UNIVERSIDADE FEDERAL DE PELOTAS
INSTITUTO DE FÍSICA E MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Modelagem de um *Data Webhouse* voltado a Produção e Comercialização de sementes

ANA MARILZA DA ROSA PERNAS

Pelotas, março de 2003.

UNIVERSIDADE FEDERAL DE PELOTAS
INSTITUTO DE FÍSICA E MATEMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Modelagem de um *Data Webhouse* voltado a Produção e Comercialização de sementes

por

ANA MARILZA DA ROSA PERNAS

MONOGRAFIA apresentada ao Curso de Bacharelado em Ciência da Computação do Instituto de Física e Matemática da UNIVERSIDADE FEDERAL DE PELOTAS, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Profª Flávia Braga de Azambuja, MSc./UFPel
Co-orientadora: Profª Maria da Graça Alcântara, Dra./UFPel

Pelotas, março de 2003.

Modelagem de um *Data Webhouse* voltado a Produção e Comercialização de sementes

por

ANA MARILZA DA ROSA PERNAS

Monografia defendida e aprovada em 7 de março de 2003, pela banca examinadora constituída pelos seguintes integrantes:

_____

Profª. Mscª. Flávia Braga de Azambuja - Orientadora

_____

Profª. Drª. Maria da Graça Alcântara - Co-orientadora

_____

Prof. Dr. Ricardo de Azambuja Silveira

_____

Prof. Msc. Daniel Litchnow (UCPel)

A minha mãe, por sua constante dedicação e carinho.

Ao meu noivo, pelo apoio e compreensão que vem me dedicando durante o desenvolvimento deste trabalho e em todos os anos em que estamos juntos.

Agradecimentos

Primeiramente a minha mãe, Marilza, por estar sempre presente me apoiando e me fazendo acreditar que sou capaz de realizar meus objetivos.

Ao meu noivo, Anderson, por ter me apoiado durante todos os anos da graduação e, especialmente, na criação deste trabalho.

Aos meus irmãos: Agustin, Gustavo e Jorge, por serem exemplos a serem seguidos e pelo incentivo e confiança que sempre depositaram em minhas decisões.

Às professoras Flávia Azambuja e Maria da Graça, por tudo que me ensinaram através de suas experiências, por terem me orientado e acreditado no desenvolvimento deste trabalho.

Aos professores, pela colaboração de cada um para que o curso continue seu funcionamento.

Aos colegas, que tiveram comigo uma convivência quase diária durante esses quatro anos de faculdade.

Sumário

Lista de Abreviaturas

BD	Banco de Dados
BDR	Banco de Dados Relacional
DBMS	*Data Base Management System*
DW	*Data Warehouse*
HOLAP	*Hybrid On-Line Analytic Processing*
IP	*Internet Protocol*
ISP	*Information Service Provider*
MD	Mineração de Dados
MOLAP	*Multidimencional On-Line Analytic Processing*
ODBC	*Open Data Base Conectivity*
OLAP	*On-Line Analytic Processing*
OLTP	*On-Line Transaction Processing*
PDF	*Portable Document Format*
ROLAP	*Relational On-Line Analytic Processing*
SAD	Sistema de Apoio à Decisão
SGBD	Sistema Gerenciador de Banco de Dados
SIE	Sistema de Informações Empresariais
SQL	*Structured Query Language*

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho é destinado ao desenvolvimento de um modelo de *Data Webhouse*, com o objetivo de prover a infra-estrutura necessária para que empresários, agricultores ou setores ligados à pesquisa científica na área de sementes possam pesquisar dados de diversas fontes em uma única base de dados. Neste *Data Webhouse* são disponibilizadas informações a respeito de produção e comércio de culturas produzidas a partir de tecnologias de DNA recombinante e de métodos tradicionais, com novas tecnologias associadas, que receberam aprovação reguladora no MERCOSUL e em outros países.

Neste trabalho, é descrito o processo de construção de um ambiente de uso conjunto da tecnologia *Data Warehouse* com a facilidade do acesso a dados pela *Internet*, obtido com a modelagem de um *site Web* voltado a consultas, o qual foi disponibilizado em um Sistema de Apoio à Decisão para o mercado produtor de sementes. As informações contidas no *Data Webhouse* foram coletadas de bases de dados pertencentes a organizações públicas e privadas. Os dados coletados foram padronizados de forma planejada para que seus usuários obtivessem uma consulta direcionada e clara diretamente na *Internet*, o que facilita a tarefa de tomada de decisão por parte de empresários e comerciantes da área, pois visa disponibilizar em seu computador um ambiente para uso estratégico dos dados.

Palavras-chave: Banco de Dados, *Data Warehouse*, *Data Webhouse*, Sistemas de Informação, MERCOSUL, tecnologia de sementes.

Abstract

This work is destined to the development of a Data Webhouse model, with the objective to provide the necessary infrastructure so that business men , agriculturists or sectors in the scientific research on the seeds area can search a diverse sources in an only database. In this Data Webhouse are offered information about production and commerce of seeds produced from technologies of recombinant DNA and traditional methods, with new technologies associates, who had received regulating approval in the MERCOSUL and other countries.

This work describes the process of construction of an environment which combines the Data Warehouse technology with the easiness of data access from the Internet, gotten with the modeling of a Web site, which offer a Decision Support System to seeds producer market. The information contained in the Data Webhouse had been collected from public and private databases. The collected data had been standard to a planned form so that users got a directed and clean consultation in the Internet, becoming easy the task of taking decision from business men and traders of the area, therefore aim at to offer an environment for strategical use of the data.

Key-words: Data Base, Data Warehouse, Data Webhouse, Information Systems, MERCOSUL, technology of seeds.

1. Introdução

A partir da década de 1990, com a explosão da *Internet* e crescente facilidade de acesso a dados, as organizações começaram a perceber que fatores como diferencial competitivo e ascensão no mercado são obtidos por aqueles que aplicam estratégias para a utilização de seus dados operacionais.

Tendo em vista essa diferenciação nos resultados obtidos, quando efetuadas pesquisas em dados operacionais e pesquisas em dados integrados para um determinado objetivo, e o crescimento da *Internet*, começaram a surgir novas tecnologias para o armazenamento de dados, de forma a suprir esta necessidade de integração das informações e facilidade para o seu acesso. Neste contexto, foi criado o ***Data Webhouse***, o qual consiste de um ambiente de uso integrado da tecnologia de *Data Warehouse* (DW) com a facilidade do acesso a dados via *Web* [KIR 2000]. Segundo Inmon, um DW consiste na organização do grande volume de dados corporativos operacionais, agregando as informações realmente necessárias com o objetivo de facilitar os processos de busca, interpretação e ordenação por parte dos gerentes e diretores das empresas [INM 97].

A tecnologia de *Data Webhouse* surgiu a partir de um renascimento do DW, devido a necessidade de que o DW acompanhasse os avanços da tecnologia e as exigências dos usuários, que desejam obter informação de forma cada vez mais rápida e fácil [KIR 2000].

Segundo Kimball em [KIR 2000], a forma mais sucinta de descrever o processo de publicar os dados de um DW na *Web*, formando um *Data Webhouse*, é dizendo que não se está mais em um ambiente cliente/servidor. Está-se em um ambiente compatível com a *Web* e possuí-se mais camadas na atual arquitetura do que na anterior.

De acordo com todas estas modificações, altamente necessárias, pelas quais tem passado a tecnologia na última década, percebeu-se a importância de se ter os dados a respeito de sementes e plantas geneticamente modificadas, que receberam aprovação reguladora nos países do MERCOSUL, relacionados e à disposição em um único local, pois empresários, agricultores e, até mesmo, setores ligados à pesquisa científica já

reconhecem a necessidade de se ter uma visão ampla dos dados que dispõem para que possa efetuar, com maior facilidade, a tarefa de tomada de decisão.

No entanto, não existe atualmente um ambiente de pesquisa integrada de dados a respeito de sementes geneticamente modificadas para a obtenção de informações relevantes. Os dados necessários existem, mas encontram-se dispersos em diferentes bases de dados pertencentes a organizações públicas e privadas que cuidam do assunto, como Ministério da Agricultura do Brasil, Serviço Nacional de Proteção de Cultivares / MA – Brasil, Secretaria da Agricultura do Estado/ RS, Instituto Nacional de Semillas /INASE – Uruguai, Instituto Nacional de Semillas/INASE – Argentina, Departamento de Semillas/DISE/MA – Paraguai, Asociacion Nacional de Produtores de Semillas/ANPROS – Chile, Consejo Nacional de Semillas/CNS – Bolívia, IBGE, FAO - Food and Agricultural Organization, ASTA, American Seed Trade Association, e outros.

Para atender esta necessidade de otimização na apresentação dos dados a empresários, agricultores e pesquisadores da área de sementes, é que se mostra necessária à criação de um *Data Webhouse*, pois este fornece uma visão integrada dos dados e não está preso a uma aplicação local, podendo ser acessado, via *Internet*, por seus diversos usuários.

Com base no conceito de *Data Webhouse* e nas necessidades dos usuários em questão, este trabalho objetiva, principalmente, a modelagem de um DW, ou "Armazém de Dados", como é conhecido em português, para que seja a base de dados de um *site Web* voltado a consultas, o qual será disponibilizado na *Internet* em um Sistema de Apoio à Decisão para o mercado produtor de sementes, formando assim um sistema de *Data Webhouse*.

Este DW irá prover descrições completas para cada uma das culturas que receberam aprovação reguladora no MERCOSUL, incluindo dados sobre plantas produzidas a partir do uso de tecnologias de DNA recombinante, isto é, modificadas geneticamente ou transgênicas, e também plantas produzidas através de métodos tradicionais com novas tecnologias associadas.

Abaixo, encontra-se a forma como está organizado este trabalho:

Na Introdução é apresentado o ambiente de *Data Webhouse* e de *Data Warehouse*, juntamente com o problema abordado no trabalho e como se espera solucioná-lo.

No Capítulo 2 é apresentado o conceito, importância e uso dos Sistemas de Apoio a Decisão, também é apresentada toda a Fundamentação Teórica a respeito de *Data Warehouse*s e *Data Webhouse*s.

No Capítulo 3 é descrito, com mais detalhes, o problema a que se destina este trabalho e a maneira com que será tratado.

No Capítulo 4 é mostrada a forma com que a modelagem foi desenvolvida, a metodologia utilizada e a forma com que os dados foram transformados para serem aplicados ao DW proposto. Neste capítulo também é apresentado um protótipo, desenvolvido de acordo com a modelagem apresentada, e são mostrados dos resultados que foram obtidos com a modelagem através, principalmente, do que foi observado durante a criação do protótipo.

No Capítulo 5 é feita uma conclusão geral a respeito do trabalho desenvolvido e são apresentadas sugestões para trabalhos futuros.

2. Fundamentação Teórica

2.1 Sistemas de apoio à decisão

Segundo Inmon, as origens do processamento de Sistemas de Apoio à Decisão (SAD) remontam aos primórdios dos computadores, no início da década de 1960 o processamento era feito em aplicações baseadas em relatórios e programas. Com os anos, na medida em que o volume de dados foi crescendo, a tarefa de análise dos mesmos se tornava quase impossível [INM 97]. Ainda segundo o autor, a proliferação de arquivos mestres e a massiva redundância de dados apresentaram alguns problemas, tais como:

- A necessidade de sincronizar dados a serem atualizados.
- A complexidade de manutenção de programas.
- A complexidade de desenvolvimento de novos programas.
- A quantidade de hardware necessária para manter todos os arquivos mestres.

Com a crescente evolução da tecnologia e o armazenamento de dados em disco, surgiram os Sistemas Gerenciadores de Bancos de Dados (SGBD), com o objetivo de tornar o armazenamento e o acesso a dados em disco mais fáceis para o programador. Entretanto, mesmo com as facilidades trazidas pelos SGBDs, o volume de dados continuava crescendo e, como não havia o devido planejamento no armazenamento dos dados no Banco de Dados (BD), sua compreensão se tornava mais complexa.

Todos esses problemas enfrentados no passado, e que hoje ainda são enfrentados por muitos analistas e desenvolvedores de sistemas, tem origem no fato de que, tradicionalmente, a tecnologia da computação tem se empenhado em automatizar tarefas rotineiras, melhorando a eficiência de processos existentes e coletando dados. Infelizmente, mesmo ocorrendo essa grande coleta de dados, até recentemente, o valor destes foi difícil de se compreender e usar [BER 99].

Através de um sistema de BD organizado, é possível transformar uma grande base de dados em vantagem competitiva, através da elaboração de um sistema que atue

no sentido de agrupar informações que demonstrem alterações de padrões. Por exemplo, uma cadeia de supermercado, que analisando as informações referentes às saídas de mercadorias, identifica um crescimento na venda de carnes nos finais de semana. Através de um relacionamento dos dados, a rede descobre que grande parte das pessoas que compra carne, também compra carvão e bebidas. Com base nessas informações a rede pode traçar estratégias de vendas mais elaboradas. Pode inclusive, decidir por evitar colocar o carvão em oferta nos finais de semana [ANN 2003].

Com a evolução dos SADs e, mais especificamente, com o surgimento do DW, organizações podem utilizar dados já coletados para obter grande retorno no investimento. Neste contexto, o DW consiste de um grande depósito de informações, encontrado no cerne do processamento SAD, visando efetiva integração de bases de dados operacionais em um ambiente que permita o uso estratégico dos dados [INM 97]. Ele forma assim uma base de dados na qual é efetuado um tratamento adequado a informação, o qual pode habilitar a descoberta e exploração de empreendimentos importantes.

Existem várias razões que justificam a construção de um DW, dentre as quais podem ser citadas:

- A necessidade de se tomar decisões de forma rápida, correta e clara, a partir do uso de todos os dados disponíveis;
- O fato de os usuários de sistemas de informação serem especialistas em negócios e em definir estratégias, não em computação;
- O aumento rápido do volume de dados, que afeta o tempo de resposta e incontestavelmente a habilidade em compreender o conteúdo das informações;
- O aumento diário da competição nas áreas de tecnologia da informação e de inteligência empresarial, bem como o valor dado às informações.

2.2 O ambiente de *Data Warehouse*

O ambiente de DW consiste de um grande BD, criado de forma a armazenar, estruturalmente, dados vindos de diferentes fontes em um ou mais repositórios, estes dados podem ser de sistemas locais ou não, planilhas, arquivos textos, etc.

No DW são armazenados, em sua maioria, dados derivados, também são armazenados dados primitivos, mas em menor quantidade, estes são adicionados quando necessário, juntamente com a sua variação no tempo. De acordo com os dados apresentados na tab. 2.1, dados primitivos atendem à atividade funcional, enquanto que dados derivados atendem à atividade gerencial [INM 97]. Para se entender melhor o que são dados primitivos e dados derivados, na tab.2.1 encontra-se uma relação de suas principais diferenças.

TABELA 2.1 - Principais diferenças entre dados primitivos e derivados.

Dados Primitivos/Dados Operacionais	**Dados Derivados/Dados SAD**
Baseados em aplicações	Baseados em assuntos ou negócios
Detalhados	Resumidos ou refinados
Exatos em relação ao momento do acesso	Representam valores de momentos já decorridos ou instantâneos
Atendem à comunidade funcional	Atendem à comunidade gerencial
Podem ser atualizados	Não são atualizados
São processados repetitivamente	Processados de forma heurística
Requisitos de processamento conhecidos com antecedência	Requisitos de processamento não são conhecidos com antecedência
Compatíveis com o ciclo de vida de sistemas de BD operacionais clássicos	Ciclo de vida completamente diferente
A performance é fundamental	Performance atenuada
Acessados uma unidade por vez	Acessados um conjunto por vez
Voltados para transações	Voltados para análise
O controle de atualizações é atribuição de quem tem a posse	O controle de atualizações não é problema
Alta disponibilidade	Disponibilidade atenuada
Gerenciados em sua totalidade	Gerenciados por subconjuntos
Não contemplam a redundância	A redundância não pode ser ignorada
Estrutura fixa; conteúdos variáveis	Estrutura flexível
Pequena quantidade de dados usada em um processo	Grande quantidade de dados usada em um processo

Dados Primitivos/Dados Operacionais	Dados Derivados/Dados SAD
Atendem às necessidades cotidianas	Atendem às necessidades gerenciais
Alta probabilidade de acesso	Baixa, ou modesta probabilidade de acesso

Fonte: INMON. Como construir o Data Warehouse. p.18

2.2.1 Características do ambiente

As principais características do DW são bem definidas e sobre elas concordam os mais renomados autores da área. Segundo Inmon, o DW é caracterizado por ser baseado em assuntos, integrado, não-volátil e variável em relação ao tempo [INM 97]. Existe outra característica importante que será abordada, a granularidade dos dados.

- **Baseado em assuntos**: enquanto os sistemas clássicos são modelados para atender às aplicações de empresas, toda a modelagem do DW é feita visando-se atender aos assuntos importantes a um determinado grupo de usuários, esses assuntos são as informações relativas a uma certa área estratégica de uma empresa ou organização.
- **Integrado**: os dados a serem colocados no DW normalmente são vindos de diversas aplicações diferentes, e cada projetista de aplicação pode ter padronizado a entrada de dados de várias formas, portanto as aplicações não apresentam coerência em termos de codificações, convenções de atribuição de nomes, unidades de medidas e outros conteúdos. Por este motivo, quando os dados são incluídos no DW, devem passar por uma classificação e padronização, para que se tornem integrados, livres de inconsistências e independentes da aplicação de origem.
- **Não-volatilidade de dados**: no ambiente de DW, diferente do ambiente operacional, geralmente não ocorre tarefa de atualização dos dados, estes sofrem a sua carga inicial e são consultados, sem sofrer modificações, o que ocorre é a sua passagem por filtros e o resumo de alguns dados antes de serem inseridos no DW. A maior parte dos dados é física e radicalmente alterada quando passam a fazer parte do DW, do ponto de vista de integração, não são mais os mesmos dados do ambiente operacional. À luz destes fatores, a redundância de dados

entre os dois ambientes raramente ocorre, resultando em menos de um por cento de duplicações [INM 97].

- **Variável em relação ao tempo**: o DW mantém o histórico dos seus dados por um período de tempo muito superior ao dos sistemas de bancos de dados operacionais tradicionais. Isso ocorre devido ao próprio conceito de DW, pois ele existe com o objetivo de apoiar a tarefa de tomada de decisão, para isso é necessário que os gerentes analisem o comportamento dos dados durante um grande período de tempo, para que apóiem suas decisões em fatos e não em intuições.
- **Granularidade**: representa o grau de detalhe ou de resumo que os dados pertencentes ao DW contém. Quanto maior for o grau de detalhamento do dado, menor será o seu grau de granularidade. Quanto menor o grau de detalhamento do dado, maior será o seu grau de granularidade. Quando o grau de granularidade é muito alto, o espaço em disco e o número de índices necessários para uma consulta no DW são bem menores, porém a uma perda de otimização quando se objetiva consultas detalhadas. Por essa razão a granularidade é a principal questão de projeto, o volume de dados contidos no DW é balanceado de acordo com o nível de detalhe de uma consulta [INM 97].

2.2.2 Aspectos de arquitetura

A base para que se desenvolva um produto ou projeto é a sua arquitetura. A arquitetura de dados serve para que seja estabelecida e compreendida a movimentação dos dados dentro de um sistema e qual o seu objetivo como um todo. No ambiente de DW, o objetivo é a transformação do dado em informação, para que isto ocorra, a arquitetura deve ser proposta de forma a representar as estruturas de dados, as comunicações, os processamentos e os resultados que serão apresentados aos usuários.

Existem várias propostas de arquiteturas para o DW, mas, de uma forma geral, é composta: pelas bases de dados operacionais, que consistem dos dados externos ao DW; pela sua área interna, onde ocorrem todos os processos de organização dos dados; e pela área física do DW, que é onde os dados estão realmente armazenados para consulta direta ou indireta dos usuários finais.

A área física do DW normalmente se encontra centralizada em um único local, devido à dificuldade do acesso aos dados "espalhados" em diversos *sites* locais, ou pelos dados contidos no DW serem necessários apenas na matriz de uma empresa, ou por não haver um volume tão grande de dados, tal que um único repositório de dados faz sentido [INM 97]. Mas pode também se encontrar de forma distribuída, no caso, por exemplo, de uma empresa que possui uma matriz, com processamento operacional global, e diversas filiais, com processamentos locais autônomos, onde ocasionalmente e para certos tipos de processamentos os dados são enviados das filiais para a matriz [INM 97].

Diferentemente da forma distribuída apresentada, a área física pode estar organizada não somente em um único DW, mas em vários, chamados de *Data Mart*s, subconjuntos lógicos do DW, geralmente tratados como um DW setorial [KIR 96]. Os *Data Marts* muitas vezes são vistos como uma alternativa ao uso do DW, pois levam menos tempo para serem desenvolvidos e implementados. Uma perspectiva *top-down* considera que um DW completo e centralizado deve ser desenvolvido antes que partes dele, sumarizadas, possam ser derivadas na forma de *Data Marts*. Enquanto que, uma perspectiva *botton-up* considera que um DW possa ser composto a partir de *Data Marts* previamente desenvolvidos [CAM 2002].

Outro aspecto importante com relação à arquitetura do DW é a existência dos metadados e a sua função no ambiente de DW. Os metadados armazenam o significado de cada dado, isto é, eles são dados sobre dados. A sua existência é fundamental, pois através deles a utilização mais produtiva do DW é alcançada e permitem que o usuário final/analista de SAD navegue pelas possibilidades [INM 97]. É importante que o DW possua uma variedade de metadados disponíveis, para que os usuários finais sejam capazes de acessar dados do DW sem a necessidade de saberem onde ou como os dados estão armazenados. Segundo Inmon, tipicamente, os aspectos sobre os quais metadados mantém informações são [INM 97]:

- A estrutura dos dados segundo a visão do programador.
- A estrutura dos dados segundo a visão do analista de SAD.
- A fonte de dados que alimenta o DW.
- A transformação sofrida pelos dados no momento de sua migração para o DW.
- O modelo de dados.

- O relacionamento entre o modelo de dados e o DW.
- O histórico de extrações.

Os metadados encontram-se armazenados no repositório de metadados, e são gerenciados por ele, podendo estar centralizados em um único local ou distribuídos, dependendo de sua forma de utilização.

2.2.3 Dimensionalidade dos dados

Nos Bancos de Dados Relacionais (BDRs) convencionais, a redundância dos dados é evitada, sendo aceita somente de forma controlada nos casos em que é realmente necessária. Esta redundância é eliminada através de processos de normalização, onde cada tabela do BD que possua dados redundantes é dividida em duas tabelas distintas, originando, deste modo, apenas tabelas contendo dados não-redundantes.

A normalização das tabelas traz benefícios nos casos em que muitas transações são efetuadas, pois estas se tornam mais simples e rápidas. Já no caso dos DWs ocorre o contrário, as transações operam sobre um grande volume de dados e não são simples nem freqüentes, não sendo conveniente a normalização das tabelas, pois no ambiente de DW ocorrem poucas transações concorrentes e cada transação acessa um grande número de registros.

Outro ponto que distingue os BDRs dos DWs está relacionado a modelagem dos dados. Os DWs não utilizam o modelo entidade - relacionamento, como ocorre com os BDRs, pois este modelo é utilizado no projeto de BDs com dados não redundantes. A modelagem lógica usada para o projeto de DWs é chamada de **modelagem dimensional**.

Diferente do modelo relacional, o modelo dimensional é muito assimétrico. Nele existe uma grande tabela "dominante" no centro do esquema, a qual se conecta com as demais através de múltiplas junções, enquanto que o restante das tabelas se liga à tabela central através de uma única junção. A tabela central é chamada de **tabela de fatos** e as demais tabelas são chamadas de **tabelas de dimensões** [KIR 96].

Um exemplo de modelo dimensional é apresentado na fig. 2.1. Esta é uma modelagem que representa o caso tratado neste trabalho, de um DW de informações

sobre a produção de sementes geneticamente modificadas no MERCOSUL, medindo o seu desempenho (quantidade produzida a cada colheita) através do tempo. Na tabela de fatos, cada registro representa o total de produção de uma espécie de semente específica em um país específico do MERCOSUL, e qualquer outra combinação de *espécie*, *pais* ou *ano* representa um registro diferente na tabela de fatos.

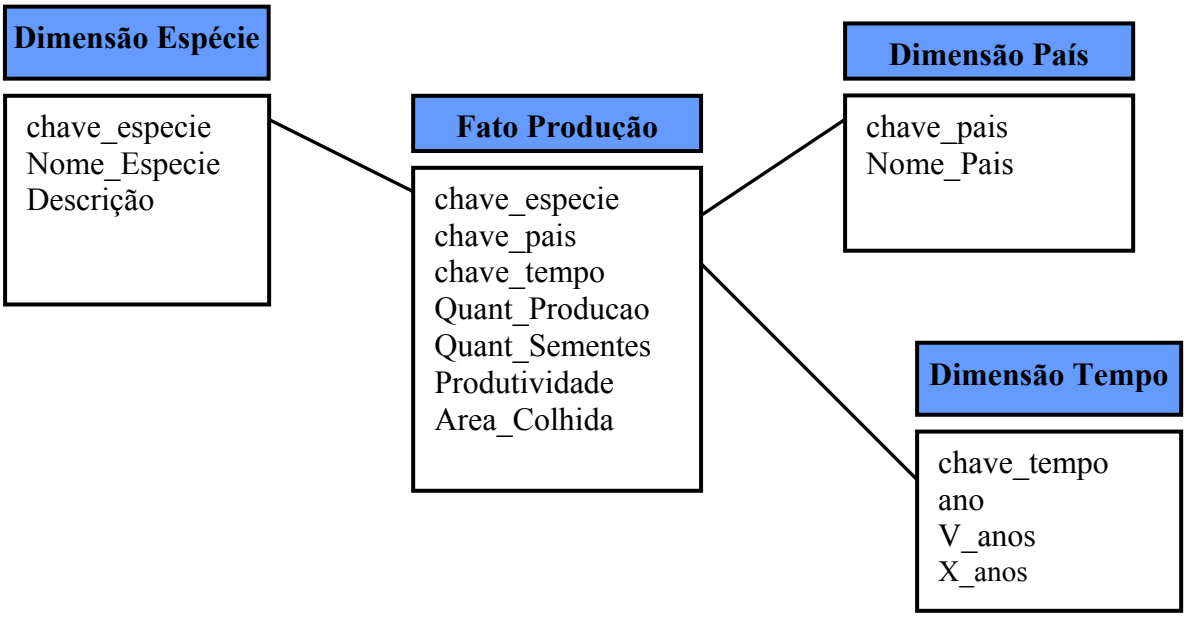

FIGURA 2.1 - Exemplo de um modelo dimensional de sementes do MERCOSUL.

Normalmente, a tabela de dimensões contém uma única chave primária e vários atributos que descrevem essa dimensão com detalhes. Na tabela de fatos, a chave primária é a combinação das demais chaves primárias das tabelas de dimensão, constituindo-se, assim, de várias chaves estrangeiras, de acordo com o número de dimensões. Os dados pertencentes à tabela de fatos são, normalmente, numéricos.

O modelo dimensional é também chamado de **esquema estrela** (*star schema*), devido ao formato com que são dispostas as tabelas do diagrama (o que pode ser percebido na fig. 2.1), com a tabela de fatos no centro e um conjunto de tabelas de dimensão nas extremidades.

Ainda dentro do modelo dimensional, outro tipo de esquema também é muito usado na construção lógica do DW, que é o chamado **esquema floco de neve** (*snowflake schema*), o qual consiste em uma extensão do esquema estrela, onde cada uma das extremidades da estrela passa a ser o centro de outras estrelas. Isto ocorre devido à necessidade de normalização das tabelas de dimensão, pela construção de hierarquias nas colunas das tabelas. No caso do exemplo da fig. 2.1, a dimensão tempo

possui uma hierarquia, onde *tempo* contém *V_anos*, e *V_anos* contém *anos* (fig. 2.2). Estes relacionamentos adicionais geram mais tabelas no esquema floco de neve.

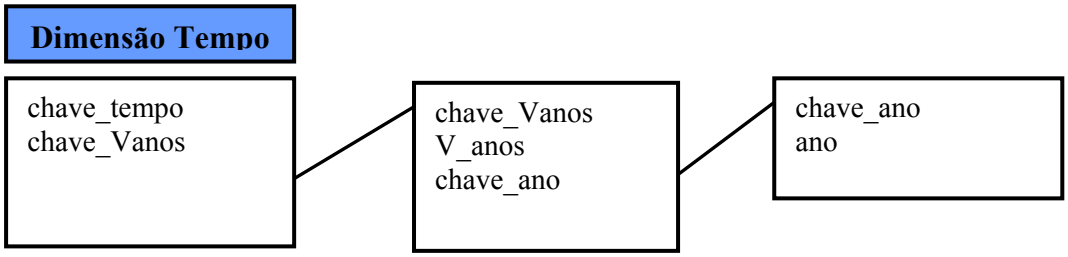

FIGURA 2.2 - Esquema floco de neve da dimensão Tempo.

No esquema floco de neve, cada um dos relacionamentos "muitos-para-um" é dividido em tabelas distintas. Apesar deste esquema parecer interessante aos olhos de muitos cientistas da computação, o seu uso não é recomendável, devido ao impacto causado ao usuário final pela complexidade deste tipo de estrutura. Um fato que levaria os projetistas ao uso deste método seria o ganho de espaço de armazenamento em disco, mas, através de estudos, constatou-se que o ganho em armazenamento que o esquema floco de neve oferece é irrelevante, não justificando seu uso [KIR 96].

2.2.4 Construção do *Data Warehouse*

A tarefa de construção do DW deve ser iniciada através da escolha da melhor estratégia, seguida pela modelagem dimensional, funcional e técnica e, finalmente, pela construção e manutenção do DW. Toda esta complexidade de aquisição, tratamento e atualização de informações é que deverá permitir às empresas determinar a necessidade e interesse de otimização na realização destes processos, através da utilização de ferramentas para análise ou padronizacão dos dados.

Uma etapa fundamental a ser executada nesta tarefa é a migração de dados de diferentes bases operacionais para uma base analítica. O principal motivo para a execução desta combinação de informações de diversas aplicações fontes é a possibilidade de cruzar referencias de dados. Ao processo de extrair, transformar e integrar toda a informação empresarial dá-se o nome de ***Data Warehousing*.**

Na construção é importante se definir o tipo de modelagem que será utilizado, para isso, a chave está em se descrever "o que" e "como" será feito o projeto de DW. A

modelagem de um DW exige a compreensão de conceitos de negócio mais abrangentes e complexos do que a modelagem de um sistema tradicional. É necessário que o implementador compreenda completamente e consiga transmitir na implementação do DW todas as idéias, conceitos e conhecimentos necessários aos usuários de sistemas de informação, para que os dados armazenados sejam realmente eficazes no processo de tomada de decisão. Segundo Yin em [YIN 89], a construção total do DW só está completa após seis meses de trabalho e, durante este período, devem haver reuniões semanais entre a equipe de planejamento do projeto, monitoração e gerência, para que o desenvolvimento do DW aconteça com ampla responsabilidade.

2.2.5 Extração, transformação e carga dos dados - ETC

Esta etapa é uma das mais críticas da construção de um DW, pois envolve toda a movimentação dos dados. A mesma se dá basicamente em três passos, conhecidos como ETC: Extração, Transformação (passo este que inclui a limpeza dos dados) e Carga dos dados.

A primeira tarefa a ser executada é a extração dos dados. Quando a operação de extração ocorre a partir de fontes externas e heterogêneas, geralmente é feita através de *gateways* e Interfaces padrão do tipo ODBC (*Open Data Base Conectivity*) - padrão para acesso a BDs do SQL Access Group Consortium adotado pela Microsoft, ou outras. Na abordagem utilizada em DWs para possibilitar o acesso a dados em fontes heterogêneas, é prevista a integração através da migração dos dados para novos SGBDs, o que caracteriza a materialização dos repositórios heterogêneos.Mas existem outros tipos de abordagens, os quais propõem a integração dos dados e das bases de dados apenas durante o processamento das consultas, somente selecionando resultados vindos destas consultas, sem materialização, isto é, sem a criação de réplicas dos repositórios locais integrados [KIM 95].

Desta última abordagem apresentada fazem parte os Sistemas Interoperáveis, que são os Mediadores e *Middlewares*, módulos de software que exploram o conhecimento representado por um conjunto ou subconjunto de dados, gerando informações para aplicações residentes em uma camada superior, provendo uma interface uniforme para o acesso aos dados [WIE 99].

Em sistemas de BDR, os dados são extraídos para um arquivo ou enviados para um aplicativo de solicitação por meio de ferramentas e aplicações utilizando SQL (*Structure Query Language* – Linguagem de Consulta Estruturada), uma linguagem que utiliza uma combinação de construtores em álgebra e cálculo relacional para a manipulação de dados em BDs, possuindo recursos para: consulta, modificação, definição da estrutura de dados e especificação de restrições de segurança [SIL 99]. Entretanto, se a formatação dos arquivos não é conhecida é necessário geração de relatório ou criação de arquivo para descarregar os dados do sistema de produção.

Em seguida, passa-se para o segundo passo do processo, o de limpeza e transformação dos dados. A limpeza consiste da análise dos dados obtidos com a finalidade de eliminar inconsistências, armazenamentos feitos de forma arbitrária apenas para preencher campos obrigatórios, e outros problemas. Pois na maioria dos casos a fonte da qual os dados foram retirados não é conhecida, portanto não se pode confiar totalmente na veracidade do seu conteúdo.

O processo de limpeza livra os dados de problemas que passaram despercebidos nos sistemas de origem, como códigos inválidos e preenchimento de vários campos com valores incompatíveis entre si. Normalmente as ações de correção das anomalias encontradas ocorrem através de rotinas que listam estes dados para que uma pessoa responsável procure solucionar as pendências caso a caso, corrigindo inclusive a base de dados original.

Em seguida, ainda no segundo passo do processo de ETC, é feita a transformação dos dados, a qual ocorre devido a vários motivos. Um deles é a necessidade de padronização dos dados, pois, uma vez que a origem dos dados pode ser de sistemas diferentes, isto é, de fontes de dados heterogêneas, é necessário padronizar os diferentes formatos, pois campos iguais em duas bases de dados distintas podem ter diferentes formatações para entrada de dados. É necessário, antes do momento da carga no DW, haver uma padronização, ou seja, quando o usuário for executar uma consulta não poderá ver informações iguais em formatos distintos.

Um exemplo simples desta falta de integração dos dados é a codificação para gênero feminino e masculino. Em uma base de dados, a entrada deste dado pode estar codificada em “m/f” e em outra base de dados em “h/m”, é preciso a padronização da entrada deste dado, não importa como será feita esta padronização no DW, deste que seja de forma coerente, como, por exemplo, não especificando a codificação como “0/1”.

Outro motivo para a transformação dos dados é o tratamento de valores perdidos, pois muitas vezes o dado a ser carregado no DW pode estar, devido a vários motivos, incompleto ou contendo valores que não podem ser manipulados corretamente. Por esta razão é importante que, durante a fase de transformação, sejam atribuídos valores padrão para dados perdidos ou corrompidos. E também que seja estabelecida uma forma para que os usuários do DW tomem conhecimento destes valores padrão.

O terceiro passo a ser executado é a carga dos dados. A carga do DW é uma operação efetuada por processos de carga/inserção específicos de cada SGBD ou por processos independentes de carga rápida (*fastload*) - tecnologia que obtém tempos de carga mais rápidos através do pré-processamento dos dados e ignorando operações de verificação de integridade e de registro das operações efetuadas. Esta tecnologia substitui uma função específica de carga do SGBD.

Existem três tipos de carga que podem ser feitos do ambiente operacional para o DW [INM 97]:

- O carregamento de dados históricos, o qual, como regra, representa um desafio menor por não ser feito com freqüência.
- O carregamento de dados de valor corrente no ambiente operacional, que também não constitui grande desafio por precisar ser feito apenas uma vez.
- O carregamento de alterações do DW a partir de atualizações que tenham ocorrido no ambiente operacional desde a última atualização do DW. Este constitui o maior desafio ao arquiteto de dados, pois o rastreamento eficiente e o tratamento dessas alterações não são tarefas fáceis de serem realizadas.

Dois fatos importantes a serem acrescentados, que devem ser levados em conta no momento da carga de alterações no DW, são: primeiro, os dados operacionais que irão ser acrescentados no DW devem sofrer modificações antes de sua carga, pois como são valores exatos no momento atual, devem ser vinculados a um elemento de tempo; segundo, a necessidade de efetuar a condensação dos dados antes de serem inseridos no DW, de outro modo, o volume de dados contidos no DW logo ficará grande demais para ser controlado.

2.2.6 Acesso aos dados

Depois de projetado o escopo do projeto e construído o ambiente na qual os dados serão colocados é que são definidas as técnicas que serão utilizadas para efetuar a comunicação entre as bases de dados e os usuários finais. Elas devem ser de fácil utilização, porque não serão utilizadas por profissionais da área técnica, mas precisam dar agilidade suficiente no acesso às informações necessárias para o processo de tomada de decisão e definição de estratégias.

Para o acesso aos dados podem ser utilizadas diversas técnicas, como as que estão apresentadas na tab. 2.2, de acordo com o objetivo que se tem para o DW.

Na tab. 2.2 são descritas as ferramentas mais utilizadas para o acesso aos dados em um DW.

TABELA 2.2 - Ferramentas para acesso aos dados no DW.

Tipo de Ferramenta	Questão Básica	Exemplo de Resposta	Usuário típico e suas necessidades
Pesquisa e Relatórios	"O que aconteceu?"	Relatórios mensais de vendas, histórico do inventário.	Dados históricos, habilidade técnica limitada.
OLAP	"O que aconteceu e por quê?"	Vendas mensais versus mudança de preço dos competidores.	Visões estáticas da informação para uma visão multidimensional; tecnicamente astuto.
SIE	"O que eu preciso saber agora?"	Memorandos, centros de comando.	Informações de alto nível ou resumidas; pode não ser tecnicamente astuto.
Mineração de Dados	"O que é interessante?" "O que pode acontecer?"	Modelos de previsão.	Tendências e relações obscuras entre os dados; tecnicamente astuto.

Fonte: Revista Byte Brasil, citada por CAMPOS, 2003.

Analisando-se a tab. 2.2, percebe-se que quando os objetivos exigem ferramentas que suportem perguntas mais estruturadas e complexas, as ferramentas mais utilizadas são OLAP (*On-Line Analytic Processing*) e Mineração de Dados (MD). Estas serão explicadas nas seções a seguir.

2.2.6.1 Ferramentas OLAP

As ferramentas OLAP foram criadas para prover o processamento analítico on-line das informações de uma empresa/corporação, contrastando com uma ferramenta anteriormente usada, a OLTP (*On-Line Transaction Processing*), para, como o próprio nome sugere, a análise de qualquer processo transacional de uma empresa a partir de dados estáticos, geralmente encontrados em bases de dados relacionais.

O uso de ferramentas OLAP se baseia na modelagem dimensional dos dados do DW, pois, através da divisão das hierarquias dos dados em dimensões, é possível que os dados sejam apresentados e analisados sob a ótica do gerente ou do tomador de decisão, facilitando a análise de dados através de sumarizações das informações contidas no modelo, aliando esta análise à possibilidade de visualizar qualquer intervalo de tempo definido no DW. Entretanto, o fato de refletirem a representação da realidade dos dados sob a ótica de quem irá analisá-los torna a OLAP uma solução não imediata, pois configurar o programa de OLAP e ter acesso aos dados requer uma clara compreensão dos modelos de dados da empresa e das funções analíticas necessárias aos executivos e outros analistas de dados [CAM 2003].

Típicas aplicações de negócios para as ferramentas OLAP incluem o desempenho e proficiência de produtos, a eficácia de um programa de vendas ou de uma campanha de *marketing*, a realização da previsão das vendas, e capacidade de planejamento [BER 97].

As mais freqüentes funcionalidades oferecidas pelas ferramentas OLAP são [KIR 96]:

- **Tabelas Cruzadas**: são as tradicionais tabelas eletrônicas, a diferença é que os dados são apresentados em planilhas com mais de duas dimensões, normalmente quatro ou mais;
- ***Drill-across***: é a requisição de dados das tabelas de dimensão com o valor das consultas modificado, ocorre quando o usuário pula um nível intermediário dentro de uma mesma dimensão. Por exemplo, a "Dimensão Tempo" mostrada na fig. 2.2 é composta por *V_anos* e *ano*, o usuário executará um *Drill-across* quando, numa consulta, passar de *V_anos* direto para *ano*;
- ***Drill-down***: aumento do nível de detalhe da informação, por parte do usuário, para solicitar uma visão mais detalhada em um conjunto de dados;

- ***Drill-up***: é o contrário do *Drill-down*, ocorre quando o usuário diminui o nível de detalhe da informação, solicitando uma visão menos detalhada de um conjunto de dados;
- ***Drill-trought***: é parecido com o *Drill-across*, mas neste o usuário passa de uma informação contida em uma dimensão para outra dimensão. Por exemplo, na fig. 2.1, o usuário está analisando a informação através da "Dimensão Tempo" e, em seguida, opta por passar a analisar através da "Dimensão Espécie";
- ***Slice-dice***: é a descrição padrão para a habilidade de acessar os dados do DW através de qualquer uma das dimensões de forma igual. Serve para modificar a posição de uma informação, alterar linhas por colunas de maneira a facilitar a compreensão dos usuários e mudar de dimensão sempre que necessário;
- ***Pivoting*** (pivoteamento): é a mudança do arranjo das linhas e colunas em um relatório tabular, onde freqüentemente as linhas ou as colunas são derivadas de dimensões diferentes. É a inversão dos eixos das dimensões, para obter-se novas visões de consultas.

Existem divisões dentro da OLAP, usadas de acordo com o BD a ser explorado. A ROLAP (OLAP relacional) é ligada aos conceitos básicos de BDRs para a análise dos dados, transformando consultas em rotinas SQL padrão. A partir dela, o usuário recebe resultados cruzados de tabelas em forma de planilha multidimensional ou de outra forma que suporte a rotação, *Drill-down* e manipulação [CAM 2003].

A MOLAP (OLAP multidimensional) difere da ROLAP no armazenamento dos dados. Na MOLAP os dados são processados a partir de um servidor multidimensional, onde o acesso aos dados ocorre diretamente no BD. Através dela o usuário trabalha, monta e manipula os dados diretamente no servidor, o que traz benefícios com relação à performance que os usuários podem atingir, mas é uma modelagem mais cara para a aquisição, por ser elaborada para um BDMD (Banco de Dados Multidimensional).

Existe também a HOLAP (OLAP híbrido), que consiste de uma mistura das tecnologias usadas na ROLAP e na MOLAP. Nela os dados de agregação, isto é, dados de baixo nível da tabela de fatos que são resumidos e armazenados em tabelas intermediárias para agilizar as consultas, são armazenados em MOLAP, enquanto que

os dados de base são armazenados no BDR. Então, em consultas que utilizam resumos, a HOLAP é análoga ao MOLAP, enquanto que, nas consultas aos dados de base, não é necessário que os usuários manipulem diretamente os dados, podendo interagir através de consultas simples, como no ROLAP. Além disso, o BD pode ser modelado ou como um BDMD ou como um BDMR (Banco de Dados Multirelacional – BD relacional projetado para aceitar propriedades multidimensionais).

2.2.6.2 Ferramentas de Mineração de Dados

Com o surgimento e o crescimento do DW, as ferramentas de Mineração de Dados (MD) ganharam grande interesse por parte do mercado, pois a tecnologia de DW somente é usada plenamente se possuir boas ferramentas na exploração dos seus dados.

Como o DW possui bases de dados bem organizadas e consolidadas, as ferramentas de MD ganharam grande importância e utilidade. Essa técnica oferece uma poderosa alternativa para as empresas descobrirem novas oportunidades de negócio e acima de tudo, traçarem novas estratégias para o futuro [CIE 2002].

Segundo Thearling, a MD consiste em um processo de descoberta eficiente de valores padrões não-óbvios a partir de um grande volume de dados. Estes valores muitas vezes se encontram implícitos ou desconhecidos, deixando, assim, de serem úteis na descoberta de conhecimento. O objetivo da MD é descobrir com antecedência características dos dados, tendências ou dependências desconhecidas, para, a partir destas informações descobertas, prever o comportamento futuro de acordo com experiências passadas [THE 2002].

O uso estratégico dos dados pode resultar em oportunidades apresentadas a partir da descoberta de dados escondidos, previamente não-detectáveis, e fatos extremamente valiosos sobre consumidores, varejistas e fornecedores. Tendo essas informações, as organizações podem formular seus negócios, marketing, e estratégias eficazes de vendas, para atingir precisamente a atividade promocional, para descobrir e penetrar mercados novos e para competir com sucesso no mercado [BER 97].

A MD é composta de tarefas e técnicas, utilizadas de acordo com o objetivo de cada aplicação. De acordo com Petrovic as tarefas da MD são [PET 2001]:

- **Classificação**: processo constantemente usado por todos os seres humanos para melhor entender o mundo em torno deles. Durante este processo, classificam-se objetos semelhantes em classes de objetos ou são examinadas

as características de um objeto específico para atribuí-lo a um dos conjuntos de classes conhecidos. Na MD, o processo de classificação é caracterizado por uma definição de todas as classes existentes e de um conjunto predefinido que abrange todos os dados previamente classificados. (Existe sempre um número limitado de classes.) O objetivo da classificação é construir um modelo que possa ser usado para classificar dados novos, ainda não classificados.

- **Previsão**: semelhante à classificação, exceto que ela agrupa os dados de acordo com algum comportamento previsto. Usando dados existentes (históricos), é construído um modelo que explica o comportamento atual. Esse comportamento é usado, então, para prever o comportamento futuro. Por exemplo, em uma mercearia, essa tarefa é usada para descobrir quais itens provavelmente serão comprados juntos.
- **Agrupamento**: processo que particiona dados heterogêneos em vários subgrupos mais homogêneos (tais subgrupos são chamados clusters). Embora à primeira vista o processo de agrupamento pareça idêntico à classificação, existe uma diferença significativa: o agrupamento não usa nenhuma das classes predefinidas. Isso significa que os dados existentes são agrupados de acordo com a auto-semelhança (é tarefa do desenvolvedor determinar o significado dos subgrupos resultantes).

Foram acima explicadas as tarefas da MD, dentre as técnicas de MD são destacadas:

- **Detecção Automática de *Cluster***: essa técnica procura grupos de itens que sejam semelhantes entre si, esperando que os itens semelhantes se comportem de maneiras semelhantes. É uma técnica que tem com vantagens: a facilidade de aplicação, pois como não exige um modelo previamente classificado, é mais fácil de aplicar do que outras técnicas de MD; a possibilidade de ser usada com dados numéricos e textuais [PET 2001].
- **Árvores de decisão**: são uma técnica poderosa que consiste da divisão de itens de um conjunto previamente classificado em grupos desunidos, sendo que cada grupo é descrito a partir do uso de uma regra. Assim, uma árvore de decisão representa uma série de perguntas onde a resposta a cada uma

delas determina qual caminho será seguido (ou seja, qual será a próxima pergunta). Uma das vantagens desta técnica é de ser a de melhor uso para classificação, pois dividem os itens em subconjuntos e, como não exigem muitos cálculos, facilitam a realização de classificações. Outra vantagem é que as regras geradas pelas árvores de decisão são fáceis de entender e traduzir, sendo facilmente traduzidas para o SQL [PET 2001].

- **Regras de associação**: também chamadas por Petrovic de "Análise tipo Cesta de Compras" [PET 2001]. São regras na forma "SE x ENTÃO y" que associam eventos em uma base de dados [THE 2002]. Os algoritmos aplicados para descoberta de regras de associação identificam afinidades entre itens de um subconjunto de dados, estas afinidades são representadas na forma de regras. No exemplo apresentado e explicado por Gimenes, a regra "72% de todos os registros que contém os itens A, B, e C também contém D e E" apresenta porcentagem de ocorrência 72%, o que representa o fator de confiança da regra, e costuma ser usado para eliminar tendências fracas, mantendo apenas as regras mais fortes [GIM 2000]. Ainda segundo o autor, as regras de associação tratam-se de algoritmos tipicamente endereçados à análise de mercado, onde o objetivo é encontrar tendências dentro de um grande número de registros de compras, por exemplo, expressas como transações. Essas tendências podem ajudar a entender e explorar padrões de compra naturais, e podem ser usadas para ajustar mostruários, modificar prateleiras ou propagandas, e introduzir atividades promocionais específicas.
- **Redes neurais**: são modelos computacionais baseados na arquitetura do cérebro humano, que consistem de múltiplas unidades de processos simples conectadas por pesos adaptativos [THE 2002]. Consistem em uma boa maneira de se classificar e predizer regras, quando os resultados do modelo são mais importantes do que entender seu funcionamento. Redes neurais não possuem bom funcionamento nos casos em que existem muitos dados e muitas entradas. Muitos pesos na aplicação a qual é submetida uma rede neural resultam em muito tempo de treinamento da aplicação, que nunca converge para uma solução [PEL 2000].

Na fig. 2.3 encontra-se uma representação da totalidade dos elementos presentes na arquitetura de um DW, os quais foram separadamente apresentados ao longo deste capítulo.

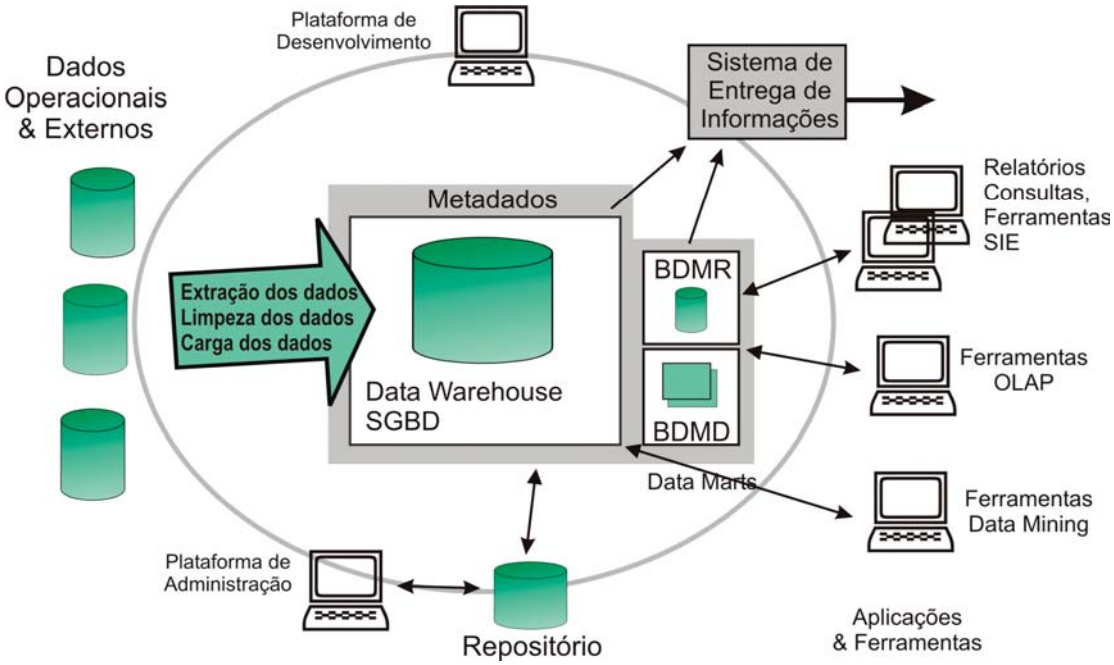

FIGURA 2.3 - Arquitetura do Data Warehouse.
Fonte: BERSON. Building Data Mining Applications for CRM. p.29, 57.

2.3 *Data Webhouse*

O principal enfoque deste trabalho está no *Webhouse*, pois foi proposto que o DW modelado seja publicado na *Internet*, se tornando não somente a modelagem de um DW, mas de um *Data Webhouse*.

Ele consiste de uma instanciação de *Web* do DW [KIR 2000]. Ou seja, um ambiente o qual disponibiliza o DW na *Web*.

Existem duas propostas, apresentadas por Kimbal, para a construção do *Data Webhouse*. Na primeira delas, a qual será posteriormente apresentada na seção 2.3.1, objetiva-se trazer a *Web* para dentro do DW, através do estudo do comportamento do usuário na *Web*. Na segunda proposta, a qual será posteriormente explicada na seção 2.3.2, objetiva-se trazer o DW para a *Web*, através da disponibilização dos dados do DW na *Web* [KIR 2000].

O *Webhouse* possui um papel central e crucial nas operações de um negócio capacitado para a *Web*, e para cumprir este potencial, o *Webhouse* [KIR 2000]:

- Armazena e publica dados de seqüência de cliques e outros dados comportamentais da *Web*, que guiam uma compreensão do comportamento do cliente;
- É adaptado a outros *data marts* distribuídos no DW da empresa e a *data marts* acima e abaixo na cadeia de fornecedores, de modo que todos esses *data marts* podem ser utilizados em conjunto;
- É uma fonte adaptável e flexível de informações. À medida que novas perguntas de negócio surgem e que novas origens de dados tornam-se disponíveis, o *Webhouse* responde elegantemente à novidade. Uma resposta elegante é aquela que permite que aplicativos antigos continuem sendo executados sem interrupção e sem reprogramação, mas também permite que as novas perguntas e os novos dados coexistam;
- É extensível aos novos meios da *Web*, incluindo imagens paradas (*still images*), imagens gráficas, áudio e vídeo;
- É um meio seguro que publica dados para clientes, parceiros de negócio e funcionários de forma adequada, mas que, ao mesmo tempo, protege os ativos de dados da empresa contra utilização não intencionada;
- É a base para as decisões de conversões para a *Web*. O *Webhouse* deve permitir que os usuários tomem decisões sobre a *Web* enquanto estiverem utilizando a *Web*.

2.3.1 A *Web* no *Data Warehouse*

Foram citadas as duas propostas para o uso do DW, na primeira, onde a *Web* é trazida para o DW, é usada uma técnica chamada *Clickstream* (seqüência de cliques) para a exploração de informações de acesso a *Web*. Para o funcionamento desta técnica os dados são obtidos através de *logs*, como é mostrado na fig. 2.4, mantidos pelo servidor *Web*, de todos os acessos efetuados. Obtém-se também através de seqüências de cliques de parceiros direcionadores (*referring*) ou de ISPs (*Information Service*

Providers), ou através de serviços de estatísticas da *Web*, os quais são empregados para colocar um controle sobre páginas *Web* que alertam quando um usuário acessa a página.

Esses dados coletados são importantes para fornecer informações ricas sobre a dinâmica de acesso, ajudando a melhorar a qualidade das interações com os usuários, podendo levar a maior lealdade e, conseqüentemente, aumento de receita [CAM 2002]. Porém, esses dados não podem ser usados de forma indiscriminada, primeiro por não possuírem informações suficientes, e segundo porque uma seqüência de cliques bruta não é uma descrição útil de comportamento, podendo levar a conclusões precipitadas.

A limpeza e a transformação destes dados são ações fundamentais e requerem conhecimento da estrutura do *site* e da aplicação. Aliás, neste ponto, a construção do *Webhouse* é muito parecida com a construção do DW, pois a partir do momento em que os dados foram obtidos, serão carregados no *Webhouse* somente após sua transformação, passando, assim, pelas etapas de ETC demonstradas no capítulo anterior. O autor Ralph Kimbal em [KIR 2000] chama as etapas de ETC, no *Webhouse*, de "pós-processador de seqüência de cliques", o qual é apresentado na fig. 2.4, e diz que neste aplicativo de pós-processamento devem ser encontradas, além da extração de chaves de dimensão para sessões, usuários e *hosts*, as seguintes tarefas:

- **Filtragem dos registros não necessários**, onde são mesclados dados associados e excluídos os registros que não serão passados para o *Webhouse*, reduzindo o máximo possível o volume de transações presentes, sem comprometer a integridade e a completitude dos dados necessários ao suporte da granularidade do projeto do DW;
- **Identificação de sessões**, onde são marcados os registros associados de seqüência de cliques com um identificador único de sessão. Também é verificado se os tempos de eventos são logicamente consistentes entre si e entre os registros que descrevem a sessão;
- **Identificação de usuários**, onde é feita a correspondência entre o usuário e um identificador existente de usuário, se possível. Caso contrário, atribui-se um identificador anônimo único de usuário se a identidade for desconhecida;
- **Identificação de *hosts***, onde são convertidos (para a granularidade desejada) os endereços de IP (*Internet Protocol*) de clientes e de origens de conexões. É retido o país de origem de dados canônicos de domínio;

- **Consolidação dos dados em um formato uniforme**, onde os dados de seqüência de cliques são colocados em um formato definido, aceitável para o software de carregamento de DW.

Assim que os dados são carregados no *Webhouse*, é preciso efetuar sua análise. Ferramentas OLAP, já mencionadas na seção 2.2.6.1, são muito utilizadas na análise dos dados, elas fornecem visões dos dados segundo diferentes perspectivas e diferentes níveis conceituais, respondendo perguntas do tipo: Quais componentes ou serviços são os mais e menos utilizados? Qual a distribuição do tráfego na rede ao longo do tempo? Quais as diferenças de acesso entre os usuários de diferentes regiões geográficas? [CAM 2002].

Quando o desejado é uma exploração mais detalhada dos dados, é muito usada a ferramenta de Mineração de Dados (MD), mencionada na seção 2.2.6.2, por consistir de um conjunto de técnicas utilizadas com o objetivo de descobrir informações implícitas e potencialmente úteis nos dados armazenados. A MD provê análises de séries de tempo, associações, classificações e outros, ela é usada para responder perguntas do tipo: Em que circunstâncias são os componentes ou serviços usados? Quais as seqüências típicas de eventos? Existem padrões de comportamento entre todos os usuários? O comportamento de usuários muda ao longo do tempo e como? [CAM 2002].

Na figura 2.4, encontra-se uma ilustração geral da tarefa de trazer a *Web* para o DW.

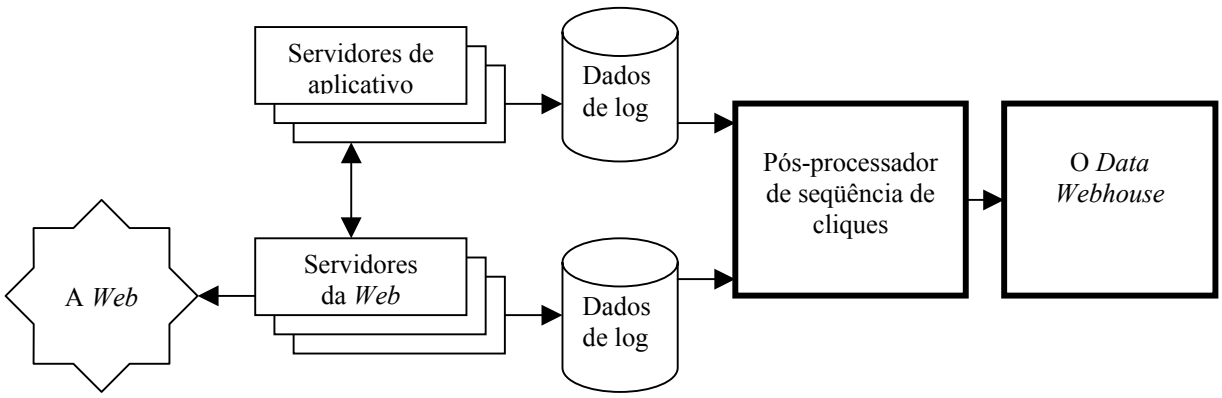

FIGURA 2.4 - Mecanismo de criação do Webhouse através da *Web*.
Fonte: KIMBALL. Data Webhouse: construindo o Data Warehouse para a Web. p.176.

2.3.2 O *Data Warehouse* na *Web*

Até o momento foram explicados todos os passos que devem ser seguidos para a construção do DW, todos os seus componentes e sua arquitetura, até mesmo como construir o DW a partir de dados obtidos pela *Web*. Mas dois pontos importantes que ainda não foram focados tratam da necessidade de divulgação e do movimento do DW. Tendo em vista esta necessidade, nesta seção será apresentada a segunda proposta de uso do DW, onde a tarefa é de trazer o DW para a *Web*.

No aspecto de divulgação e movimentação do DW, é impossível não se pensar no uso da *Internet* e em todos os benefícios que ela oferece. Ela permite que as pessoas tenham acesso rápido, a princípio, a dados dos mais variados assuntos, 24 horas por dia, sete dias por semana, e constantemente atualizados. Sem contar com a facilidade do acesso aos dados da *Web*, pois qualquer tipo de usuário pode preencher formulários simples e utilizar os navegadores *Web*.

Para o DW, a *Web* vem a facilitar grandemente a divulgação e a movimentação dos dados, pois, por exemplo, no caso de uma empresa que tem seus dados organizados em um DW, a divulgação pela *Web* é ótima pelo fato de seus clientes, em grande maioria, estarem conectados à *Internet*, podendo, independentemente de sua localidade, interagirem com o DW, neste caso, *Data Webhouse*.

Mas, para que o DW seja disponibilizado na *Web*, deve seguir várias novas regras [KIR 2000]. Pois a *Web* impõe algumas regras de usabilidade que devem ser seguidas para se obter sucesso no projeto do *Data Webhouse*. Abaixo estão descritas algumas delas:

- **Facilidade de utilização pelos usuários**. Até um certo tempo, os projetistas de interface com o usuário não tinham valores diretos das atividades dos usuários, as novas idéias com relação à usabilidade eram pouco consideradas e as sugestões individuais de usuários eram tidas como insignificantes. Com a *Web*, uma pressão bem maior é imposta sobre as ferramentas, no sentido de serem de fácil utilização [KIR 2000];
- **Vocabulário fácil**. A linguagem utilizada não pode ser estritamente técnica, pois, mesmo que os acessos sejam em grande parte feitos por conhecedores da área, o *Webhouse* está localizado em um ambiente de amplo acesso por diversos tipos de usuários, não podendo ser vinculado a um único grupo;

- **Velocidade no acesso aos dados**. Um tempo de resposta eficaz é indispensável para o sucesso do projeto de *Webhouse*, apesar de os dados no *Webhouse* serem históricos e de acesso mais demorado, a modelagem deve ser efetuada de forma que o usuário não fique indefinidamente aguardando a resposta de uma consulta. Quando se está trabalhando com um DW local, até pode-se permitir um maior tempo de resposta a consultas, mas quando se está acessando um *Webhouse*, onde o usuário depende, em média, de uma conexão de 56 Kilobytes por segundo, um tempo de resposta lento não é aceitável;
- **Natureza Multicultural da *Web***. As informações dispostas no *Webhouse*, tanto às contidas na interface com o usuário quanto às contidas no BD, devem estar em padronização internacional, tendo em vista o uso global que terá o *Webhouse*. Devem ter padronização internacional, por exemplo, os nomes, endereços, telefones, datas, horários e valores monetários;
- **Formato multimídia**. O *Data Webhouse* deve entregar seus resultados em um ambiente multimídia. É importante fornecer imagens gráficas, animação, imagens ampliáveis, mapas, videoclipes e sons [KIR 2000]. Além disso, os usuários querem a possibilidade de armazenar os dados pesquisados em diversos formatos, tanto em arquivos extensão PDF (*Portable Document Format*), quanto em planilhas ou arquivos texto, o *Webhouse* deve fornecer esta possibilidade;
- **Segurança e privacidade dos dados**. No ambiente *Web* é muito difícil certificar-se se os dados serão acessados somente por aqueles que tenham a devida permissão para o acesso. A única maneira de certificar-se da total segurança no acesso aos dados do *Webhouse* é tendo na equipe de projeto, desde o início, um especialista em segurança de rede dedicado em tempo integral, pois este especialista entenderá e especificará um sistema de segurança de um início de sessão e de um único console que será construído desde a base como um sistema instalado/distribuído para a *Web* [KIR 2000].

Após esta apresentação dos tipos de uso do *Data Webhouse*, neste trabalho pretende-se executar a criação do seu modelo a partir da segunda proposta definida por Kimball em [KIR 2000], que é trazer o DW para a *Web*.

3. Definição do Problema

Os DWs podem ser utilizados para os mais diversos assuntos, por se tratarem de uma nova metodologia para o armazenamento e acesso a informação. Esta nova metodologia, adotada na forma com que os dados são dimensionados e armazenados no DW, é que o torna mais complexo e, ao mesmo tempo, mais atraente do que os BDs tradicionais. Devido a sua estrutura planejada, a qual já foi devidamente detalhada neste trabalho, o tratamento de grandes volumes de dados históricos não se torna uma tarefa tão desgastante e, muitas vezes, quase impossível de ser feita. Fato este que deve ser mostrado de forma a encorajar empresários de várias áreas a adotarem o uso do DW para tornar eficiente o uso dos dados que possuem, pois, devido a grande quantidade de dados que estes usuários precisam dominar, sentem-se incapazes de alcançar diferencial competitivo no mercado.

Neste trabalho, busca-se focalizar o problema do armazenamento de dados históricos no caso específico do mercado de sementes geneticamente modificadas do MERCOSUL. Pois, nos últimos anos, o número de descobertas científicas e inovações tecnológicas na área de sementes geneticamente modificadas tem crescido muito, em decorrência deste e de outros fatores, tem crescido também a competição para a criação de novos cultivares, pois esta é incentivada pelo fato de seus inventores possuírem os direitos de propriedade intelectual sobre suas invenções, encorajando o desenvolvimento industrial e o crescimento econômico [ALC 2000].

Nos países do Mercosul, esse número elevado de descobertas científicas e criação de novos cultivares tem aberto caminho para uma nova mentalidade a respeito do desenvolvimento, comercialização e consumo de plantas produzidas usando tecnologias de DNA recombinante (modificadas geneticamente ou plantas transgênicas), e também plantas produzidas através de métodos tradicionais com novas tecnologias associadas. Os resultados, obtidos através de pesquisa e desenvolvimento na área de biotecnologia, começam a chegar ao mercado, por meio de produtos de primeira geração, que apresentam tolerância a herbicidas e resistência a insetos. As conseqüências possíveis a serem geradas nos próximos anos estão diretamente relacionadas aos aspectos ligados especificamente ao mercado agrícola, no que se refere

ao fluxo de oferta de insumos, basicamente e principalmente às sementes, e ainda ás questões ligadas à segurança alimentar [ALC 2000].

Com o rápido crescimento do número de empresas vinculadas à pesquisa, desenvolvimento e comercialização de produtos geneticamente alterados e a constante busca da descoberta de novos cultivares, aumentam, em conseqüência, as informações vinculadas a esses produtos, as quais devem ser plenamente conhecidas e dominadas pelos empresário e agricultores do ramo, para permanecerem à frente do mercado. Porém, atualmente, não existe ambiente que reúna todas estas informações de uma forma otimizada, para que estes usuários efetuem suas consultas na área de sementes. O que torna a tomada de decisão, por parte desses usuários, uma tarefa difícil e que exige pesquisa em diversas fontes diferentes, a qual não oferece uma visão global dos dados necessários.

Tendo em vista a otimização na análise de grandes volumes de dados históricos que os DWs oferecem, neste trabalho procurou-se estudar qual a melhor forma de disponibilizar os dados, relativos a sementes modificadas, em um DW, para que venha a solucionar o problema de análise dos dados, apresentado acima. E, como os usuários que necessitam destas informações se localizam em vários países e pertencem a diversos ramos, procurou-se também estudar a forma de modelar estes dados buscando rapidez no acesso, de forma a viabilizá-los na *Internet*, na forma de um *Data Webhouse*.

4. Modelagem do *Data Webhouse*

4.1 Introdução

Neste capítulo estão descritos todos os passos que foram seguidos para a criação de um modelo de DW de dados de sementes geneticamente modificadas do MERCOSUL, juntamente com a descrição de um protótipo deste DW, desenvolvido para comprovar a viabilidade da aplicação da modelagem.

Para a criação da modelagem buscou-se, primeiramente, ter uma visão geral de todas as tarefas necessárias para a criação. Em seguida, estas tarefas foram divididas em etapas a serem executadas, com o objetivo de criação de um protótipo da modelagem proposta, as etapas foram as seguintes:

- Definição dos dados relevantes para a composição do DW;
- Localização e extração destes dados necessários das bases de dados onde estavam armazenados;
- Definição geral da padronização a ser aplicada nos dados, esta realizada a partir de estudo de como os usuários que lidam com estas informações costumam tratá-las;
- Tratamento dos dados, etapa na qual foi realizada: a limpeza dos dados, o tratamento de dados nulos (dados que não foram encontrados nas fontes utilizadas para extração), a padronização dos dados e a especificação dos metadados;
- Definição final de como os dados deveriam ser dispostos na modelagem, etapa onde foram especificadas as tabelas de dimensão e de fatos e quais seriam os registros de cada tabela, especificados através de sua relevância e formato (nas tabelas de dimensões encontram-se dados textuais, enquanto que nas tabelas de fatos encontram-se dados numéricos);
- A partir dos dados já extraídos, organizados e transformados, definição da viabilidade da construção de um protótipo da modelagem criada, juntamente com a definição do ambiente em que seria construído o protótipo.

4.2 Especificação e busca dos dados

Para a descoberta dos dados relevantes que iriam constituir o DW foi feito um estudo nas principais fontes da área, como Ministério da Agricultura do Brasil, Secretaria da Agricultura do Estado do RS, Instituto Nacional de Semillas /INASE – Uruguai, Instituto Nacional de Semillas/INASE – Argentina, Departamento de Semillas/DISE/MA – Paraguai, dentre outros. Foram pesquisadas quais as informações que são necessárias pelos usuários para a tomada de decisão e, segundo estudo desenvolvido por Alcântara em [ALC 2000] e os resultados da pesquisa desenvolvida, chegou-se à conclusão que os usuários necessitam de informações relacionadas a:

- **Produção de cada espécie de semente, por ano, em cada país**. Nestas informações, a respeito da produção, destacam-se os dados relativos a: quantidade de sementes produzidas no ano, às próprias sementes, produtividade e área colhida.
- **Comercialização de cada espécie de semente, por ano, em cada país**. Onde se destacam dados relativos a: valor total de importações obtidas, valor total de exportações obtidas, quantidade importada e quantidade exportada.
- **Cultivares derivados de cada espécie**. Constam de listagens a respeito da criação do cultivar (gen inserido), se é um cultivar que possui o cultivo somente registrado ou se é protegido, se possui certificação, quem possui a posse do cultivar, quando obteve concessão para o cultivo e comércio e até quando possui esta concessão.

Com os dados já especificados, foi iniciada a tarefa de busca e extração dos dados. A busca ocorreu nos *sites* das organizações, já mencionadas, e através de pedidos, via *e-mail*, para que as organizações, que não possuíam a totalidade de seus dados *on line*, disponibilizassem seus dados.

Tendo em vista o grande volume de dados que teriam que ser extraídos para a criação do protótipo do DW, decidiu-se somente extrair dados a respeito da produção e comercialização de sementes ocorridas no território brasileiro. O motivo se deve, principalmente, ao fato do Brasil ser o maior dos países do MERCOSUL em termos de território, e porque a modelagem do DW com dados brasileiros já abrange toda a

problemática que seria enfrentada em uma modelagem total dos dados dos quatro países. Pois o modelo aplicado é o mesmo, o que difere é o volume de dados a serem analisados e processados pelo DW, o que possivelmente afetaria somente o tempo gasto para o SGBD efetuar as consultas.

Por motivos análogos aos acima apresentados, decidiu-se também efetuar a extração somente de dados a respeito das sementes de Arroz, Milho e Soja, pois esses produtos possuem grande significância, tanto em termos de produção quanto de comercialização, nos quatro países do MERCOSUL [ALC 2000].

A tarefa se extração ocorreu, basicamente, através de busca exaustiva nos *sites* das organizações, devido à dificuldade de se obter estes dados diretamente das bases onde estão alocados. Estes dados foram extraídos e organizados de forma a permitir o seu tratamento e padronização.

4.3 Preparação dos dados

Nesta etapa, os dados foram agrupados de forma a facilitar a compreensão. Para a sua padronização, optou-se pela formatação internacional, com o valor da moeda em dólares americanos, as medidas de área em hectares (ha) e os volumes em toneladas métricas (tm), tendo em vista ser a formatação já familiar aos usuários em questão e ser a mesma utilizada pelas fontes de onde foram extraídos os metadados utilizados na modelagem.

O tratamento de dados nulos é uma tarefa complexa e que exige um amplo conhecimento da área em questão, pois não se pode apenas adotar um valor médio, obtido nos últimos anos, como padrão para valores nulos, pois este poderá diferir seriamente do valor real. Neste caso, a modelagem ficaria comprometida, pois se afastaria do objetivo central da criação do DW, que é automatizar a análise dos dados existentes, facilitando a tomada de decisão. Portanto, optou-se por apenas adotar o valor zero e informar aos usuários, através dos metadados, esta medida padrão adotada.

Na tab. 4.1 estão descritos os dados de produção de arroz, obtidos pelo Brasil nos anos de 1991 a 2001, e na tab. 4.2 estão descritos os dados de comercialização de arroz, obtidos também pelo Brasil nos anos de 1990 a 2000. Nas tabelas estes já estão organizados e padronizados na forma com que foram armazenados no BD criado, o qual será posteriormente apresentado.

Como já foi mencionado, na busca dos dados foram extraídos dados das espécies: arroz, milho e soja, dos anos de 1961 a 2001. Entretanto, devido a grande quantidade dos dados, nas tabs. 4.1 e 4.2 estes estão parcialmente descritos.

TABELA 4.1 - Dados de produção de Arroz no Brasil, de 1991 a 2001.

	Produção (tm)	Sementes (tm)	Produtividade (hg/ha)	Área Colhida (ha)
1991	9488007	370626	23020	4121600
1992	10006292	352956	21349	4687020
1993	10107310	339977	22912	4411320
1994	10540789	335972	23876	4414800
1995	11226064	248494	25668	4373540
1996	8643803	235129	26566	3253770
1997	8351665	239803	27310	3058130
1998	7716090	292690	25198	3062200
1999	11709700	281569	30708	3813270
2000	11089800	238764	30339	3655290
2001	10195400	241288	32453	3141630

TABELA 4.2 - Dados de comercialização de Arroz no Brasil, de 1990 a 2000.

	Valor Importação ($)	Quantidade Importação (tm)	Valor Exportação ($)	Quantidade Exportação (tm)
1990	$144,011	413825	$871	1248
1991	$372,332	960193	$1,087	1606
1992	$154,435	583904	$2,379	3565
1993	$214,127	700724	$5,432	12330
1994	$323,429	987120	$1,435	3307
1995	$293,560	870506	$4,673	18537
1996	$301,401	792463	$4,384	21811
1997	$323,392	816116	$2,397	9158
1998	$545,370	1304958	$3,857	6566
1999	$275,115	984265	$13,735	47639
2000	$140,746	659508	$6,505	26380

4.3.1 Metadados

Os metadados possuem extrema importância no entendimento dos dados e a sua criação deve ser planejada e ter em vista o grupo de usuários a que se destina, para determinar o tipo de linguagem que deve ser usado. Para este trabalho buscou-se adaptar os metadados extraídos de [FAO 2000], para que se tornassem mais claros aos usuários pertencentes aos países do MERCOSUL. Para a criação dos metadados também foram utilizadas como fonte as informações apresentadas em [ALC 2000].

4.3.1.1 Descrição dos elementos de comercialização

- **Quantidade das Importações e Quantidade das Exportações:** Estes dois elementos reportam o comércio estrangeiro (importação e exportação) de forma quantitativa. A unidade de medida é a mesma (toneladas métricas) para todos os produtos, com exceção de animais vivos, que são relatados em unidades (cabeças), exceto as aves domésticas, pombos e coelhos, que são relatadas em '000 unidades. Em regra geral, os dados de comércio se referem à quantidade líquida. Todos os produtos florestais são relatados em volumes contínuos, com exceção da polpa da madeira e da polpa de outra fibra, que são dados em peso, o papel e carvão para lenha são expressos em toneladas métricas (tm). O valor "0" (zero) obtido no resultado da consulta representa a não existência da informação requerida no banco de dados.

- **Valor das Importações e Valor das Exportações**: Ambos estes elementos expressam o comércio estrangeiro em termos do valor das importações e exportações. Os dados são armazenados em mil dólares americanos. As moedas correntes nacionais, usadas pelos países durante as transações legais, são convertidas usando-se a taxa média de troca anual, fornecida pelo Fundo Monetário Internacional (FMI). Somente em alguns casos as taxas de troca foram extraídas de fontes nacionais. Para os países que não seguem esta regra geral, é necessário consulta às notas de comércio. O valor "0" (zero) obtido no resultado da consulta representa a não existência da informação requerida no banco de dados.

4.3.1.2 Descrição dos elementos de produção

- **Área Colhida:** Os dados referem-se à área na qual as culturas foram colhidas. Estão excluídas as áreas nas quais, embora tenham sido semeadas ou plantadas, não tenha havido colheita devido a danos, perdas, etc. Em geral estes valores são líquidos para culturas temporárias e brutos para culturas permanentes. Com relação a culturas mistas ou associadas, os países foram requisitados a informar a área semeada relativa a cada cultura separadamente. Quando a mistura refere-se a uma cultura em particular, geralmente grãos, foi considerada uma única cultura; área semeada é informada somente para a cultura reportada. Portanto, no caso de culturas com diferentes formas, a área semeada refere-se ao total.

  Se a cultura em consideração é colhida mais de uma vez durante o ano, como conseqüência de sucessivas semeaduras, a área é contada tantas vezes quanto foi colhida. A área colhida é expressa em hectares (ha).

- **Produtividade:** Os dados representam a produção colhida por unidade de área plantada das culturas. Em muitos casos os dados de rendimento não foram registrados, mas obtidos dividindo os dados de produção pela área plantada. Os dados estão registrados em hectograma (hg - 100 gramas) por hectare (hg/ha).

- **Produção:** Em princípio, os dados de produção relacionam-se a produção interna total não importando se dentro ou fora do setor agrícola, ou seja, incluem produção comercial e não comercial. Em geral, os dados de produção são registrados em nível de propriedade agrícola para produção vegetal e produção animal, incluindo no caso de culturas vegetais as perdas de colheita. A produção, portanto, inclui as quantidades da cultura vendida no mercado (produção comercializada) e as quantidades consumidas ou mesmo usadas pelos produtores (autoconsumo).

  Quando os dados de produção disponíveis referem-se a um período de produção correspondente a dois anos sucessivos e não é possível alocar a produção a cada um deles, são tomados como referência os dados de produção no período o qual recai a maior parte da produção. Os dados de produção agrícola são apresentados em toneladas métricas (tm).

- **Sementes:** Os dados incluem a quantidade da cultura deixada para semear durante o ano, tanto a produzida internamente como a importada. Foi considerado o número de semeadura no caso de culturas que são realizados mais de um plantio por ano.

  Os dados sobre sementes incluem também, dependendo do caso, as quantidades necessárias para semeadura ou plantio de áreas relacionadas à parte dos cultivos destinados para ração animal. Quando dados oficiais não estão disponíveis, o volume de sementes foi estimado em relação à área cultivada no ano subseqüente. Os dados de sementes são apresentados em toneladas métricas (tm).

4.4 Modelagem e dimensionalidade dos dados

A modelagem do DW proposto neste trabalho foi feita de acordo com as regras apresentadas na fundamentação teórica deste trabalho. No modelo existem, primeiramente, duas Tabelas de Fatos, as quais são dominantes e possuem a maior parte dos dados a serem consultados, estas tabelas são: **Fato Produção** e **Fato Comércio**. Existem também as Tabelas de Dimensão, as quais qualificam as Tabelas de Fatos, que são: **Dimensão Espécie**, **Dimensão País** e **Dimensão Tempo**.

Existem também as listagens de **Cultivares Protegidos** e **Cultivares Registrados**, as quais não fazem parte do DW por consistirem apenas de listagens informativas. Por serem consultas importantes aos usuários que irão consultar o DW, resolveu-se propor a sua existência paralela ao DW, no mesmo BD, mas com a estrutura de armazenamento mais simples e linear, não tendo seus valores cruzados para as consultas, como ocorre com as outras tabelas.

Na fig. 4.1, a modelagem do DW é demonstrada na forma de um cubo, o qual consiste apenas de uma representação visual do modelo, para facilitar o entendimento. As dimensões do cubo são independentes umas das outras, e são as mesmas para cada ponto no cubo, portanto, o dado existente em cada ponto do cubo é uma combinação dos dados de cada uma das suas dimensões.

No cubo, as tabelas de dimensão País, Espécie e Tempo se cruzam e fornecem dados às tabelas de fatos Produção e Comércio.

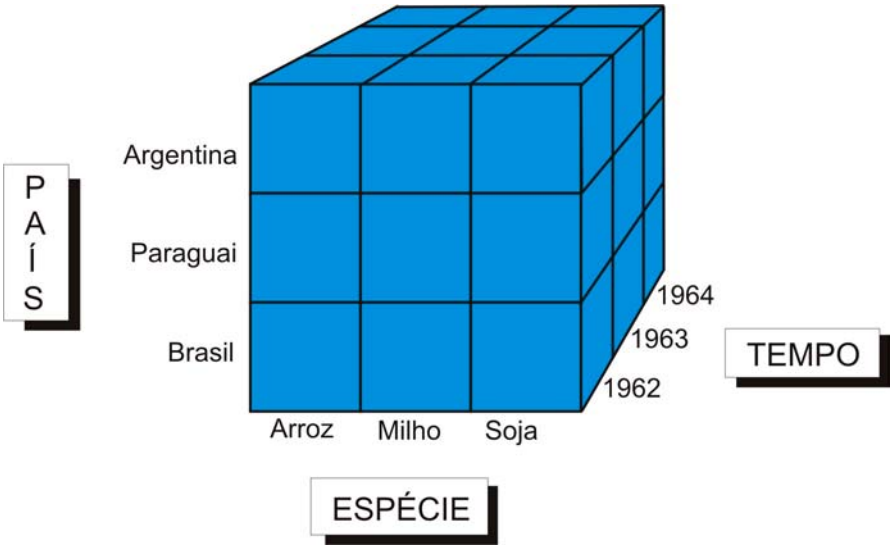

FIGURA 4.1 - Representação da disposição dos dados.

A modelagem foi feita de acordo com as regras ditadas pelo Esquema Estrela, explicado no capítulo 2 deste trabalho, o qual, na verdade, é uma mistura de modelagem conceitual com modelagem lógica, pois já é bastante voltada para a abordagem relacional, tratando o armazenamento como em tabelas [CAM 2002]. Na fig. 4.2 encontra-se a modelagem do DW no Esquema Estrela, somente com a tabela Fato Comércio, pois a modelagem da tabela Fato Produção já foi mostrada no capítulo 2, fig. 2.1.

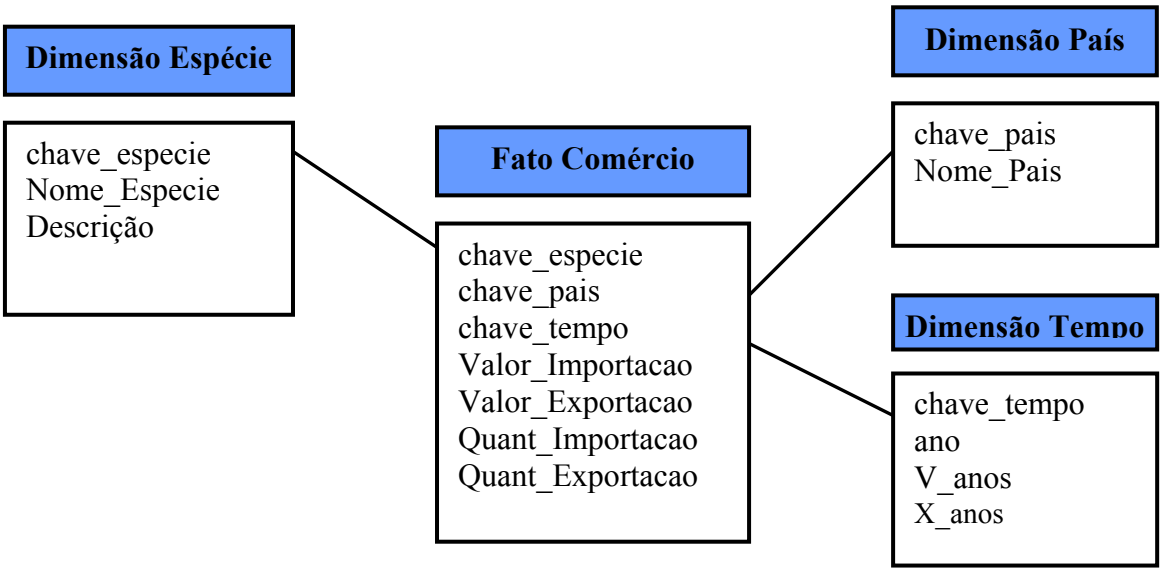

FIGURA 4.2 - Esquema Estrela da tabela de Fatos Comércio.

4.4.1 Protótipo da modelagem criada

Neste trabalho será demonstrado, paralelamente ao detalhamento da modelagem, um protótipo do DW proposto, o qual foi desenvolvido para demonstrar a viabilidade da implementação da modelagem apresentada. Este protótipo foi desenvolvido no ambiente descrito abaixo:

- Sistema Operacional Windows 2000 Server;
- SGBD Microsoft SQL Server 2000 - versão para experimentação, 120 dias;
- Microsoft SQL Server Analysis Services - versão para experimentação, 120 dias;
- Microsoft Commerce Server 2002 - versão para experimentação, 120 dias.

Optou-se por usar o ambiente Windows porque o Microsoft SQL Server foi o SGBD, com recursos para a criação de DWs, encontrado em versão para experimentação, e também por ter-se a disposição uma versão registrada do Sistema Operacional Windows 2000 Server, exigido pelo SGBD para a criação de DWs. Utilizou-se uma versão para experimentação de um software pago por não ter sido encontrado um software livre com suporte a criação de DWs.

O SGBD Microsoft SQL Server 2000 possui sua plataforma para o desenvolvimento de DWs baseada na mesma metodologia proposta pelo autor Ralph Kimball em [KIR 96], a qual foi seguida neste trabalho. Esta metodologia apresenta o desenvolvimento de DWs a partir da criação de cubos de dados, que são originados pelas tabelas de dimensão e de fatos, para a análise dos dados.

O SGBD Microsoft SQL Server 2000 foi utilizado para a criação de um BD, o qual foi necessário como fonte de alimentação de dados para o DW, as tabelas deste BD serão descritas mais tarde.

Com a ferramenta Microsoft SQL Server Analysis Services, a qual é usada para o processamento dos cubos multidimensionais e para o envio de resultados de consultas feitas pelos clientes [PET 2001], foram criados dois cubos de dados, um para a tabela de fatos Produção, apresentado na fig. 4.3, e outro para a tabela de fatos Comércio, apresentado na fig. 4.4. As dimensões também foram criadas através do uso da ferramenta Microsoft SQL Server Analysis Services, e estão descritas na fig. 4.5, onde é mostrada a disposição dos seus dados no cubo criado.

Para o desenvolvimento foi utilizada a modelagem multidimensional, por ser a que melhor se adapta a análise de dados no cubo. No armazenamento foi utilizado o tipo ROLAP, por oferecer melhor desempenho de consultas simples em DWs multidimensionais.

A ferramenta Microsoft Commerce Server 2002 permite, através do servidor, a publicação na *Internet* do DW criado, sem a necessidade do uso de linguagens de programação no intermédio do *site Web* com o BD. A ferramenta também oferece meios para que o desenvolvedor utilize o DW criado tanto para publicar na *Internet* dados vindos do DW quanto para importar dados de "seqüências de cliques" de usuários da *Internet* para o DW, como foi apresentado no capítulo 2, onde foi explicado o porquê de trazer a *Web* para o DW. A maior parte do protótipo foi realizada por meio desta ferramenta, por ela possuir o componente Commerce Server Manager, o qual permite o uso do Microsoft SQL Server, do Microsoft SQL Server Analysis Services e do próprio Commerce Server em um único local.

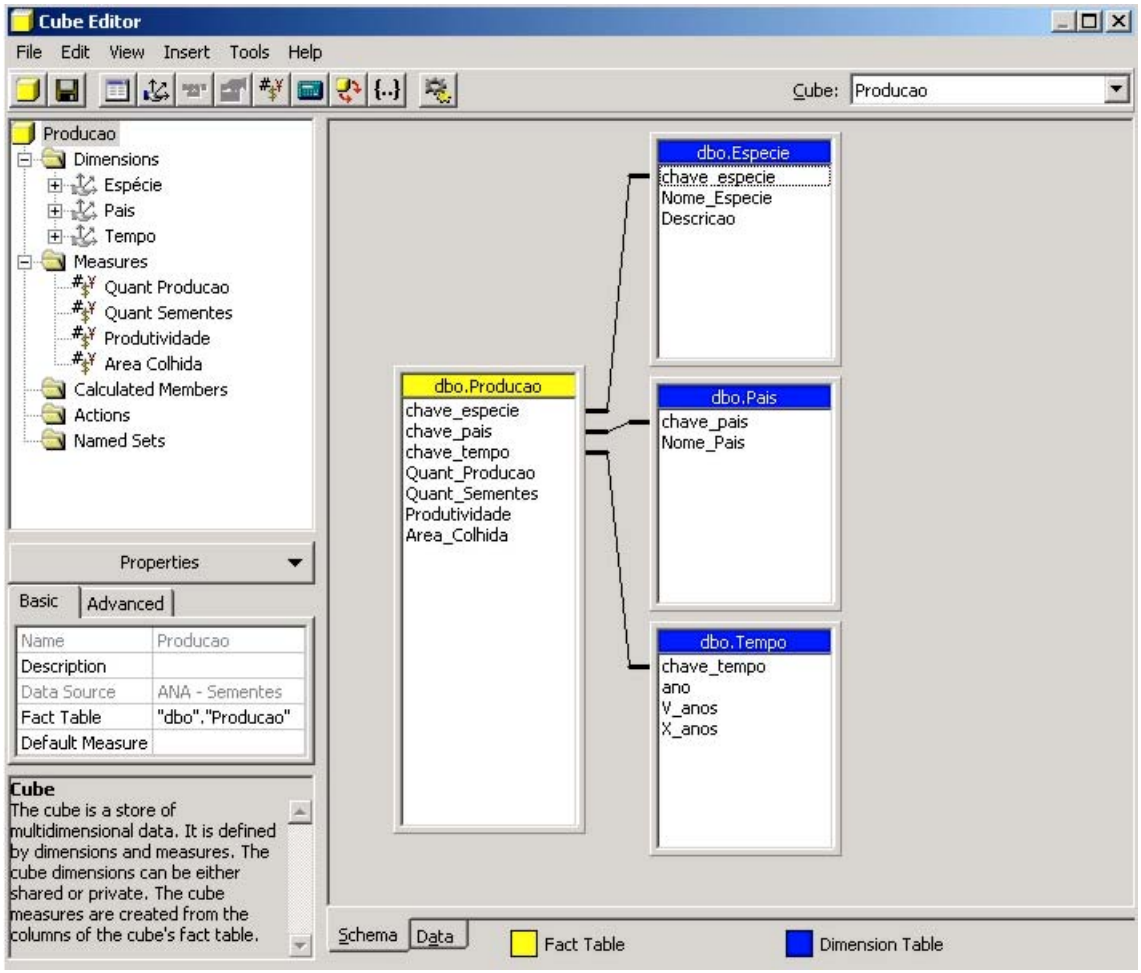

FIGURA 4.3 - Descrição da estrutura do cubo Produção.

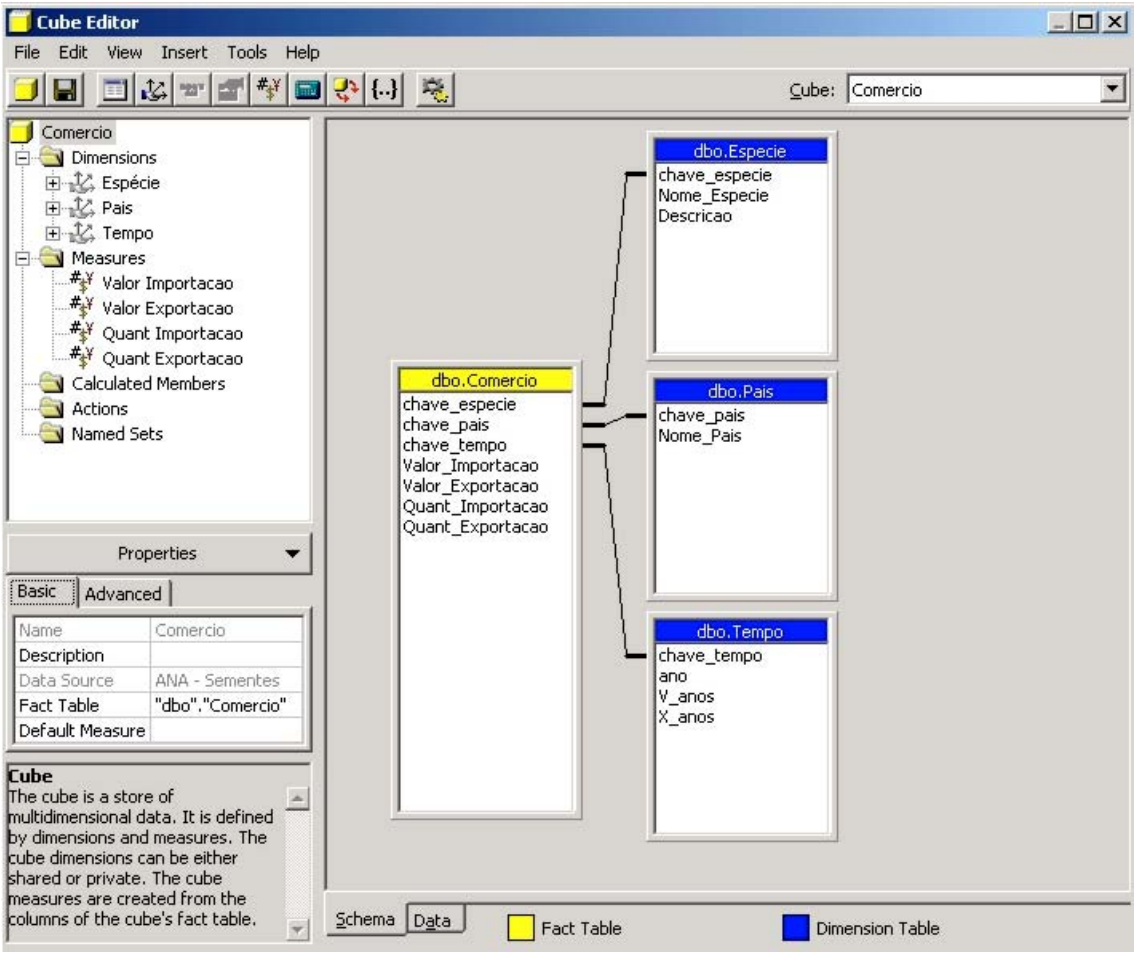

FIGURA 4.4 - Descrição da estrutura do cubo Comércio.

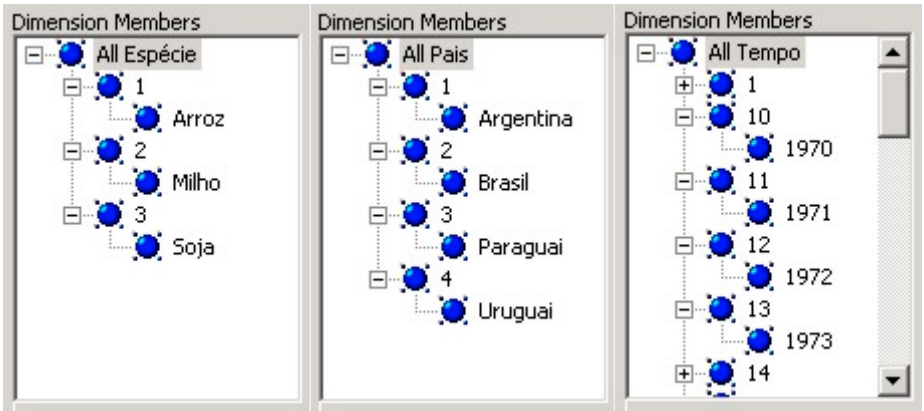

FIGURA 4.5 - Estrutura dos dados das dimensões no cubo.

Para o protótipo, também foi criado o *layout* da pagina principal e das demais páginas do *site* que irá publicar os dados do DW na *Web*. Com o objetivo de demonstrar como o usuário optará pelos dados a serem pesquisados nas consultas e como terá acesso aos metadados. O *layout* do *site* foi desenvolvido de forma a se adaptar aos conceitos ditados por Kimbal em [KIR 2000] e apresentados no capítulo 2, onde é apresentada a tarefa de trazer o DW para a *Web*.

As figuras que apresentam as páginas *Web* criadas encontram-se no Anexo I.

4.4.2 Descrição das tabelas

- **Cultivar_Protegido**(cod_cultp, especie, cultivar, num_certificado, titular, nome_titular, data_concessão, validade);
  cod_cultp: chave-primária da entidade Cultivar_Protegido.
- **Cultivar_Registrado** (cod_cultr, especie, num_certificado);
  cod_cultr: chave-primária da entidade Cultivar_Registrado.
- **Dimensão Espécie** (chave_especie, Nome_Especie, Descrição);
  chave_especie: chave-primária da Dimensão Espécie.
- **Dimensão País** (chave_pais, Nome_Pais);
  chave_país: chave-primária da Dimensão País.
- **Dimensão Tempo** (chave_tempo, ano, V_anos, X_anos);
  chave_tempo: chave-primária da Dimensão Tempo.
- **Fato Comercio** (chave_especie, chave_pais, chave_tempo, Valor_Importacao,Valor_Exportacao,Quant_Importacao, Quant_Exportacao);
  chave_ especie, chave_ pais, chave_tempo: chave-primária composta de Fato Comercio.
- **Fato Producao** (chave_especie, chave_pais, chave_tempo, Quant_Producao, Quant_Sementes, Produtividade, Área_Plantada);
  chave_ especie, chave_ pais, chave_tempo: chave-primária composta de Fato Producao.

Na fig. 4.6 encontra-se o diagrama de relacionamentos gerado entre as tabelas, acima descritas, criadas no Microsoft SQL Server 2000. Já neste pode-se visualizar a forma em estrela em que as tabelas estão dispostas. As tabelas Cultivar_Protegido e Cultivar_Registrado se encontram a parte, sem relacionamentos, por serem independentes do DW, pois são usadas apenas para listagens simples.

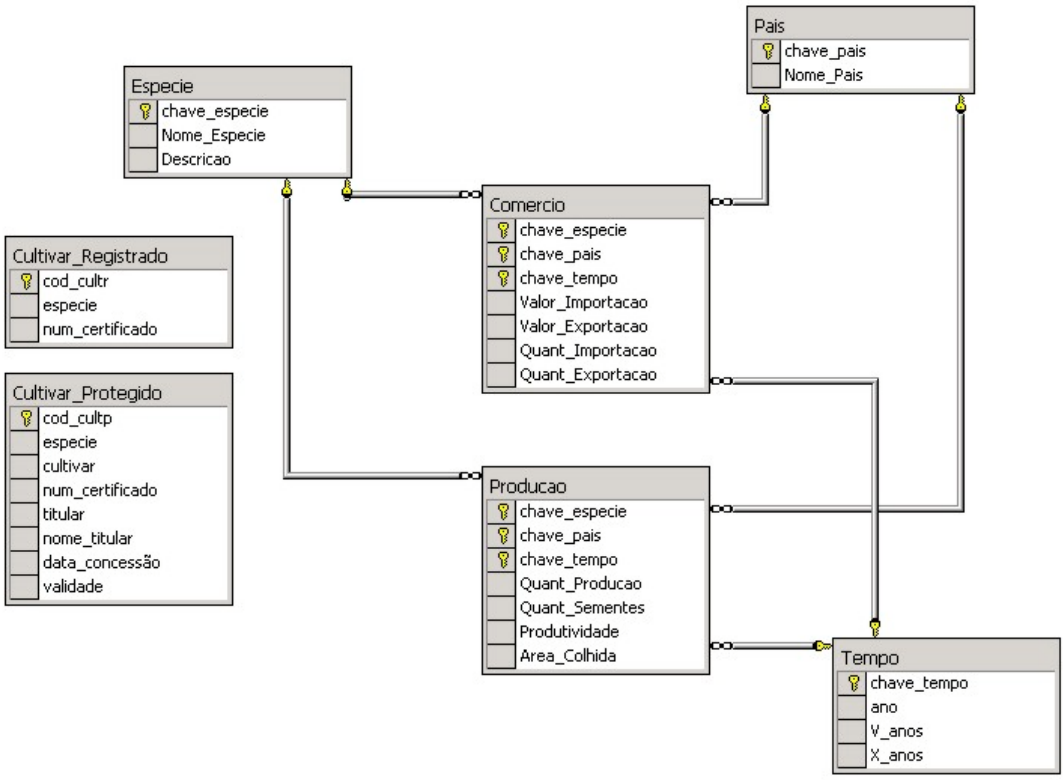

FIGURA 4.6 - Diagrama de relacionamentos das tabelas do BD.

4.4.3 Descrição das dimensões, fatos e campos

- Descrição das Tabelas de Dimensões:
  - *Nome*: **Espécie**;

    Onde é usada: nas pesquisas de produção e comércio;

    Como é usada: na indexação das pesquisas, pois possui os códigos das espécies pesquisadas.

  - *Nome*: **País**;

    Onde é usada: nas pesquisas de produção e comércio;

    Como é usada: na indexação destas pesquisas, por ter os códigos dos países.

- *Nome*: **Tempo**;

  Onde é usada: nas pesquisas de produção e comércio;

  Como é usada: na indexação das pesquisas, por ter os códigos dos anos, ou intervalos de tempo, escolhidos na consulta.

- Descrição das Tabelas de Fatos:
  - *Nome*: **Producao**;

    Onde é usada: nas pesquisas gerais a respeito da produção das espécies de sementes nos respectivos países e datas escolhidos;

    Como é usada: diretamente na pesquisa, cruzando, a partir da consulta dos usuários, os respectivos valores de espécie, país e tempo existentes nas tabelas de dimensão.

  - *Nome*: **Comercio**;

    Onde é usada: nas pesquisas gerais a respeito das importações e exportações das espécies de sementes nos países e datas escolhidos;

    Como é usada: diretamente na pesquisa, através do cruzamento dos dados de espécie, país e tempo, escolhidos pelo usuário para a pesquisa.

- Descrição das demais tabelas do BD:
  - *Nome*: **Cultivar_Protegido**;

    Onde é usada: nas listagens de dados sobre cultivares protegidos;

    Como é usada: na apresentação da listagem dos cultivares protegidos.

  - *Nome*: **Cultivar_Registrado**;

    Onde é usada: nas listagens de dados sobre cultivares registrados;

    Como é usada: na apresentação da listagem dos cultivares registrados.

- Descrição dos campos das tabelas de Dimensão e de Fatos:
  - *Nome*: **ano**;

    Tipo: Inteiro;

    Onde é usado: na tabela de dimensão Tempo;

    Como é usado: separa os registros, pois os dados são todos relativos ao ano em que ocorreram, são dados temporais;

    Conteúdo: contém os anos de ocorrência dos dados.

- *Nome*: **Area_Colhida**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Produção;

  Como é usado: informa à pesquisa a área em que as culturas foram colhidas, durante o ano ou intervalo de tempo pesquisado;

  Conteúdo: os dados referem-se à área onde as culturas foram colhidas. Estão excluídas as áreas onde não tenha havido colheita devido a danos, perdas, etc. A área colhida é expressa em hectares (ha).

- *Nome*: **chave_especie**;

  Tipo: Inteiro;

  Onde é usado: na tabela de dimensão Espécie e nas tabelas de fatos Producao e Comercio;

  Como é usado: relaciona o código da espécie pesquisada com seu nome;

  Conteúdo: contém os códigos das espécies, pertencentes à Dimensão Espécie.

- *Nome*: **chave_pais**;

  Tipo: Inteiro;

  Onde é usado: na tabela de dimensão País e nas tabelas de fatos Producao e Comercio;

  Como é usado: relaciona o código do país pesquisado com seu nome;

  Conteúdo: códigos dos países, pertencentes à tabela de dimensão País.

- *Nome*: **chave_tempo**;

  Tipo: Inteiro;

  Onde é usado: na tabela de dimensão Tempo e nas tabelas de fatos Producao e Comercio;

  Como é usado: relaciona o código pesquisado ao ano a que se refere;

  Conteúdo: contém os códigos dos anos.

- *Nome*: **Descrição**;

  Tipo: Caracter (Tamanho 255);

  Onde é usado: na tabela de dimensão Espécie;

  Como é usado: devolve ao usuário uma descrição geral da espécie a qual está pesquisando;

Conteúdo: contém uma descrição geral da espécie, como: de onde é originária, tamanho, cor.

- *Nome*: **Nome_Especie**;

  Tipo: Caracter (Tamanho10);

  Onde é usado: na tabela de dimensão Espécie;

  Como é usado: devolve à pesquisa o nome da espécie escolhida;

  Conteúdo: contém os nomes das espécies.

- *Nome*: **Nome_Pais**;

  Tipo: Caracter (Tamanho 10);

  Onde é usado: na tabela de dimensão País;

  Como é usado: devolve à pesquisa o nome do país escolhido;

  Conteúdo: contem os nomes dos países que integram o MERCOSUL.

- *Nome*: **Produtividade**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Producao;

  Como é usado: informa ao usuário os dados referentes à produção colhida por unidade de área plantada do cultivar, durante o ano, ou intervalo de tempo, pesquisado;

  Conteúdo: os dados representam a produção colhida por unidade de área plantada das culturas. Estão registrados em hectograma (hg - 100 gramas) por hectare (hg/ha).

- *Nome*: **Quant_Exportacao**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Comercio;

  Como é usado: informa dados a respeito da quantidade de semente exportada no país e no ano, ou intervalo de tempo, escolhidos pelo usuário no momento da consulta;

  Conteúdo: contém os dados sobre o comércio estrangeiro, especificamente sobre exportação, de forma quantitativa. A unidade de medida é a mesma (toneladas métricas) para quase todos os produtos.Em regra geral, os dados de exportação se referem à quantidade líquida.

- *Nome*: **Quant_Importacao**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Comercio;

  Como é usado: informa dados a respeito da quantidade de semente importada no país e no ano, ou intervalo de tempo, escolhidos pelo usuário no momento da consulta;

  Conteúdo: este elemento reporta o comércio estrangeiro, relativo a importação, de forma quantitativa. A unidade de medida é a mesma (toneladas métricas) para quase todos os produtos.Em regra geral, os dados de importação se referem à quantidade líquida.

- *Nome*: **Quant_Produção**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Producao;

  Como é usado: informa ao usuário como foi a produção interna total do cultivar durante o ano, ou intervalo de tempo, pesquisado;

  Conteúdo: os dados relacionam-se a produção interna total, não importando se dentro ou fora do setor agrícola, ou seja, incluem produção comercial e não comercial. Os dados são apresentados em toneladas métricas (tm).

- *Nome*: **Quant_Sementes**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Producao;

  Como é usado: informa aos usuários os dados referentes à quantidade de cultura usada para semear durante o ano escolhido na pesquisa;

  Conteúdo: Contém os dados de quantidade da cultura deixada para semear durante o ano, tanto a produzida internamente como a importada. São apresentados em toneladas métricas (tm).

- *Nome*: **V_anos**;

  Tipo: Inteiro;

  Onde é usado: na tabela de dimensões Tempo;

  Como é usado: separa os registros com os resultados totais relativos aos cinco anos pesquisados, com o objetivo de diminuir o tempo de resposta das consultas ao DW;

Conteúdo: somatória dos valores de produção e comercialização obtidos em um tempo de cinco anos.

- *Nome*: **Valor_Exportacao**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Comercio;

  Como é usado: informa ao usuário dados a respeito do valor que o país escolhido obteve com exportações feitas de determinado cultivar, no ano, ou intervalo de tempo, escolhido pelo usuário;

  Conteúdo: expressa o comércio estrangeiro em termos do valor das exportações. Os dados são armazenados em mil dólares americanos.

- *Nome*: **Valor_Importacao**;

  Tipo: Inteiro Longo;

  Onde é usado: na tabela de fatos Comercio;

  Como é usado: devolve ao usuário dados a respeito do valor que o país escolhido despendeu com importações feitas de determinado cultivar, no ano, ou intervalo de tempo, também escolhido pelo usuário;

  Conteúdo: expressa o comércio estrangeiro em termos do valor das importações. Os dados são armazenados em mil dólares americanos.

- *Nome*: **X_anos**;

  Tipo: Inteiro;

  Onde é usado: na tabela de dimensão Tempo;

  Como é usado: separa os registros com os resultados totais relativos aos dez anos pesquisados, com o objetivo de diminuir o tempo de resposta das consultas ao DW;

  Conteúdo: somatória dos valores de produção e comercialização obtidos em um tempo de dez anos.

4.5 Análise dos resultados obtidos

Em termos de aceitação pelo software utilizado e criação em geral, a modelagem desenvolvida neste trabalho foi bem sucedida ao ser usada no protótipo do DW apresentado.

Buscou-se desenvolver o protótipo ocupando o mínimo de espaço em disco, tendo em vista o grande volume de dados existente. Neste aspecto, acredita-se que também foi obtido êxito, pois, após povoar o BD com os dados de produção e comercialização das sementes de arroz, milho e soja, obtidos pelo Brasil dos anos de 1960 até 2001, ocupou-se um espaço de 3.1 Megabytes em disco, contando-se o espaço ocupado pelos registros armazenados por cada tabela e pelo esquema do BD em si. Sem ser considerado o espaço ocupado pelos cubos de dados, que é mínimo.

Ao final, cada tabela de fatos e cada tabela para listagens apresentaram a seguinte quantidade de registros: tabela Fato Produção - 492 registros; tabela Fato Comércio - 492 registros; tabela Cultivar_Protegido - 410 registros; tabela Cultivar_Registrado - 588 registros. Importante salientar que a totalidade dos registros das tabelas de fatos correspondem a dados numéricos, e que os registros das tabelas para listagens variam entre dados numéricos e textuais.

Este pouco espaço de armazenamento obtido se deve, além do fato da diminuição do escopo da aplicação para dados somente brasileiros e de três tipos de sementes, ao caráter dos dados, que são, em sua grande maioria, numéricos. Mas, como foi visto, o espaço armazenado é associado diretamente ao nível de granularidade dos dados e detalhamento das consultas, e não a velocidade com que os dados são processados e analisados. Com um volume menor de dados, possivelmente as consultas serão processadas mais rapidamente, mas não se tendo um DW funcionando com todos os dados de todos os países, para ser feito um comparativo, não há como se ter absoluta certeza da diferença de desempenho apresentada.

O protótipo foi criado e, tendo em vista a estruturação dos dados, foi bem sucedido. Os cubos foram processados pelo software Microsoft SQL Server Analysis Services e apresentaram boa performance. Os testes foram feitos em um aplicativo existente no próprio Microsoft SQL Server Analysis Services, chamada MDX Sample Application, o qual permite que as consultas sejam executadas e os resultados visualizados no mesmo ambiente. Foram feitas consultas aos cubos criados simulando as consultas feitas por usuários ao DW, estando ele localizado em um servidor real.

Nestas consultas foi utilizada a linguagem SQL padrão, com apenas algumas modificações impostas pelo aplicativo.

Estes testes realizados apresentaram resultados positivos a respeito da estruturação dos dados nas tabelas do BD, pois as consultas apresentaram os resultados esperados e demonstraram clareza no entendimento dos dados. Na fig 4.7 é mostrado o código em SQL utilizado em uma consulta feita ao cubo de dados Produção e o seu resultado, esta consulta apresenta o total obtido pelo Brasil, com relação à Produção, Quantidade de Sementes, Produtividade e Área Colhida, nos anos de 1960 até 2001 com as sementes de Arroz, Milho e Soja.

Consulta1:

```
select

  {[Measures].[Quant Producao], [Measures].[Quant Sementes],
  [Measures].[Produtividade], [Measures].[Area Colhida]} on columns,

  {TopCount([Espécie].[Nome Especie].members,3,
  [Measures].[Produtividade])} on rows

from Producao
```

Resultado:

	Quant Producao	Quant Sementes	Produtividade	Area Colhida
Arroz	346,203,325	14,676,725	765,538	191,518,134
Milho	855,470,208	11,699,791	750,742	454,264,413
Soja	551,190,573	12,187,411	675,988	288,066,490

FIGURA 4.7 - Consulta feita ao cubo Produção.

Na fig. 4.8 também é mostrado o código em SQL e o resultado de uma consulta utilizada, mas ao cubo de dados Comércio, nesta consulta é apresentado o total obtido pelo Brasil, com relação ao Valor de Importações, Quantidade de Importações, Valor de Exportações e Quantidade de Exportações, nos anos de 1981 à 1986 com as sementes de Arroz, Milho e Soja.

Consulta 2:

```
Select

  {[Measures].[Valor Importacao],[Measures].[Valor
  Exportacao],[Measures]. [QuantImportacao], [Measures].[Quant
  Exportacao] } on columns,

  order(Intersect
  ([Tempo].[Ano].members,{[1981],[1982],[1983],[1984],[1985],[1986]}),
  [Espécie]. [AllEspécie], ASC ) on rows

from Comercio

where ( [Pais].[Brasil])
```

Resultado:

	1984	1983	1985	1982	1981	1986
Valor Importacao	78,823	158,421	162,600	353,264	510,530	659,218
Valor Exportacao	478,119	381,637	765,052	181,225	424,912	245,148
Quant Importacao	388,748	562,264	793,421	1,388,476	1,975,861	4,097,509
Quant Exportacao	1,740,723	2,069,038	3,495,789	1,056,621	1,506,821	1,204,800

FIGURA 4.8 - Consulta feita ao cubo Comércio.

Não foram executados testes reais porque o software exige que o DW esteja localizado em um servidor, e não foi possível obter a infra-estrutura exigida pelo ambiente para o funcionamento do DW. Por esse motivo foi proposta a modelagem de um DW, e desenvolvido apenas um protótipo. Posteriormente, com incentivo cedido por alguma Instituição Educacional ou pela UFPel (Universidade Federal de Pelotas), poderão ser obtidos manuais, livros técnicos, softwares registrados e acesso ao servidor, provendo a infra-estrutura necessária para que este protótipo venha a compor efetivamente um DW.

Foi obtido êxito também na criação do *layout* do *site Web* para a publicação dos dados do DW, pois cada página possui, em média, 10 Kilobytes e não apresentou demora quando acessada pela *Internet* em uma velocidade de aproximadamente 40 Kilobytes por segundo, em uma conexão discada.

5. Conclusões e Trabalhos Futuros

5.1 Conclusões

Concluiu-se neste trabalho que, independentemente do assunto ou área em questão, a criação de *Data Webhouses*, tanto para divulgação de dados quanto para a captura de dados de usuários da *Web*, é o futuro de todo o projeto que se destina realmente a auxiliar na descoberta de conhecimento ou descobrir conhecimento. Pois atualmente as empresas, principalmente comerciais, devem expandir sua área de atuação para não perder lugar no mercado, e não há forma mais adequada do que através da *Internet*.

O uso de DWs vem a otimizar este crescimento das empresas, pois no momento em que estas possuem seus dados armazenados de forma planejada, não há motivo para temerem um crescimento no volume de dados a serem analisados para tomarem decisões, e estas também podem, a partir da experiência adquirida e de informações bem fundamentadas, planejarem o futuro.

O mercado de sementes do MERCOSUL, assunto específico tratado neste trabalho, é um setor que está em constante crescimento e, junto com ele, cresce também o volume de pessoas e informações ligadas a este mercado. Em virtude deste crescimento, a modelagem do *Data Webhouse* proposto neste trabalho mostrou-se necessária e foi bem aplicada, juntamente com a criação do seu protótipo.

O desenvolvimento, tanto da modelagem quanto do protótipo, foi bem sucedido, as consultas executadas no protótipo do DW mostraram resultados satisfatórios, comprovando a possibilidade da sua implementação definitiva. Infelizmente, por falta de recursos, não foi possível a sua implementação real. Mas espera-se que este trabalho venha a motivar e facilitar futuros projetos desenvolvidos na área.

5.2 Trabalhos Futuros

As tarefas de modelagem e criação de um DW são extensas e, para a execução de um projeto de sucesso, exigem grande dedicação por parte do desenvolvedor, o que pode levar vários meses, dependendo do projeto.

Mas estas não são as únicas tarefas executadas em um projeto de DW, pois após a sua carga ainda é necessário recarregar novos dados de atualização no DW, e esta tarefa exige a busca de novos dados, análise, limpeza, padronização, enfim, o projeto de DW consiste de uma execução cíclica de vários processos.

Portanto, para trabalhos futuros são sugeridos:

- Automatização dos processos de recarga de dados em DWs e *Data Webhouses*, através do desenvolvimento de metodologias de projeto e ferramentas para a extração dos dados de fontes heterogêneas e distribuídas, como a *Web*, visando incentivar o planejamento e a construção de Sistemas de Apoio a Decisão relacionados a assuntos em que seus dados não se encontram localmente armazenados por uma única fonte.
- Desenvolvimento de aplicativos que executem o direcionamento das transformações que devem ser feitas nos dados após serem extraídos, isto é, que a partir da escolha do tipo de padronização dos dados, analise os dados extraídos e aponte as possíveis modificações pelas quais deverão passar. Tornar os dados totalmente prontos para a recarga é uma tarefa muito ambiciosa para ser executada em um único trabalho, mas o desenvolvimento de um aplicativo que explicite os dados que devem ser transformados, juntamente com o padrão adotado no DW, seria de grande valia no auxilio a manutenção de DWs e *Data Webhouses*.

Anexo I - Telas do *site Web* criado

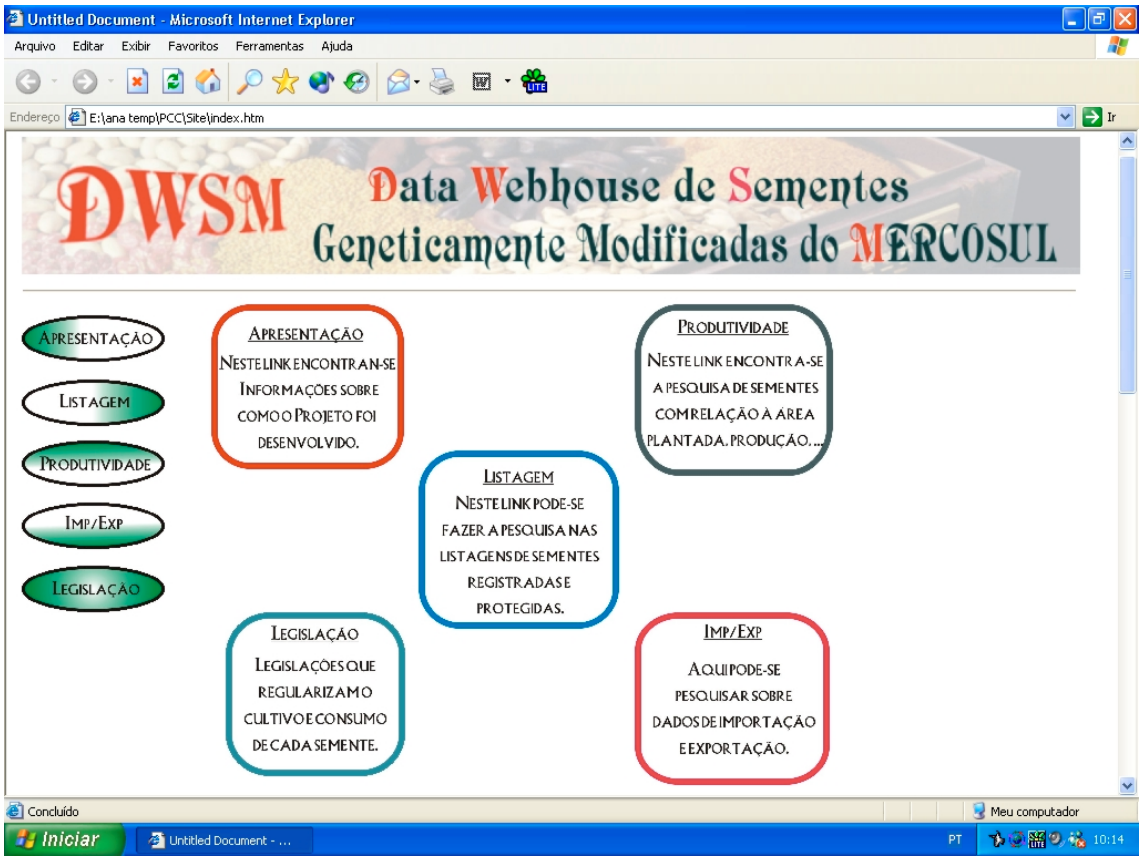

FIGURA I.1 - Tela inicial do *site*.

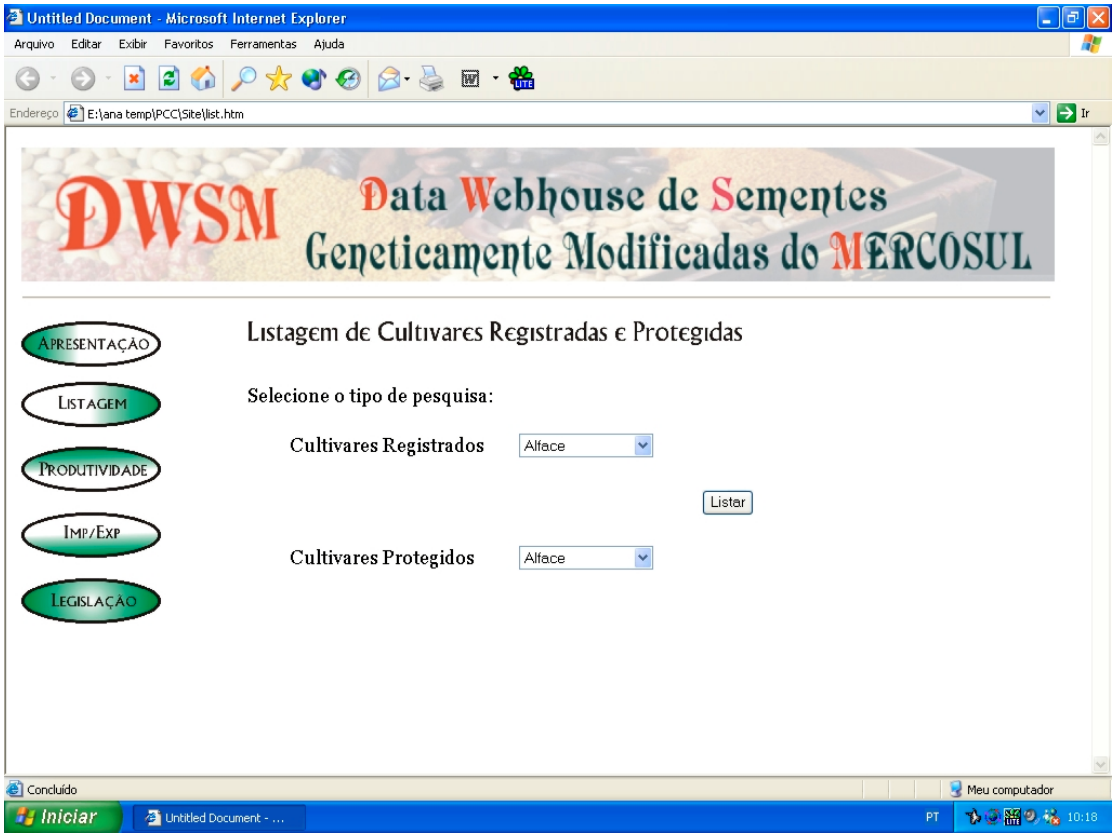

FIGURA I.2 - Tela para listagem dos Cultivares Registrados e Protegidos.

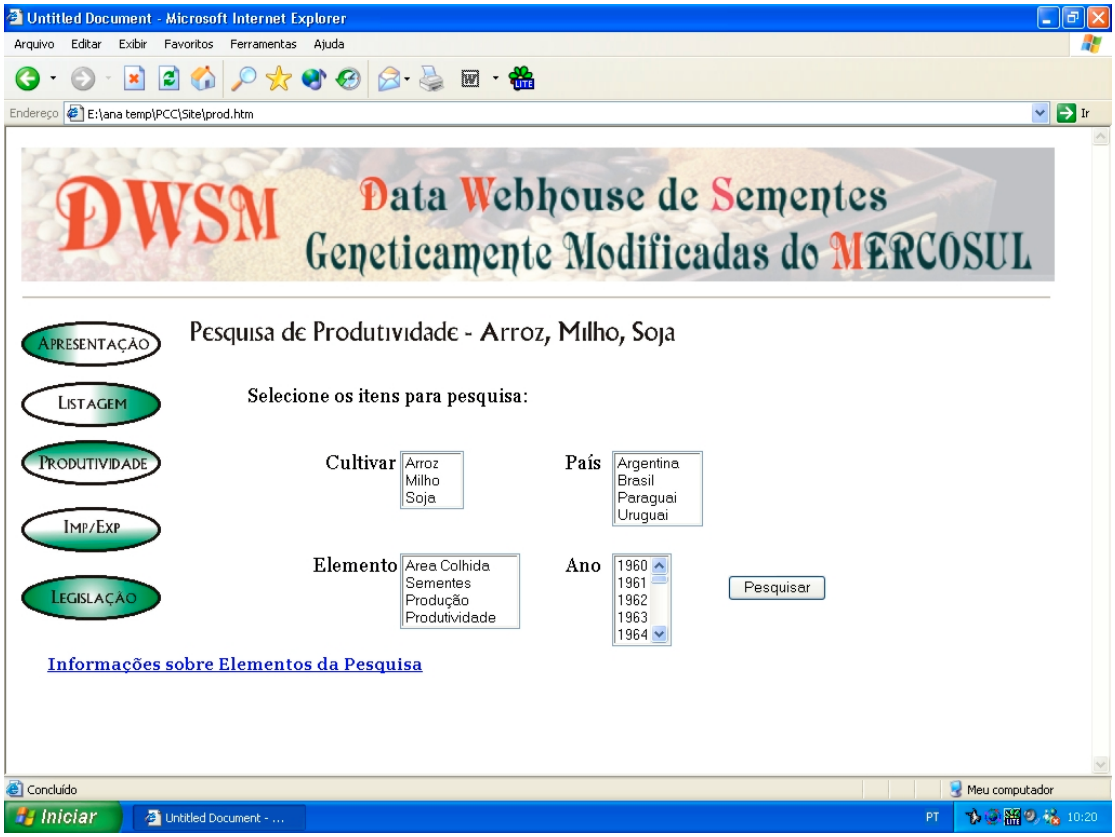

FIGURA I.3 - Tela de pesquisa sobre produção de sementes.

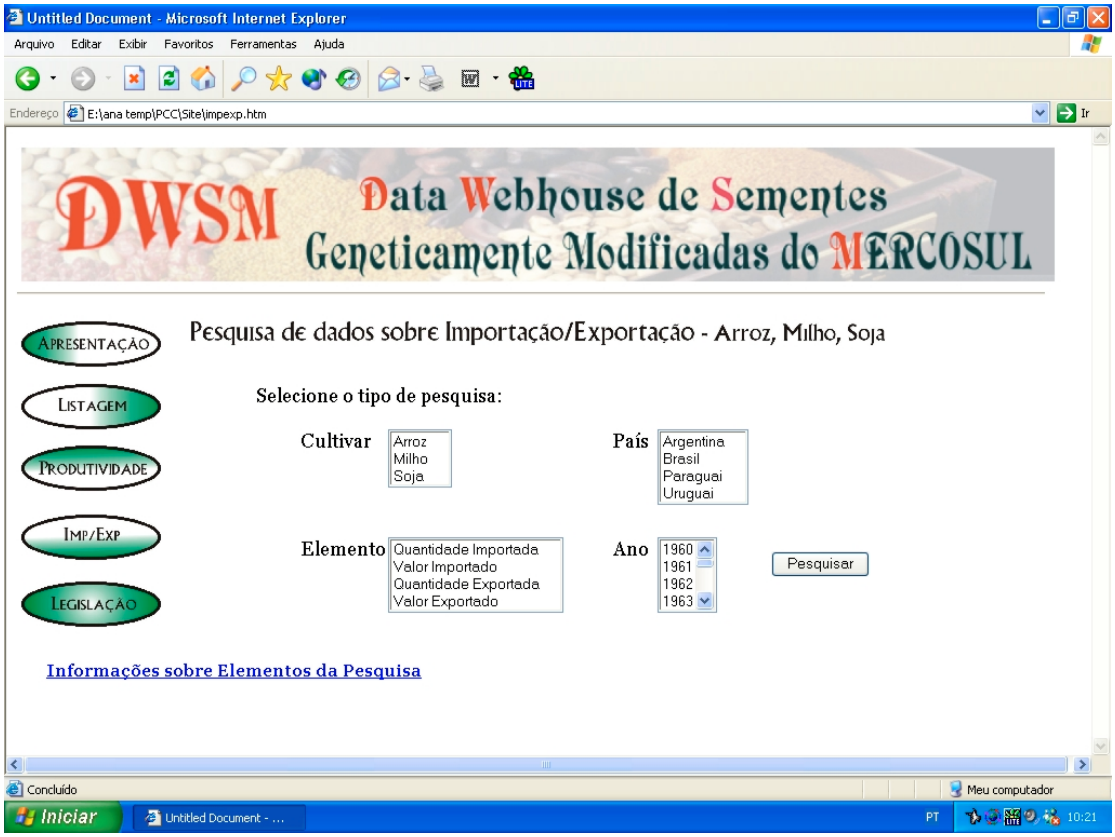

FIGURA I.4 - Tela de pesquisa sobre comercialização de sementes.

Bibliografia

[ALC 2000] ALCANTARA, Maria G. C. **Mercado de Sementes no MERCOSUL**. Pelotas, 2000. Tese (Doutorado) – Universidade Federal de Pelotas.

[ANN 2003] ANNES, Ricardo. **Sistemas de Apoio a Decisão**. Disponível por WWW em http://www.pucrs.campus2.br/~annes/pos_sad.html . (28 Jan. 2003).

[BER 97] BERSON, Alex, SMITH, Stephen. **Data Warehousing, Data Mining, and OLAP**. McGraw-Hill, 1997.

[BER 99] BERSON, Alex, et. al. **Building Data Mining Applications for CRM**. McGraw-Hill, 1999.

[CAM 2002] CAMPOS, M.L. Data Warehouse. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 17., 2002, Gramado. **Mini-Curso de Data Warehouse.** Porto Alegre: Instituto de Informática da UFRGS, 2002. p.2-56.

[CAM 2003] CAMPOS, M.L. ROCHA, Arnaldo V. Filho. **Tutorial de Data Warehouse**. Disponível por WWW em http://genesis.nce.ufrj.br/dataware/tutorial/home.html. (21 Jan. 2003).

[CIE 2002] CIELO, Ivã. **Data Mining**. Disponível por WWW em http://www.datawarehouse.inf.br. (18 Nov. 2002).

[FAO 2000] FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **FAOSTAT – Agriculture**. Disponível por WWW em http://www.fao.org. (15 Dez. 2002).

[GIM 2000] GIMENES, Eduardo. **"Data Mining – Data Warehouse" A Importância da Mineração de Dados em Tomadas de Decisões**. Taquaritinga, 2000. Monografia (Curso de Tecnólogo em Processamento de Dados) – Faculdade de Tecnologia de Taquaritinga. Disponível por WWW em http://geocities.yahoo.com.br/dugimenes/. (10 Jan. 2003).

[INM 97] INMON, W.H., **Como Construir o Data Warehouse.** Rio de Janeiro: Editora Campus, 1997.

[KIM 95] KIM, Won. **Modern Database Systems: the Object Model, Interoperability and Beyond**. ACM press, Addison Wesley, 1995.

[KIR 96] KIMBALL, Ralph**, The Data Warehouse Toolkit**. John Wiley & Sons, Inc, 1996.

[KIR 2000] KIMBALL, Ralph, MERZ, Richard, **Data Webhouse: construindo o Data Warehouse para a Web**. Rio de Janeiro: Editora Campus, 2000.

[PEL 2000] PELLEGRINI, Gisele F., COLLAZOS, Katia. **Extração de Conhecimento a partir dos Sistemas de Informação**. In VII Congresso Brasileiro de Informática em Saúde - CBIS 2000, outubro/2000. Disponível por WWW em http://www.inf.ufsc.br/~l3c/artigos/Pellegrini00.pdf . (21 Jan. 2003).

[PET 2001] PETROVIC, Dusan. **SQL SERVER 2000 – Guia Prático**. São Paulo: MAKRON Books Ltda, 2001.

[SIL 99] SILBERSCHATZ, Abraham, KORTH, Henry F., SUDARSHAN, S., **Sistemas de Banco de Dados**. 3ª ed. São Paulo: Makron Books, 1999.

[THE 2002] THEARLING, Kurt. **Glossary of Data Mining**. Disponível por WWW em http://www.thearling.com. (21 Jan. 2003).

[YIN 89] YIN, Robert K. **CASE Study Research: Design and Methods**. Sage Publications, Newbury Park, CA, 1989.

[WIE 99] WIEDERHOLD, Gio. **Mediation to Deal with Heterogeneous Data Sources**. In: Proceed of Interop'99. Zurich, 1999, pgs. 1-16. Disponível por WWW em http://hake.stanford.edu/pub/gio/. (10 Dez. 2002).