela Arquitetura até a Modelagem. Posteriormente, como forma de consolidação da teoria, é apresentado no Capítulo 3 a definição e o funcionamento de um pequeno Data Mart que trata-se de um caso real. Para a publicação foi necessário a substituição dos dados reais por dados fictícios para manter o sigilo das informações da empresa, mas este fato não trouxe qualquer prejuízo, pois a intenção é mostrar a sua construção.

2 Revisão Bibliográfica

Data Warehouse, segundo INMON (1997, p.33), "é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, que serve de apoio às decisões gerenciais."

O que se quis dizer com:

- dados baseado em assunto -> um Data Warehouse direciona o seu interesse nas principais entidades do negócio enquanto que um sistema aplicativo é direcionado ao processamento do dado, ao seu registro e a sua manutenção. Exemplo é dado por SINGH (2001, p.14) ao citar um Data Warehouse de uma entidade de ensino superior com assuntos como alunos, cursos, departamentos e professores, enquanto que no sistema aplicativo estão os registros de alunos.
- Integrado -> significa que em um Data Warehouse os dados oriundos dos aplicativos e que referem-se a mesma informação terão o mesmo domínio. Assim se no Data Warehouse padronizou-se, por exemplo, que para determinar que o registro trata-se de um indivíduo do sexo masculino a notação é "m" e para o feminino é "f", então todos os dados vindo das aplicações referentes a esta informação serão transformados seguindo estes domínios.
- Não volátil -> os dados carregados no Data Warehouse não sofrem atualização. São carregados e acessados.
- Variável em relação ao tempo -> significa que os dados de um Data Warehouse estão relacionados com um período de tempo, isto é, estes dados são fotografias de um momento e em outra carga se terá outra imagem destes mesmos dados e assim por diante. Tanto isto é importante que a estrutura que é chave num Data Warehouse sempre tem um elemento de tempo, por exemplo, ano, mês, dia, etc.

Então, pode-se dizer que um Data Warehouse é um grande repositório de dados obtidos das áreas de negócio e trabalhados de maneira que as informações possam ser obtidas pelos tomadores de decisão.

2.1 FUNCIONALIDADES

Cada vez mais as empresas necessitam de informações que ajudem no processo de tomada de decisão. A informação rápida e consistente é um diferencial capaz de tornar uma empresa mais competitiva e eficaz que outra. E a função de um Data Warehouse é fornecer estas informações que foram coletadas de todos os sistemas que integram a organização permitindo análises gerenciais que irão auxiliar no seu crescimento e sobrevivência.

2.2 BENEFÍCIOS – Por que Construir um Data Warehouse?

Vivemos um tempo de intensa competitividade entre as empresas e ter informações sobre o negócio é crucial para a sobrevivência e desenvolvimento da organização. É necessário transformar dados em conhecimento.

Um Data Warehouse permite que as informações sejam obtidas rapidamente e de forma integrada, isto é, por agrupar todos os dados da empresa é possível fazer análises e comparações em todos os setores, além de trabalhar com históricos permitindo acompanhar os resultados de qualquer setor ou atividade que se julgar necessário.

Outro aspecto a ser considerado é a interdependência entre as diversas áreas que compõem uma organização e a necessidade de analisar as informações vindas de cada uma delas de maneira integrada, por exemplo, para se fazer um orçamento é necessário prever receitas e despesas e estas informações estão espalhadas por diversos setores, mas no Data

Warehouse elas estarão num mesmo repositório e podem ser agrupadas e comparadas com maior agilidade pelos administradores.

2.3 COMPONENTES DE UM DATA WAREHOUSE

Existem diferentes descrições dos componentes de um Data Warehouse e uma delas é a apresentada por KIMBALL e ROSS (2002, pg.8-17) e condensada a seguir.

É preciso considerar quatro componentes separados e distintos quando analisamos o ambiente de DW: sistemas operacionais de origem, data staging area, área de apresentação de dados e ferramentas de acesso a dados.

- **Sistemas Operacionais de Origem**

  Esses são os sistemas operacionais de registro que capturam as transações da empresa. Os sistemas de origem devem ser considerados como externos ao Data Warehouse porque presume-se que se tenha pouco ou nenhum controle sobre o conteúdo e o formato dos dados nesses sistemas operacionais legados. As principais prioridades dos sistemas de origem são o desempenho e a disponibilidade de processamento. As consultas feitas nos sistemas de origem são consultas limitadas, feitas em um registro por vez, que fazem parte do fluxo normal de transações e com uma demanda no sistema operacional extremamente limitada. Parte-se do forte princípio de que os sistemas de origem não são consultados da forma ampla e inesperada que os Data Warehouses costumam ser. Os sistemas de origem mantêm um volume pequeno de dados históricos e se estiver com um Data Warehouse adequado , a responsabilidade de os sistemas operacionais representarem o passado será significativamente reduzida. Normalmente, cada sistema de origem é uma aplicação naturalmente independente natural, em que foi

feito um investimento mínimo no compartilhamento de dados comuns como produto, cliente, geografia ou agenda com outros sistemas operacionais da empresa.

- **Data Staging Area**

A Data Staging Area do Data Warehouse é tanto uma área de armazenamento com um conjunto de processo e normalmente denomina-se ETL (Extract-Transformation-Load). A Data Staging Area abrange tudo entre os sistemas operacionais de origem e a área de apresentação dos dados.

A extração é a primeira etapa do processo de obtenção de dados no ambiente de Data Warehouse. O processo de extração envolve a leitura e a compreensão de dados de origem e cópia dos dados necessários ao Data Warehouse na Data Staging Area para que sejam manipulados posteriormente.

Depois que os dados são extraídos para a Data Staging Area, ocorrem muitas transformações em potencial, como filtragem dos dados (correções de erros de digitação, solução de conflitos de domínio, tratamento de elementos ausentes ou a divisão em formatos padrão), combinação de dados das várias origens, cancelamento de dados duplicados e atribuição de chaves de Data Warehouse. Essas transformações são todas precursoras para carregar os dados na área de apresentação do Data Warehouse.

Um banco de dados normalizado para o armazenamento da Data Staging Area é aceitável. No entanto, ainda se tem algumas restrições quanto a esse método. A criação de estruturas normalizadas para Data Staging e de estruturas dimensionais para apresentação significa que os dados são extraídos, transformados e carregados duas vezes – uma no banco de dados normalizado e outra quando carregamos o modelo dimensional...

Independentemente de estarmos trabalhando com vários arquivos simples ou com uma estrutura de dados normalizada na Data Staging Area, a última etapa do processo ETL é carregar os dados. O processo de carga no ambiente do Data Warehouse normalmente assume a forma de um processo de apresentação de tabelas dimensionais garantidas por controle de qualidade para os vários recursos de carga de cada Data Mart. O Data Mart de destino precisa então indexar os dados que acabaram de chegar para execução da consulta. Depois que cada Data Mart tiver acabado de ser carregado, indexado e suprido com as agregações apropriadas e verificado para controle de qualidade, a comunidade de usuários será avisada que os novos dados foram publicados. Nesse caso, a publicação inclui a comunicação da natureza de qualquer alteração que tenha ocorrido nas dimensões e novas conclusões que tenham sido introduzidas nos fatos medidos ou calculados.

- **Apresentação dos Dados**

  A área de apresentação dos dados é o local em que os dados ficam organizados, armazenados e tornam-se disponíveis para serem consultados diretamente pelos usuários, por criadores de relatórios e por outras aplicações de análise...

  Normalmente, nos referimos à área de apresentação de dados como uma série de Data Marts integrados. Um Data Mart é uma parte do todo que compõe a área de apresentação. Em sua forma mais simples, um Data Mart representa os dados de um único processo do negócio. Esses processos de negócio cruzam os limites das funções organizacionais.

- **Ferramentas de Acesso a Dados**

  O último componente principal do ambiente de Data Warehouse são a(s) ferramenta(s) de acesso a dados. Usa-se o termo "ferramenta" para designar a variedade de recursos com que usuários de negócio podem contar para melhorar a tomada de decisões analíticas. Por definição, todas as ferramentas de acesso de dados consultam os dados na área de apresentação do Data Warehouse.

  Uma ferramenta de acesso a dados pode ser tão simples como uma ferramenta de consulta específica ou tão complexa quanto uma aplicação sofisticada de modelagem ou exploração de dados. As ferramentas de consulta específicas, eficientes como são, podem ser compreendidas e usadas de modo produtivo apenas por uma pequena porcentagem da população de usuários em potencial do Data Warehouse. A grande maioria da base de usuários provavelmente acessará os dados através de aplicações analíticas orientadas a parâmetros. Aproximadamente entre 80% e 90% dos usuários em potencial serão atendidos por essas aplicações enlatadas que nada mais são do que modelos prontos que não exigem que os usuários construam consultas relacionais diretamente.

Os dados são um importante componente de um Data Warehouse e eles podem estar divididos segundo a seguinte estrutura (SINGH. 2001, p.21):

- Dados Atuais:

  Referem-se aos acontecimentos mais recentes.

  São armazenados no menor nível de granularidade.

  Representam grandes volumes.

São armazenados em disco para facilitar o acesso, mas isso acarreta alto custo de gerenciamento.

- Dados Antigos

  Referem-se aos dados históricos que podem vir a ser usados para compor uma série histórica.

  Devem ser mantidos no mesmo nível de granularidade dos dados atuais.

  Podem ser armazenados em um meio mais barato, por exemplo, em cartuchos, e recuperados quando necessário

- Dados Sumarizados

  São dados extraídos do nível mais baixo de detalhe e sumarizados conforme parâmetros pré-definidos.

  Podem ser levemente sumarizados ou altamente sumarizados facilitando com isso o acesso, pois assim podem diminuir o volume do dado no seu nível mais baixo.

  Podem ser ou não armazenados no Data Warehouse dependendo da freqüência da sua utilização.

- Metadados

São dados sobre os dados, tais como, estrutura segundo a visão do programador e segundo a visão do usuário final, algoritmos usados para a sumarização, mapeamento das transformações sofridas na passagem destes dados do ambiente operacional para o ambiente do Data Warehouse e estatísticas de utilização do dado.

Possuem um catálogo dos dados que estão no Data Warehouse e os indicadores e ponteiros para acessá-los.

Mapeam as entidades (Modelo de Dados) e seu relacionamento com o Data Warehouse e caso ocorram alterações é capaz de auxiliar na alteração da aplicação do usuário final.

Segundo SINGH (2001, p. 291), "O metadado funciona, de certo modo, como o coração do ambiente do Data Warehouse. Criar definições de metadado completas e eficientes pode ser um processo demorado, mas quanto melhores as definições, melhor será a compreensão da comunidade de usuários."

Outra citação que enfatiza a importância do metadado:

" Quando os dados são armazenados ao longo do tempo, não é suficiente apenas armazenar o conteúdo, pois o contexto da informação é tão importante quanto o conteúdo para ser possível compreender e interpretar a informação. O contexto da informação é armazenado e gerenciado com o passar do tempo em forma de metadados."
(INMON; WELCH; GLASSEY, 1999, p.97).

2.4 ARQUITETURA

A arquitetura de um Data Warehouse demonstra como é feita a armazenagem, integração, comunicação, processamento e apresentação dos dados e depende de como é implementado o projeto em função das necessidades e possibilidades de cada organização.

Segundo MACHADO (2000, p.31), "Muitas variáveis afetam a escolha da implementação e arquitetura, entre elas o tempo para a execução do projeto, o retorno do investimento a ser realizado, a velocidade dos benefícios da utilização das informações, a satisfação do usuário executivo e os recursos necessários à implementação de uma arquitetura."

A escolha da arquitetura será determinada por onde o Data Warehouse irá residir. Baseado nisso MACHADO (2000, p.32-36), apresenta três tipos de arquiteturas: Global, Independente e Integrada que estão condensadas a seguir:

Global: O Data Warehouse é projetado e construído baseado nas necessidades da empresa como um todo. É considerado como um repositório comum de dados de suporte à decisão, disponível para toda a empresa, ou melhor, em toda a empresa. A arquitetura global pode ser fisicamente centralizada ou fisicamente distribuída nas instalações de uma empresa.

Os dados são extraídos de sistemas operacionais e possivelmente de fontes de dados externas por meio de processos batch em horários fora do pico de operações. Eles são filtrados, eliminam-se os dados não necessários e realiza-se a transformação para a qualidade e necessidade dos requisitos levantados para o projeto. Eles são então carregados nas bases de dados apropriadas de DW para acesso aos usuários finais.

A arquitetura global habilita os usuários finais a utilizar visões corporativas de dados, que normalmente são requisitos de negócio; entretanto, este tipo de ambiente consome muito tempo de desenvolvimento e administração, assim como seu custo de implementação é muito alto.

Independente: A arquitetura independente implica em Data Marts stand-alone controlados por um grupo específico de usuários e que atende somente às suas necessidades específicas e departamentais, sem foco corporativo algum.

Por exemplo, os dados são extraídos dos sistemas operacionais por meio de geração interna do departamento, com auxílio da área ou departamento de tecnologia da informação ... Este tipo de arquitetura raramente tem impacto nos recursos de tecnologia da informação e resulta sempre em implementação rápida. Entretanto, sua restrição é que possui um mínimo de

integração corporativa e não permite nenhuma visão global. Normalmente, este tipo de Data Mart está acessível somente ao pessoal do departamento específico proprietário do Data Mart.

Integrado: A arquitetura de Data Marts integrados é basicamente uma distribuição de implementação. Apesar de os Data Marts serem implementados separadamente por grupos de trabalho ou departamentos, eles são integrados ou interconectados, provendo uma visão corporativa maior dos dados e informações. De fato, o alto nível de integração é similar ao da arquitetura global. Por outro lado, os usuários de um departamento podem acessar e utilizar os dados de um Data Mart de outro departamento.

Neste caso a atuação da área de tecnologia da informação deve ser bem maior que na arquitetura independente, ficando sob sua responsabilidade o controle e administração dos Data Marts.

Cada departamento é proprietário de seus dados, porém as ferramentas e recursos necessários para implementação são providos e administrados por tecnologia da informação.

Um resumo esclarecedor sobre a Arquitetura de um Data Warehouse foi feito por SINGH (2001, p.77), "Cada Data Warehouse é diferente do outro, porém todos são caracterizados por alguns componentes chave:

- Um modelo de dados para definir o conteúdo do Data Warehouse.
- Um banco de dados cuidadosamente projetado, seja hierárquico, relacional ou multidimensional, separado dos bancos de dados operacionais.
- Diversos utilitários para limpeza de dados, gerenciamento de cópias, transporte de dados, replicação de dados e comunicação entre plataformas.
- Um servidor de DW otimizado para processamento rápido de relatórios e consultas.
- Um sistema front-end para suporte à decisão para relatório e análise de tendências."

Um último tópico deve ser considerado ao se falar sobre a arquitetura de um Data Warehouse é o acesso aos dados que evoluíram para a utilização de uma tecnologia chamada OLAP (On-line Analytical Processing) ao contrário dos bancos de dados operacionais que têm um processamento OLTP (On-line Transaction Processing). Esta tecnologias OLAP permitem o acesso a dados e consultas escolhidas pelo usuário sem a dependência de um técnico da informática. A idéia de acessar de maneira eficiente dados multidimensionais é o ponto crucial do Data Warehouse. Estes dados estão armazenados em dimensões e estes arquivos são conhecidos como cubos.

2.5 MODELAGEM

Um modelo de dados é uma representação gráfica dos dados de uma área ou assunto de uma empresa feita para auxiliar o seu entendimento.

Ao se fazer a modelagem de um Data Warehouse não se utiliza um Modelo Relacional e sim de um Modelo Dimensional. Segundo KIMBALL e ROSS (2002, pg.13), "Modelagem dimensional é um novo nome para uma técnica antiga que permitia tornar os banco de dados fáceis e compreensíveis. Caso após caso, a partir da década de 1970, as empresas de TI, as consultorias, os usuários finais e os fornecedores migraram para uma estrutura dimensional simples para atender à necessidade humana fundamental de simplicidade."

O Modelo Relacional é aquele que se utiliza ao desenhar um aplicativo operacional e nele existe a preocupação com a informação como uma transação individual. O modelo é concebido a partir da identificação das Entidades envolvidas e os seus relacionamentos. É um modelo estruturado. Segundo KIMBALL e ROSS (2002, pg.14), "A modelagem normalizada é extremamente útil para o desempenho do processamento operacional porque uma transação de

atualização ou inserção só precisa atingir o banco de dados em um local. No entanto, os modelos normalizados são muito complicados para as consultas do Data Warehouse."

O Modelo Dimensional representa as diversas visões pelas quais um conjunto de informações pode ser analisado. É composto de uma tabela com chave composta chamada Tabela Fato e um conjunto de tabelas menores chamadas Tabelas Dimensões, cujas chaves compõem a chave da Tabela Fato.

No Modelo Dimensional tem-se 3 elementos:

- As **Dimensões,** que normalmente correspondem aos campos não numéricos do banco de dados, são as perspectivas através das quais uma ou mais medidas podem ser analisadas.
- Os conjuntos de **Medidas**, que correspondem normalmente aos campos numéricos dos quais podem derivar diferentes informações dependendo da visão empregada sobre os mesmos.
- As **Agregações** que são a relação entre as dimensões e as medidas e a sua sumarização ou agregação em informações que serão assim armazenadas para facilitar determinadas consultas dos usuários.

O Modelo Multidimensional é também conhecido como Modelo Estrela (Star Schema) devido a sua representação: no centro uma grande tabela e ao seu redor tabelas menores, lembrando o formato de uma estrela.

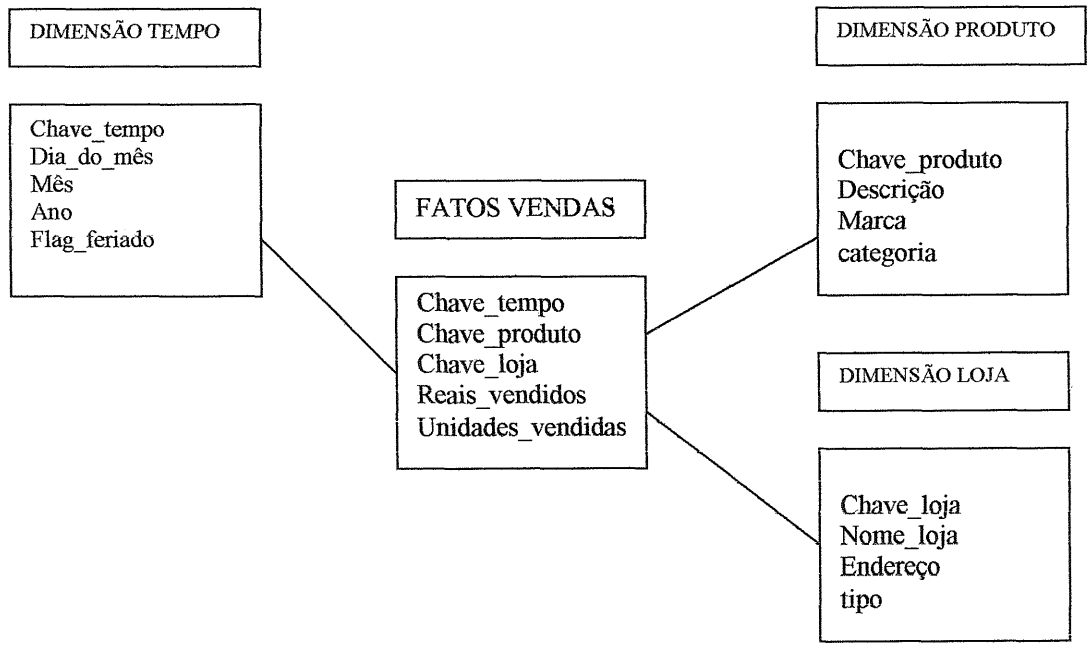
O Modelo Estrela baseia-se numa tabela central chamada de Tabela de Fatos que apresenta conexões com outras secundárias chamadas Tabelas de Dimensão.

Na Tabela de Fatos estão armazenadas todas as medidas numéricas do negócio que o Data Warehouse está atendendo. Nesta tabela ficam os "valores" resultado da ligação de todas as dimensões, enquanto que nas Tabelas de Dimensão estão as descrições textuais das

dimensões do negócio. A Tabela Fato é identificada por uma chave composta pelas chaves das Dimensões associadas. Uma linha da Tabela Fato corresponde a uma combinação única dos domínios de todas as chaves de todas as Dimensões.

Um exemplo simples de um Modelo Estrela é apresentado na figura 1, extraído de OLIVEIRA (2002, p.63):

FIGURA 1 – EXEMPLO DE MODELO ESTRELA



Outros conceitos importantes de serem mencionados ao se falar em Modelagem:

- **Atributo da Dimensão:** são dados, normalmente textos e não aditivos, que descrevem e qualificam uma Dimensão. Através de cada atributo de uma Dimensão é possível visualizar as medidas da Tabela Fato. O poder de análise está intimamente relacionado à qualidade dos Atributos das Dimensões. Por exemplo,

Cliente é uma Dimensão e Sexo do Cliente é um atributo que pode gerar importantes análises e conclusões.

- **Chave Artificial** (SURROGATE KEY): a chave artificial é gerada automaticamente para identificar univocamente cada ocorrência de uma Dimensão. Geralmente corresponde a um número seqüencial que não possui um significado próprio e é transparente para o usuário.

  A vantagem de usá-la é a maior performance em "joins" e a independência e estabilidade das Dimensões em relação às atualizações das respectivas chaves naturais no ambiente operacional.

  Sobre este assunto é importante citar KIMBALL e ROSS (2002, pg.69), "Incentivamos o uso de chaves substitutas em modelos dimensionais em vez de contar com códigos de produção operacionais. As chaves substitutas são conhecidas por muitos outros nomes alternativos: chaves sem significado, chaves de inteiros, chaves não-naturais, chaves artificiais, chaves sintéticas, etc. Resumindo, as chaves substitutas são inteiros atribuídos seqüencialmente conforme necessário para preencher uma dimensão. Por exemplo, o primeiro registro de produto recebe uma chave substituta do produto com o valor de 1, o próximo registro de produto recebe a chave 2, e assim sucessivamente. As chaves substitutas servem simplesmente para unir as tabelas de dimensão à tabela de fatos. "

- **Hierarquia:** corresponde a uma estrutura de Atributos de uma Dimensão que possibilita a visualização em níveis. Exemplos:

DIMENSÃO	HIERARQUIA
Tempo	Ano, Mês, Quinzena, Semana, Dia
Mercado	Região, Estado, Cidade, Loja

- **Granularidade:** determina o maior nível de detalhe da informação que se deseja atingir em uma análise. Ela declara o exato significado de ocorrência da Tabela Fato. Quanto maior o nível de granularidade maior é o detalhamento do dado que se pode obter, mas isso significa um volume também maior.

3. Uma experiência na Construção de um Data Mart

Um Data Mart é uma base de dados contendo informações voltadas para atender necessidades específicas de uma determinada área ou processo de negócio, porém tendo como origem o Data Warehouse. Por exemplo, uma empresa pode ter um Data Mart de Markting, um Data Mart de Crédito, um Data Mart de Orçamento, etc. Mas, enquanto o Data Mart tem um enfoque departamental, o Data Warehouse tem uma visão corporativa. Segundo SINGH (2001, p.14), "É um subconjunto do Data Warehouse empresa-inteira. Tipicamente, desempenha o papel de um Data Mart departamental, regional ou funcional. "

O Data Mart que é o objeto deste estudo foi construído em função da necessidade de informações para auxiliar no processo de decisão da área financeira de uma certa Instituição Financeira X [1]. Esta necessidade vem ao encontro dos objetivos de um Data Warehouse e à rapidez na obtenção de informações que permitem a melhoria da gestão no negócio.

É importante ressaltar que os sistemas convencionais são geralmente projetados para subsidiar o ambiente operacional e não para auxiliar no processo de tomada de decisão e por isso surge a necessidade de sistemas de apoio à decisão.

Este Data Mart permite que o usuário personalize suas consultas de forma a melhor atender suas necessidades, proporcionando acesso rápido e fácil às informações consolidadas ou detalhadas. A sua construção utilizou uma metodologia preocupada com o levantamento adequado, qualificação e consistência dos dados.

[1] A Instituição Financeira X em questão existe de fato, mas por motivos de segurança da informação o seu nome não será divulgado.

As diversas fases da construção deste Data Mart permitem visualizar como é possível utilizar diferentes tecnologias e adequá-las ao tamanho da necessidade de cada projeto. É esta a experiência da equipe.

3.1 Escopo

O projeto a ser tratado possui as seguintes características:

Abrangência: Fornecer informações a respeito das movimentações de Fundos de Investimentos.

Requisitos do Usuário: No levantamento das necessidades da área verificaram-se as visões para o atendimento das expectativas, sendo que as duas principais foram:

- movimentações das aplicações e/ou resgate no período (diário, trimestral ou mensal) por determinadas premissas que não serão descritas por tratar-se do negócio da Instituição Financeira X.
- Evolução do patrimônio líquido no mesmo período e premissas.

Além das necessidades dos usuários também foram implementadas outras visões:

- Desempenho dos Administradores e das Agências em relação ao volume/valor aplicado e resgatado referente a um período;
- Análise por período, dos estornos realizados por Clientes, Fundos, Agência e Administradores;
- Acompanhamento dos tipos das movimentações;
- Análise dos impostos em cotas e valores;
- Movimentação acumulada por Fundo, Administrador, Agência e Cliente.

Dados: A carga inicial e incremental dos dados foi realizada a partir de dois sistemas operacionais e a partir dela sua periodicidade é semanal.

Foi feita a qualificação dos dados, verificando inconsistências e/ou distorções.

Também foi realizada a formatação dos dados visando a geração de arquivos a serem utilizados no Data Mart.

3.2 Procedimentos Operacionais / Definições

Premissas Básicas utilizadas:

- Entendimento dos conceitos que envolvem os sistemas operacionais envolvidos;
- Mapeamento das bases e dos arquivos fontes visando a construção do Data Mart;
- Definição de campos e atributos no processo de extração e transformação dos dados;
- Definição do modelo lógico e físico multidimensional para atendimento das necessidades levantadas.

Estratégia de Desenvolvimento utilizada:

- processo de extração dos dados;
- processo de transformação dos dados;
- definição e confecção dos programas cobol e rotinas para extração, qualificação e carga inicial;
- definição e confecção dos programas e rotinas para extração, qualificação e carga no processo incremental semanal;
- processo de carga dos dados no ambiente Data Mart;
- construção dos Cubos Multidimensionais.

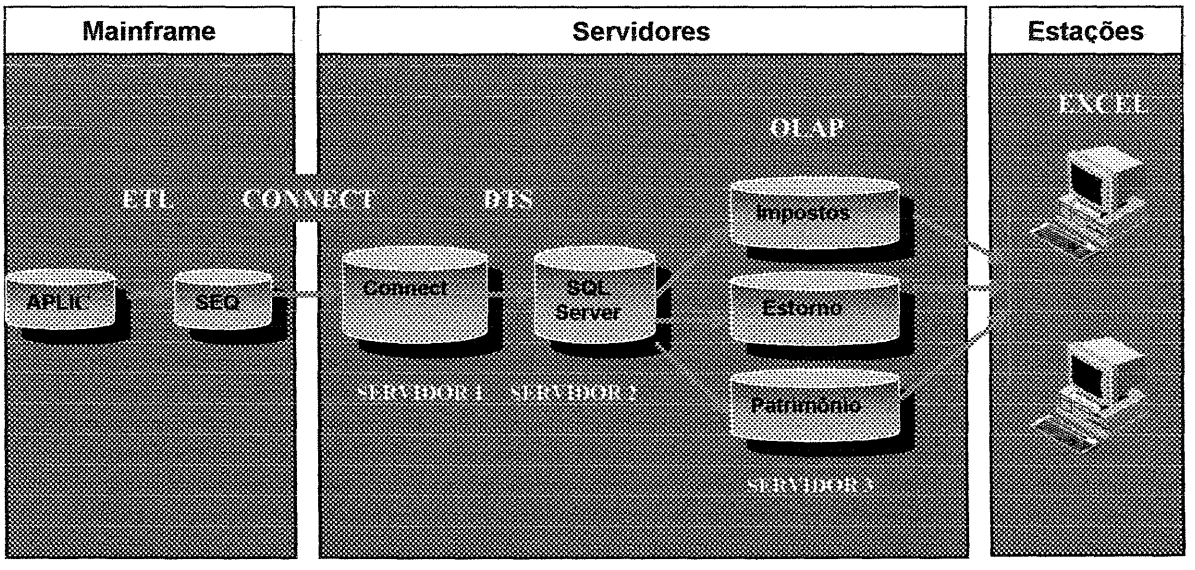
Estratégia de Implantação utilizada:

- transporte dos programas e rotinas de extração, transformação e carga para o ambiente produtivo;
- criação de tabelas em ambiente produtivo (servidores);
- encadeamento das rotinas produtivas;
- treinamento do usuário para utilização da ferramenta;
- disponibilização das informações por um período de teste para validar a integridade do processo.

3.3 Arquitetura

- Foi utilizada arquitetura mista. Primeiramente o processamento acontece no mainframe, em seguida ocorre a transmissão dos dados para um servidor Connect e a apropriação no servidor SQL Server e finalmente a geração das agregações nos cubos OLAP, sendo que a visualização é feita através do Microsoft Excel 2000 (fig. 1).

FIGURA 2 - ETL - Extração, Transformação e Load



3.4 Extração, Qualificação, Formatação e Transmissão dos Dados

Este processo é executado no Mainframe e existem programas e rotinas distintas que fazem a Extração, Qualificação e Formatação dos dados que são transmitidos via Connect-Direct. Existe uma rotina de transmissão para a baixa plataforma no Servidor 1 (fig. 2) para cada arquivo.

Todas estas rotinas estão sincronizadas com o sistema operacional e são controladas por um software chamado Control-M (ver 3.10), que garante o perfeito encadeamento de todo o processo.

Alguns exemplos da qualificação dos dados:

- Código do Administrador: para que o campo seja considerado válido deve ser diferente de zeros, deve conter dados numéricos e não conter espaços.
- Código do Fundo: para que o campo seja considerado válido deve ser diferente de zeros; deve conter dados numéricos.
- Nome do Fundo: para que o campo seja considerado válido não deve conter espaços.
- Data de Processamento: para que o campo seja considerado válido deve ser diferente de zeros, deve conter campos numéricos e deve estar contido no arquivo de datas.

3.5 Validação dos Arquivos Transmitidos e Carga no Banco de Dados SQL

Este processo é executado no Servidor 2 (fig. 2).

A Validação é feita nos arquivos gerados no Mainframe e que foram transmitidos via Connect-Direct.

É verificada a quantidade de registros dos arquivos que chegaram ao Servidor em relação à quantidade de registros gerados no Mainframe (esta quantidade está gravada no

header). É feita também a verificação de formato dos campos, através de um componente do SQL Server chamado DTS (Data Transformation Services), que foi previamente definido para cada arquivo que chega do mainframe.

A Carga das Tabelas é feita por rotinas existentes no Servidor a partir dos arquivos validados. Estas tabelas foram previamente criadas no SQL Server.

3.6 Processamento das Dimensões e Cubos

Criação das Dimensões (Visões): as dimensões são as visões que se quer ter das Medidas (dados valorados). As dimensões são geradas a partir das tabelas criadas no SQL Server , sendo que é possível estabelecer a ordem na qual se quer ver as informações, assim como sua hierarquia; por exemplo, na Dimensão Tempo a hierarquia utilizada é Ano, Trimestre, Mês e Dia. As dimensões existentes são: Administrador, Agência, Fundo, Cliente e Tempo.

Criação dos Cubos: relacionam-se as Medidas (Tabela Fato) com as Dimensões. A ferramenta OLAP onde são criados os Cubos é o Analysis Manager (PATTON; OGLER, 2002).

Ao se criar os cubos escolhe-se a forma de acesso aos dados que pode ser:

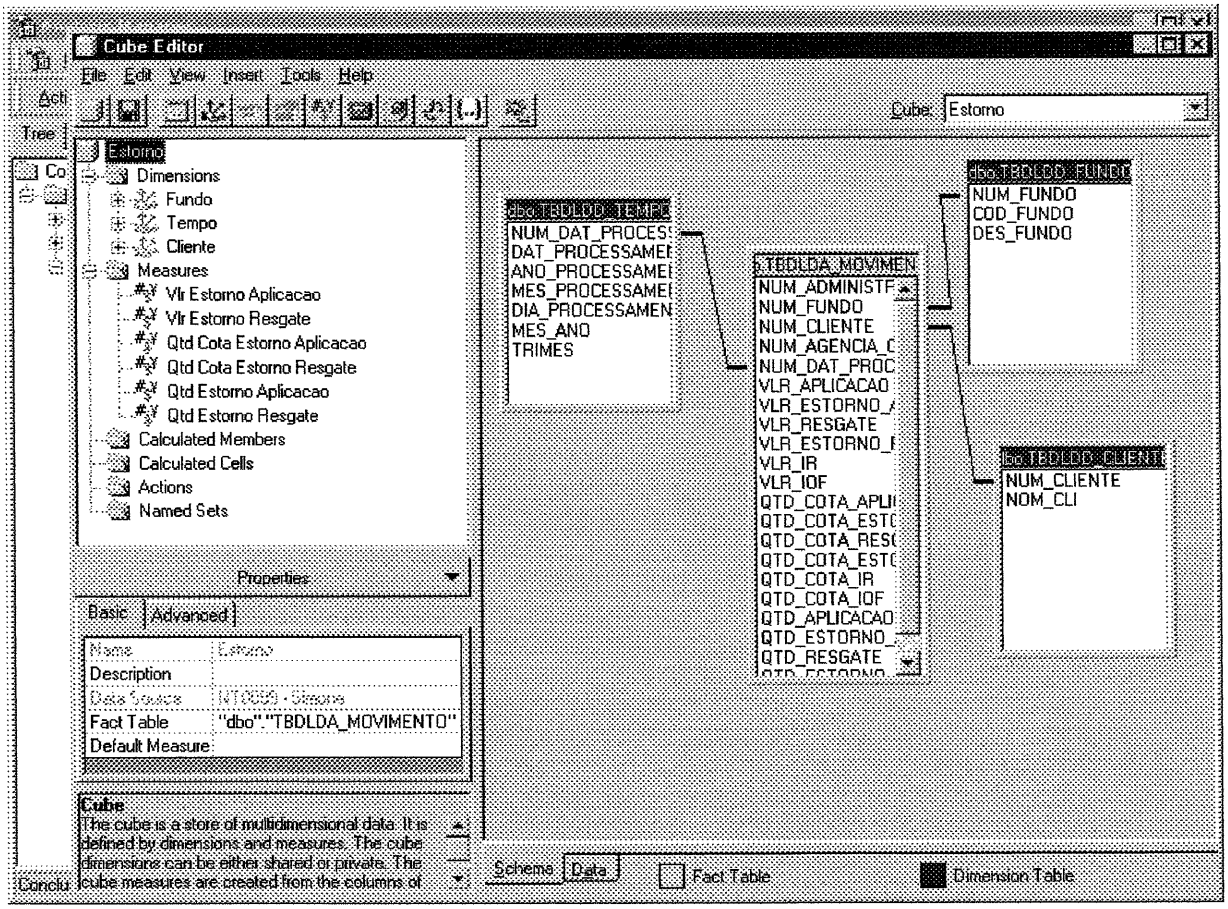
- MOLAP : busca-se os dados diretamente no cubo.
- ROLAP: os dados são obtidos no Banco de Dados.
- HOLAP: é uma forma híbrida que mistura as duas formas anteriores. É a usada neste projeto.

A solução escolhida vai depender do espaço em disco disponível e da necessidade de rapidez de acesso.

Os cubos criados estão demonstrados nas figuras 3, 4, 5 e 6.

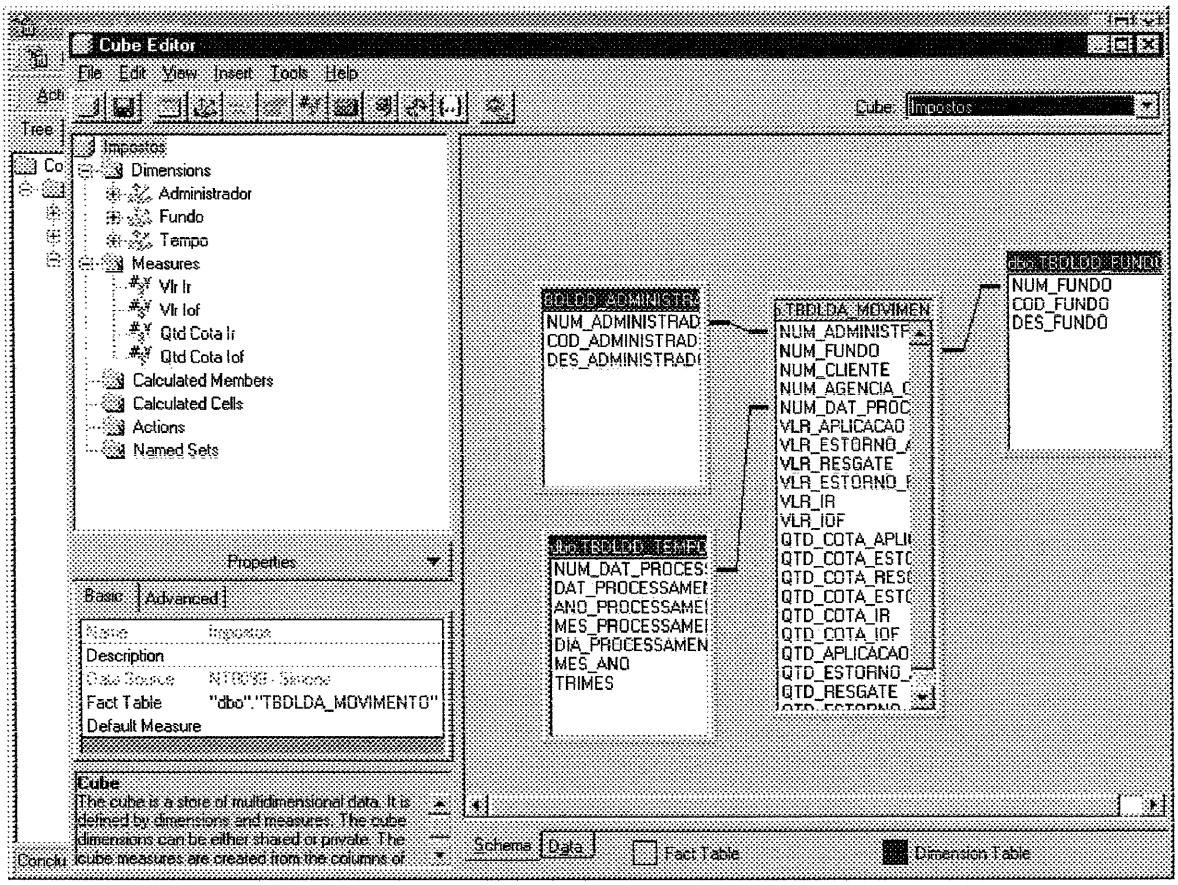
Cubo Estorno: Este cubo relaciona as dimensões Fundo, Tempo e Cliente com as medidas Valor do Estorno da Aplicação, Valor do Estorno do Resgate, Quantidade de Cota do Estorno da Aplicação, Quantidade de Cota do Estorno do Resgate, Quantidade de Estorno da Aplicação e Quantidade de Estorno do Resgate.

FIGURA 3 – CUBO ESTORNO



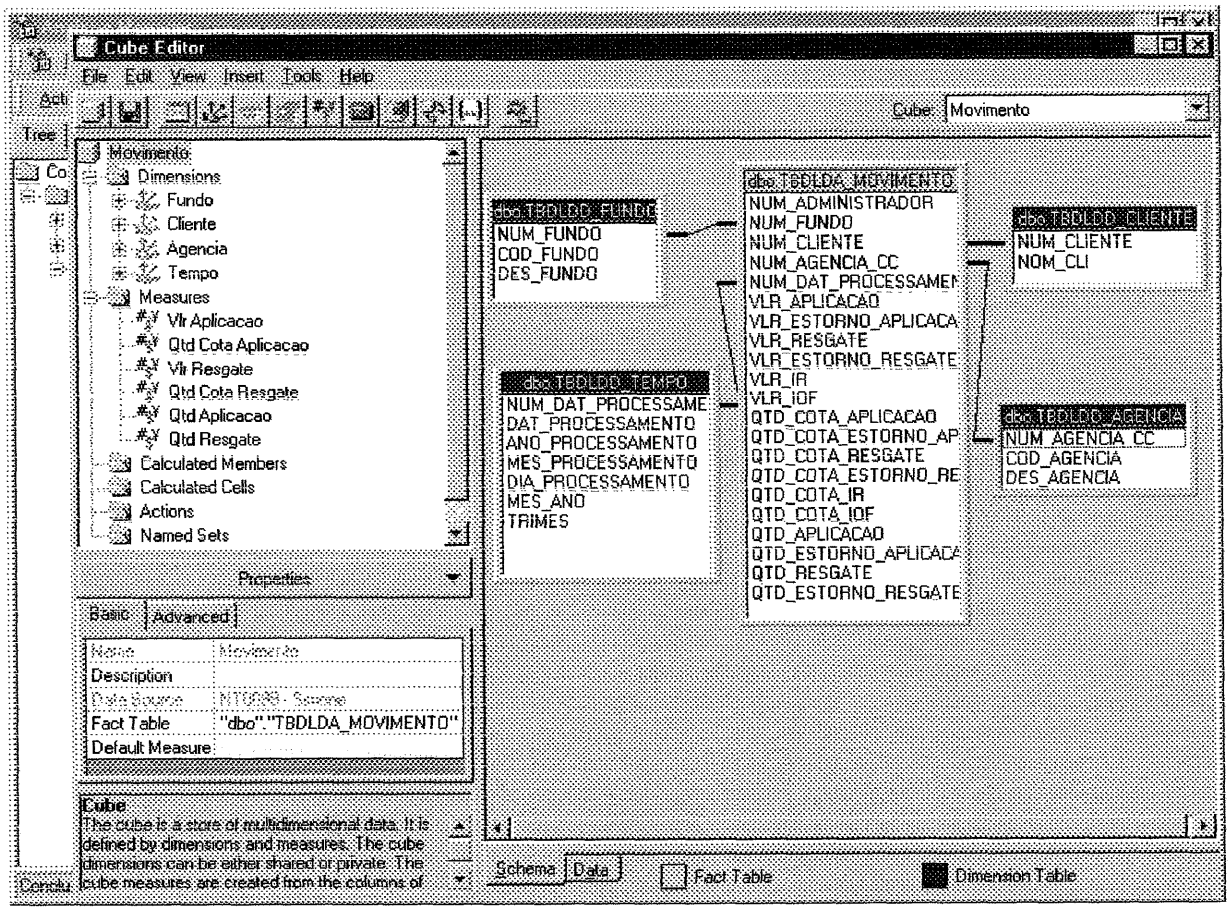
Cubo Impostos: Este cubo relaciona as dimensões Administrador, Fundo e Tempo com as medidas Valor de Imposto de Renda, Valor de Imposto sobre Operações Financeiras, Quantidade de Cota de Imposto de Renda, Quantidade de Cota de Imposto sobre Operações Financeiras.

FIGURA 4 – CUBO IMPOSTOS



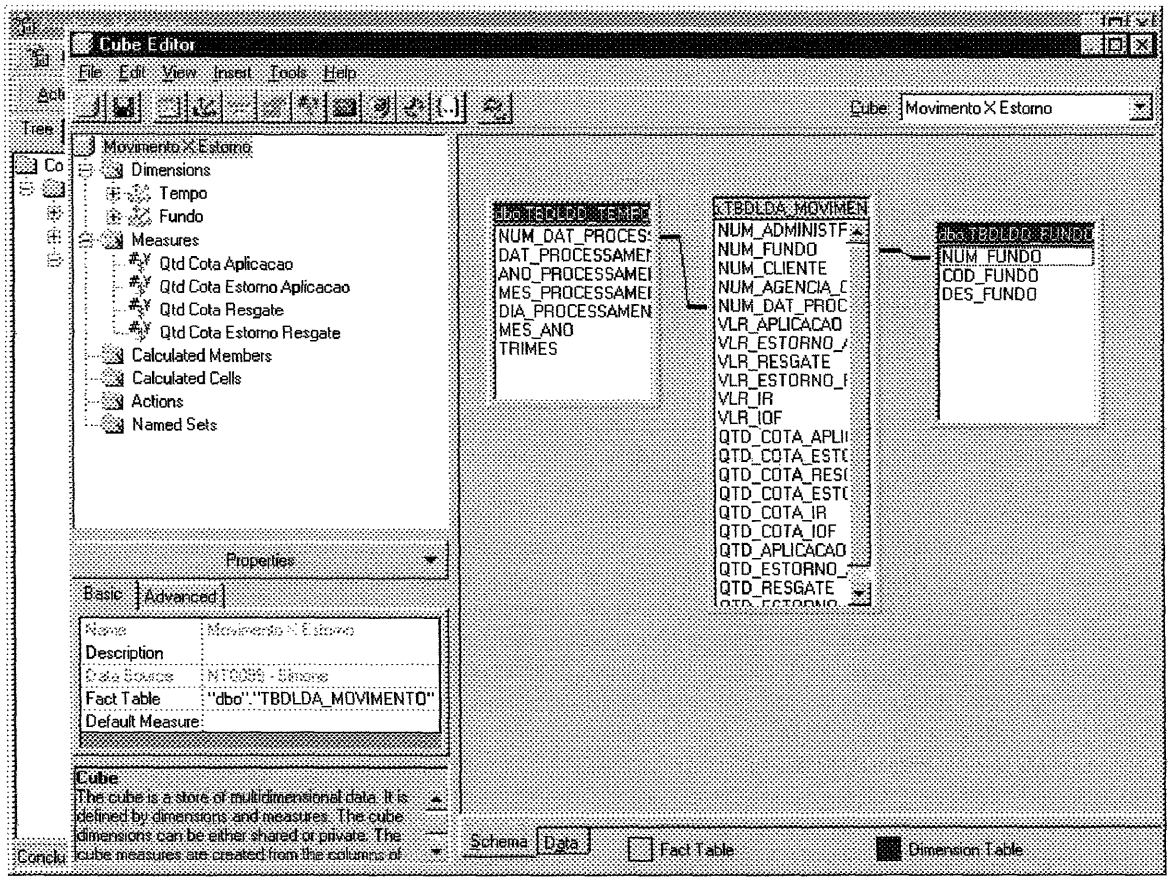
Cubo Movimento: Este cubo relaciona as dimensões Fundo, Cliente, Agência e Tempo com as medidas Valor da Aplicação, Quantidade de Cota do Estorno da Aplicação, Valor do Resgate, Quantidade de Cota do Resgate, Quantidade da Aplicação e Quantidade do Resgate.

FIGURA 5 – CUBO MOVIMENTO



Cubo Movimento x Estorno: Este cubo relaciona as dimensões Tempo e Fundo com as medidas Quantidade de Cota da Aplicação, Quantidade de Cota de Estorno da Aplicação, Quantidade de Cota do Resgate, Quantidade de Cota de Estorno do Resgate.

FIGURA 6 – CUBO MOVIMENTO X ESTORNO



3.7 Criação das Planilhas Excel

A visualização das informações é feita através de Planilhas Excel que é uma ferramenta conhecida pelo usuário.

Também é a garantia de barateamento do projeto, pois ao utilizar o SQL Server já se tem a ferramenta OLAP embutida e também a facilidade do Excel 2000.

A utilização do Excel é muito simples para o usuário, pois basta abrir uma planilha, escolher a opção Dados, depois Obter Dados Externos, em seguida Criar Nova Consulta, depois Cubo OLAP, conectar o Analysis Services e escolher um dos Cubos criados anteriormente. Ao se efetuar todos estes passos surgirá a planilha Excel correspondente ao Cubo escolhido.

Para auxiliar na compreensão a Planilha de Movimento é comentada a seguir. (figs. 7 e 8). Quanto as demais ver ANEXOS.

Planilha Movimento: apresenta as informações do Cubo Movimento. As Dimensões Fundo, Cliente, Agência e Tempo estão combinadas com os valores das medidas e estas Dimensões também podem ser combinadas entre si, retirando os seus dados da planilha com um simples arrastar do mouse. Isto pode ser observado nas figuras 7 e 8 .

Esta mobilidade permite ao usuário diferentes possibilidades de análise.

FIGURA 7 – PLANILHA MOVIMENTO COM A DIMENSÃO FUNDO

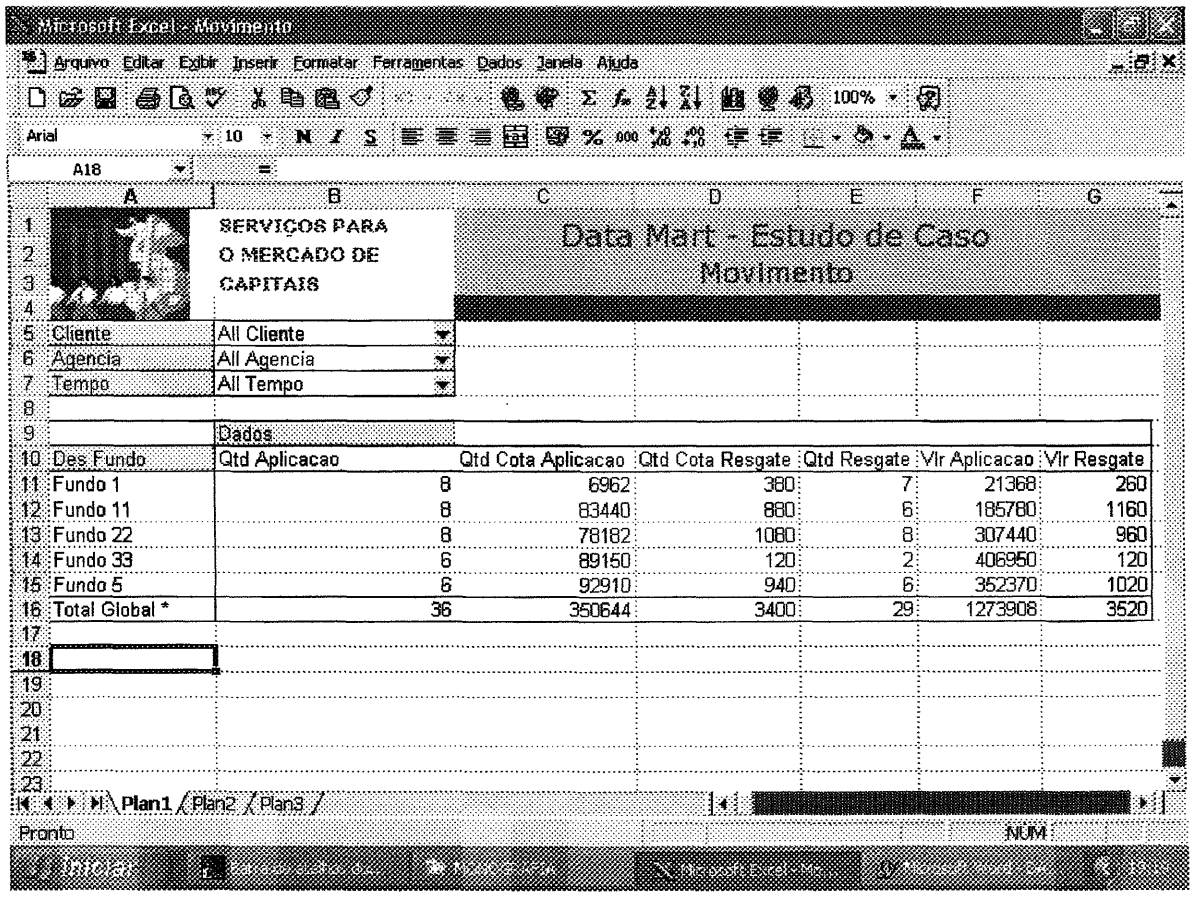
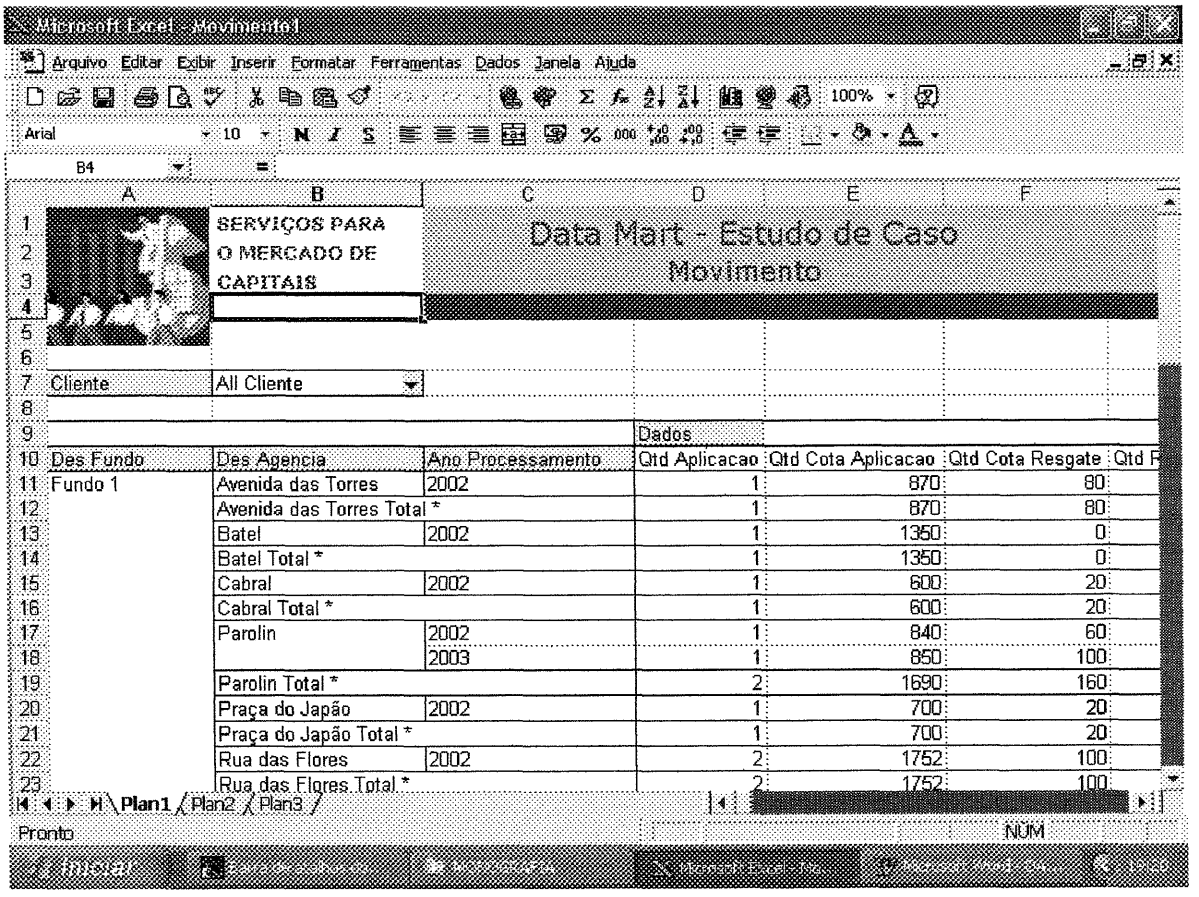


FIGURA 8 – PLANILHA MOVIMENTO COM AS DIMENSÕES FUNDO, AGÊNCIA E TEMPO



Microsoft Excel - Movimento1

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

SERVIÇOS PARA O MERCADO DE CAPITAIS

Data Mart - Estudo de Caso
Movimento

Cliente: All Cliente

Des Fundo	Des Agencia	Ano Processamento	Dados: Qtd Aplicacao	Qtd Cota Aplicacao	Qtd Cota Resgate	Qtd R
Fundo 1	Avenida das Torres	2002	1	870	80	
	Avenida das Torres Total *		1	870	80	
	Batel	2002	1	1350	0	
	Batel Total *		1	1350	0	
	Cabral	2002	1	600	20	
	Cabral Total *		1	600	20	
	Parolin	2002	1	840	60	
		2003	1	850	100	
	Parolin Total *		2	1690	160	
	Praça do Japão	2002	1	700	20	
	Praça do Japão Total *		1	700	20	
	Rua das Flores	2002	2	1752	100	
	Rua das Flores Total *		2	1752	100	

Plan1 Plan2 Plan3

Pronto

NUM

3.8 Metadados dos Cubos

É a conceituação dos dados e é utilizada pelo usuário para esclarecer dúvidas quanto aos seus conteúdos. O Metadado pode ser armazenado num banco de dados, relatório ou planilha, e deve estar acessível ao usuário.

A seguir estão alguns exemplos:

Dimensão: **Administradores**
Definição: **Registro de Administradores de Fundos**.

Nome do Atributo	Definição do Atributo
Nome Adm	Nome do Administrador dos Fundos de Investimentos
Cod Adm	Código do Administrador dos Fundos de Investimentos

Dimensão: **Fundos**
Definição: **Registro de Cadastro de Fundo**.

Nome do Atributo	Definição do Atributo
Cod Fundo	Código Fundo de Investimento
Nome Fundo	Nome do Fundo de Investimento

Dimensão: **Data**
Definição: **Registro das Datas**

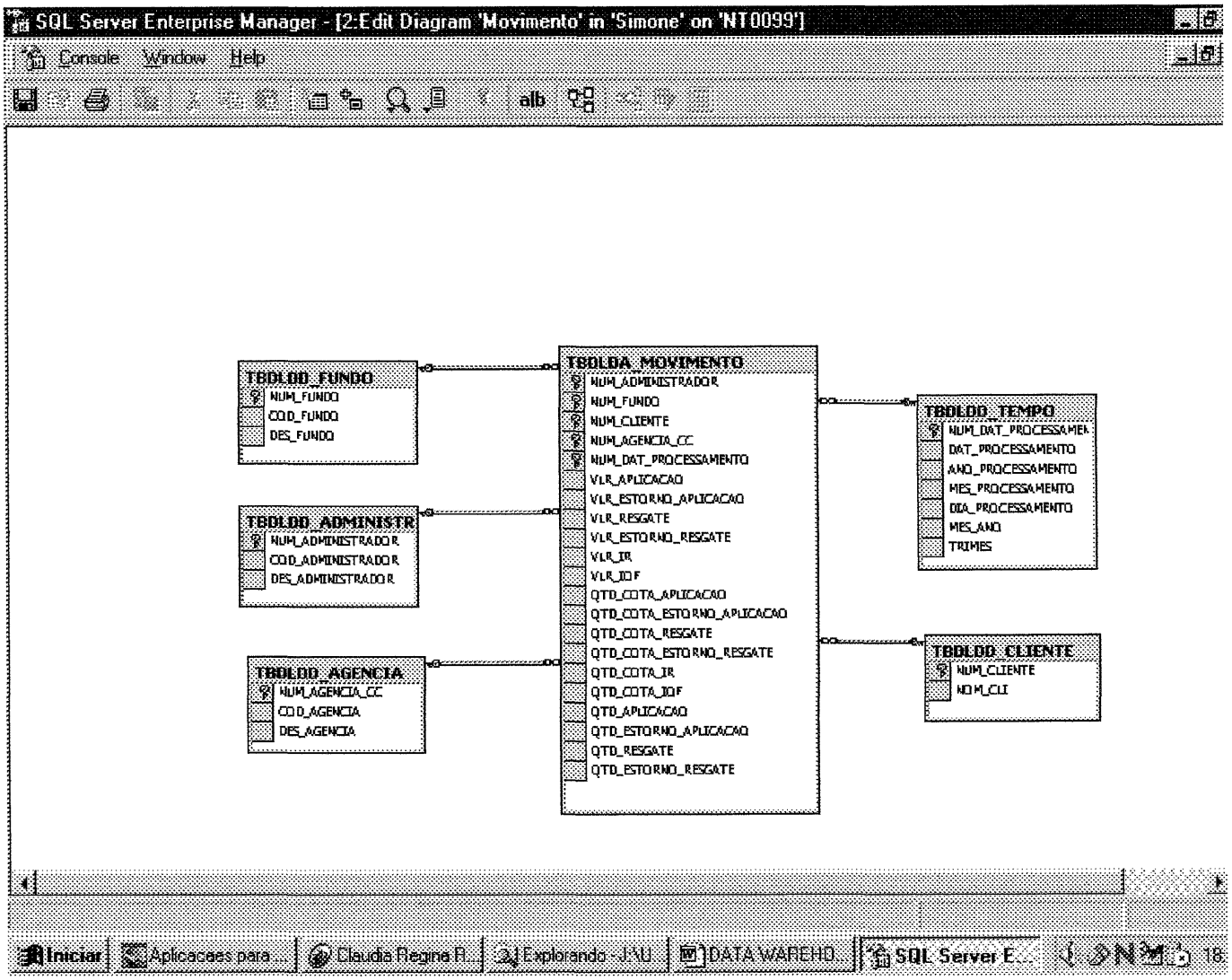
Nome do Atributo	Definição do Atributo
Ano	Ano
Trimestre	Trimestre
Mês	Mês
Dia	Dia

3.9 Modelo Estrela

A base de dados foi gerada no SQL Server utilizando a Ferramenta Enterprise Manager (PETKOVIC, 1999). Neste momento é criado um database onde existe uma área de dados e uma de log. Para isso é informado o nome do database, tamanho da área de dados, tamanho do log e então é gerado um comando SQL que é submetido no SQL SERVER.

Dentro deste Database foram criadas as tabelas dimensões Fundo, Administrador, Agência, Tempo, Cliente e a tabela fato Movimento, bem como os índices destas tabelas e os relacionamentos entre elas. As tabelas dimensões Fundo, Administrador, Agência, Tempo e Cliente possuem chaves artificiais chamadas Num_Fundo, Num_Administrador, Num_Agencia_CC, Num_Dat_Processamento e Num_Cliente respectivamente e apresentam números seqüenciais gerados nos processos no mainframe. A tabela fato Movimento tem como chave os campos que representam as chaves das tabelas dimensões e isso mantém o relacionamento entre elas (fig. 9).

FIGURA 9 – MODELO ESTRELA DO DATA MART DE FUNDOS



3.10 Tecnologia utilizada

SQL Server 2000

Os serviços OLAP do SQL Server 2000 são nomeados de Analysis Services (PATTON; OGLER, 2002).

Na essência dos recursos do OLAP do SQL Server está o uso de cubos que permitem a construção de visualizações multidimensionais de dados dinâmicos.

O software que faz a carga das tabelas no servidor a partir dos arquivos texto gerados pelos sistemas operacionais no mainframe é o DTS (PETKOVIC, 1999).

Office 2000

Microsoft Excel: como front-end para o SQL Server

Control- M

É um produto de controle de rotinas que garante o perfeito encadeamento das mesmas. É possível controlar a execução estabelecendo as predecessoras e sucessoras de cada rotina.

Connect Direct

É um produto utilizado para a transferência de arquivos em instalações com grande volume de dados. Possibilita que se faça a carga de arquivos do mainframe para os servidores Windows NT (download) ou dos Servidores Windows NT para o mainframe.

4 Conclusão

Este trabalho apresentou o estudo dos principais conceitos de Data Warehouse e um pequeno estudo de caso no desenvolvimento de um Data Mart para uma Instituição Financeira.

Ao procurar aliar a teoria e a prática foi possível consolidar o conhecimento e observar que na construção deste Data Mart aspectos como a padronização da informação e a qualificação dos dados foram muito importantes para o sucesso do projeto. Outro ponto que chamou a atenção foi a utilização de chaves artificiais nas tabelas do modelo dimensional, o que garantiu um melhor desempenho no acesso às informações.

A contribuição deste trabalho foi a demonstração de que um Data Mart pode se adaptar ao tamanho da solução que a empresa precisa para cada área, pois nele foi utilizada uma tecnologia que permitiu o barateamento do projeto com a utilização do SQL Server que já possui uma ferramenta OLAP embutida e o uso do Excel 2000 que já era uma ferramenta utilizada e conhecida pelo usuário.

Referências

1. INMON, W.H. Como Construir o Data Warehouse. Tradução da 2. Ed. Rio de Janeiro: Campus, 1997.

2. INMON, W.H.; WELCH, J.D.; GLASSEY, L. Gerenciando Data Warehouse. São Paulo: Makron Books, 1999.

3. KIMBALL, Ralph; ROSS, Margy. Data Warehouse Toolkit; o guia completo para modelagem multidimensional; tradução de Ana Beatriz Tavares, Daniela Lacerda. Rio de Janeiro: Campus, 2002.

4. MACHADO, Felipe Nery Rodrigues. Projeto de Data Warehouse: Uma Visão Multidimensional. São Paulo: Érica, 2000.

5. OLIVEIRA, Wilson José. Data Warehouse. Florianópolis: Visual Books, 2002.

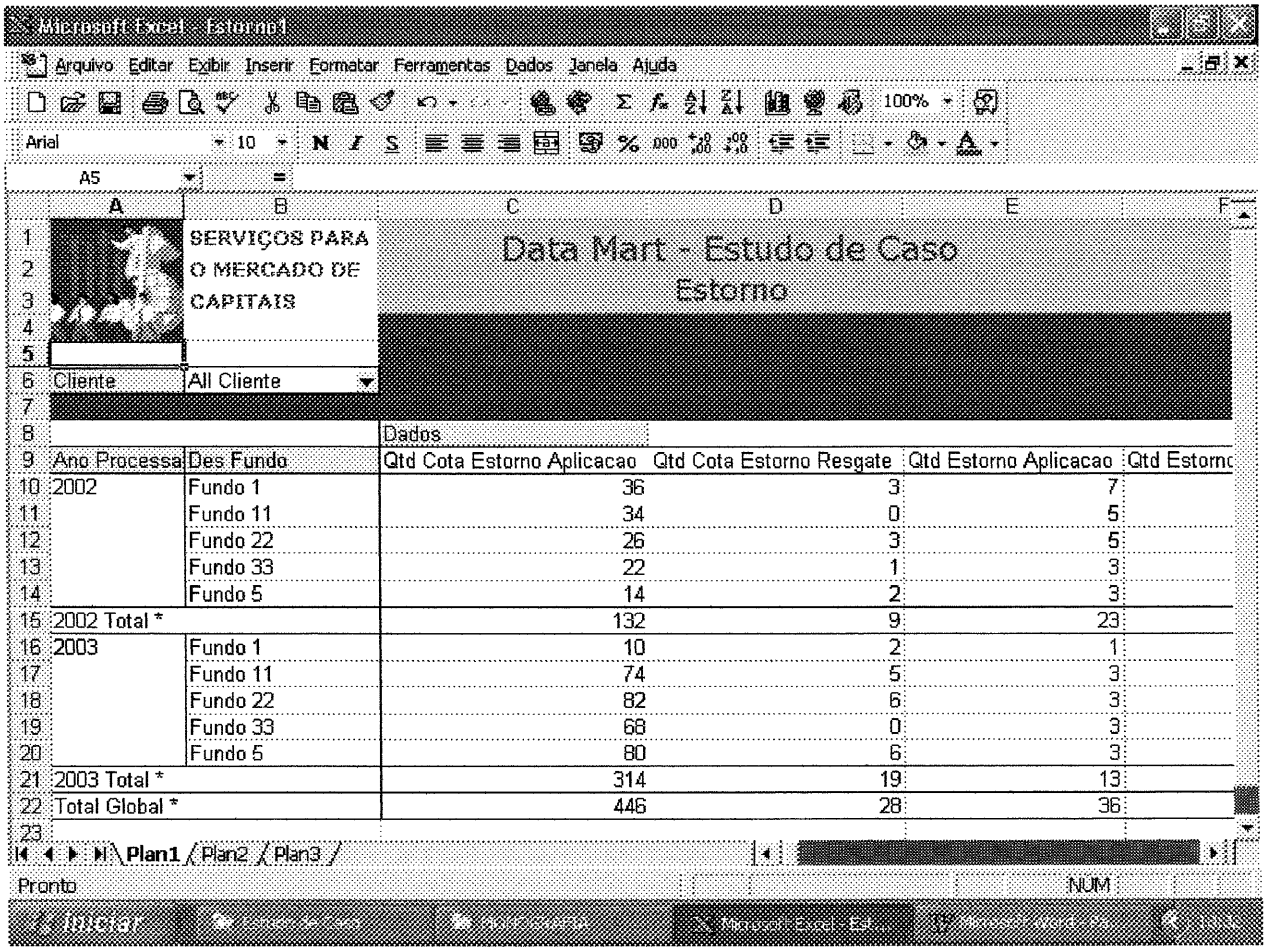
6. PATTON, Robert; OGLER, Jennifer. Projetando e Administrando Banco de Dados SQL Server 2000 como Servidor Enterprise. Rio de Janeiro: Alta Books, 2002.

7. PETKOVIC, Dusan. SQL Server 7: Guia Prático. São Paulo: Makron Books, 1999.

8. SINGH, Harry S. Data WareHouse – Conceitos, Tecnologias, Implementação e Gerenciamento. São Paulo: Makron Books, 2001.

ANEXOS

Anexo 1 – Planilha Estorno com as dimensões Fundo e Tempo



Microsoft Excel - Estorno1

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

SERVIÇOS PARA O MERCADO DE CAPITAIS

Data Mart - Estudo de Caso
Estorno

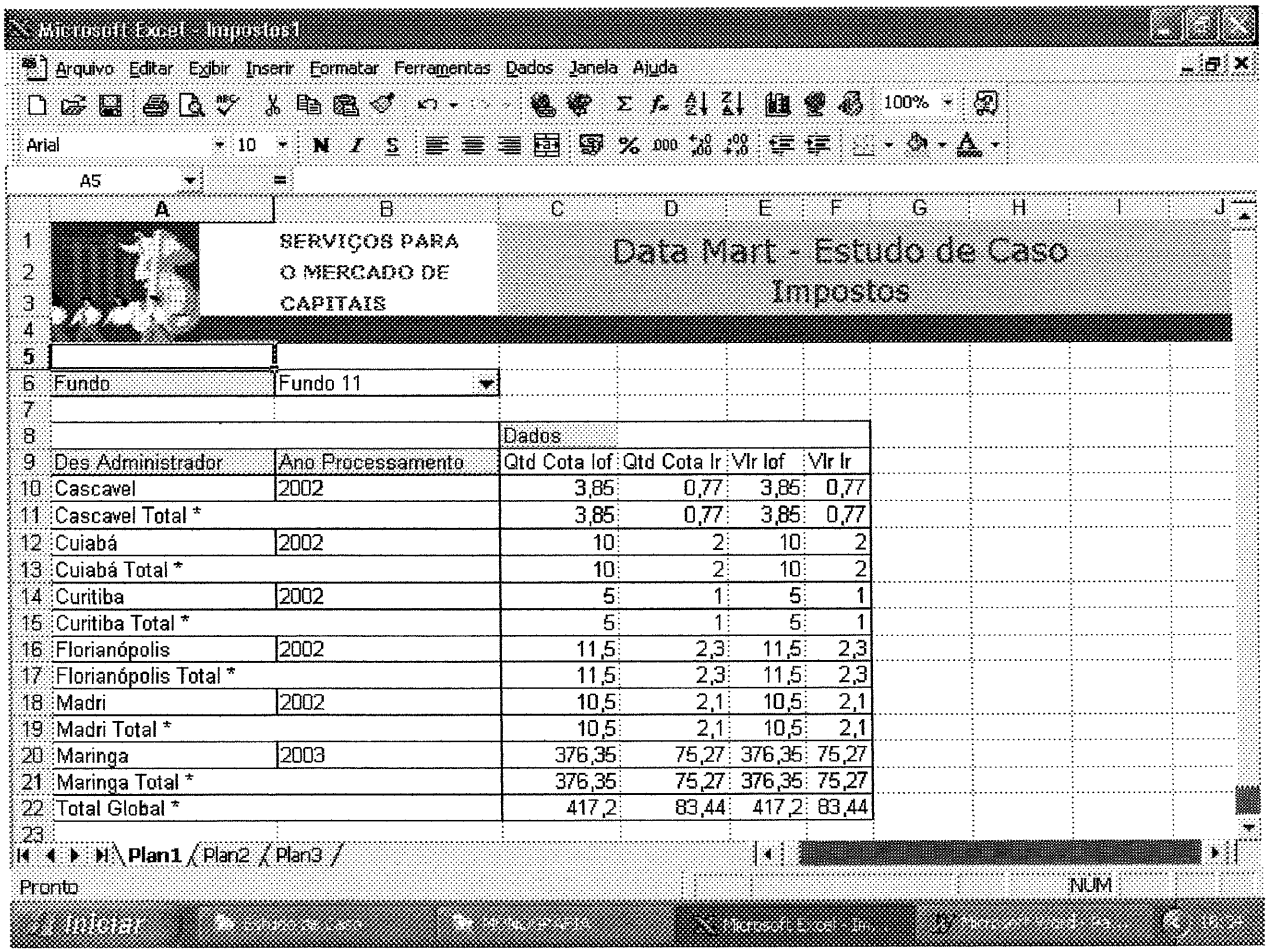
Cliente	All Cliente					
		Dados				
Ano Processa	Des Fundo	Qtd Cota Estorno Aplicacao	Qtd Cota Estorno Resgate	Qtd Estorno Aplicacao	Qtd Estorno	
2002	Fundo 1	36	3	7		
	Fundo 11	34	0	5		
	Fundo 22	26	3	5		
	Fundo 33	22	1	3		
	Fundo 5	14	2	3		
2002 Total *		132	9	23		
2003	Fundo 1	10	2	1		
	Fundo 11	74	5	3		
	Fundo 22	82	6	3		
	Fundo 33	68	0	3		
	Fundo 5	80	6	3		
2003 Total *		314	19	13		
Total Global *		446	28	36		

Plan1 / Plan2 / Plan3 /

Pronto NUM

Iniciar

Anexo 2 – Planilha Imposto com as dimensões Administrador e Tempo

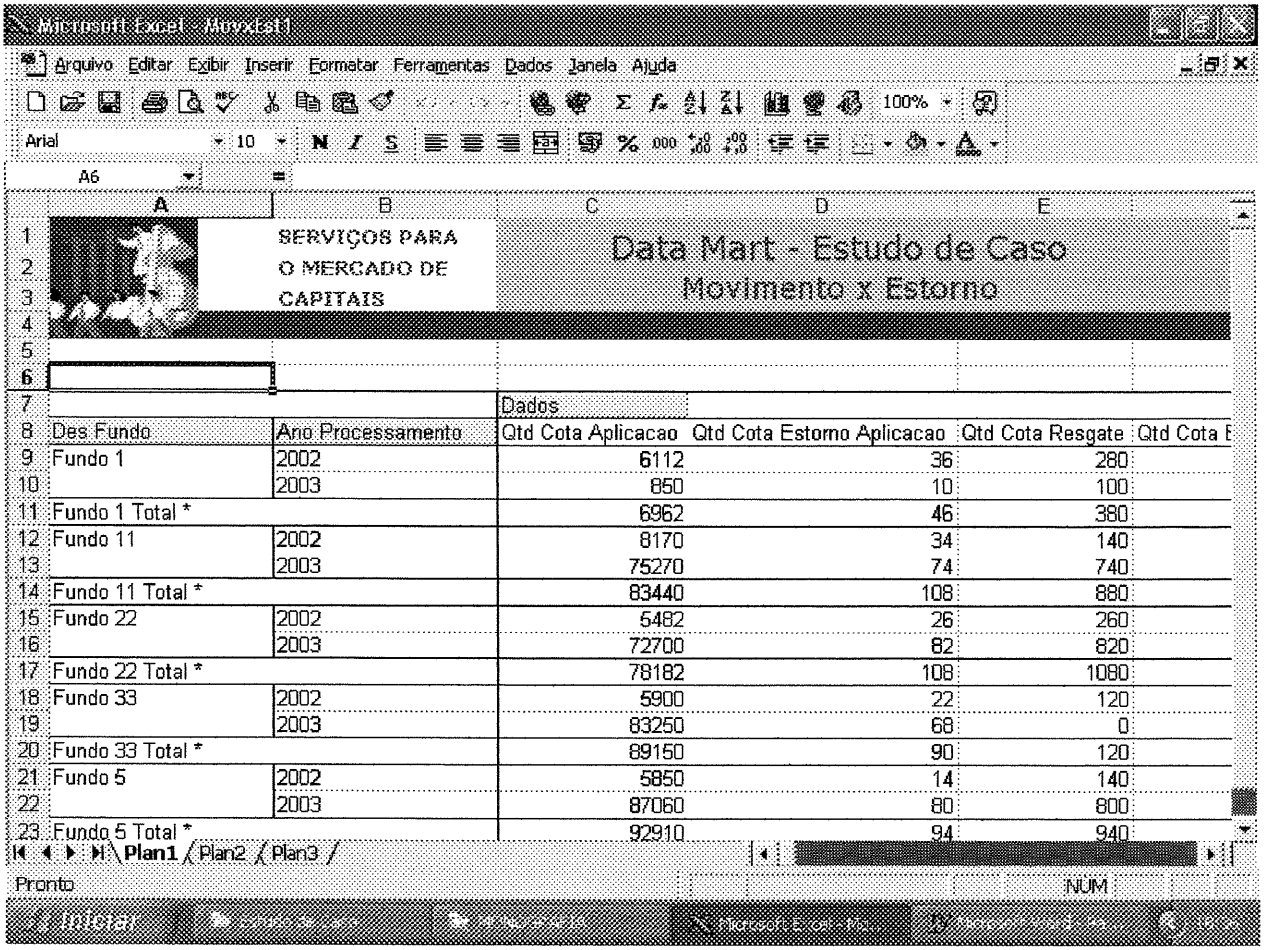


SERVIÇOS PARA O MERCADO DE CAPITAIS

Data Mart - Estudo de Caso
Impostos

Fundo	Fundo 11				
		Dados			
Des Administrador	Ano Processamento	Qtd Cota Iof	Qtd Cota Ir	Vlr Iof	Vlr Ir
Cascavel	2002	3,85	0,77	3,85	0,77
Cascavel Total *		3,85	0,77	3,85	0,77
Cuiabá	2002	10	2	10	2
Cuiabá Total *		10	2	10	2
Curitiba	2002	5	1	5	1
Curitiba Total *		5	1	5	1
Florianópolis	2002	11,5	2,3	11,5	2,3
Florianópolis Total *		11,5	2,3	11,5	2,3
Madri	2002	10,5	2,1	10,5	2,1
Madri Total *		10,5	2,1	10,5	2,1
Maringa	2003	376,35	75,27	376,35	75,27
Maringa Total *		376,35	75,27	376,35	75,27
Total Global *		417,2	83,44	417,2	83,44

Anexo 3 – Planilha Movimento x Estorno com as Dimensões Fundo e Tempo



SERVIÇOS PARA O MERCADO DE CAPITAIS

Data Mart - Estudo de Caso
Movimento x Estorno

	Des Fundo	Ano Processamento	Dados			
			Qtd Cota Aplicacao	Qtd Cota Estorno Aplicacao	Qtd Cota Resgate	Qtd Cota E
9	Fundo 1	2002	6112	36	280	
10		2003	850	10	100	
11	Fundo 1 Total *		6962	46	380	
12	Fundo 11	2002	8170	34	140	
13		2003	75270	74	740	
14	Fundo 11 Total *		83440	108	880	
15	Fundo 22	2002	5482	26	260	
16		2003	72700	82	820	
17	Fundo 22 Total *		78182	108	1080	
18	Fundo 33	2002	5900	22	120	
19		2003	83250	68	0	
20	Fundo 33 Total *		89150	90	120	
21	Fundo 5	2002	5850	14	140	
22		2003	87060	80	800	
23	Fundo 5 Total *		92910	94	940	