Analisando os Dados do Programa de Melhoramento Genético da Raça Nelore com *Data Warehousing* e *Data Mining*[1]

VALMIR FERREIRA MARQUES[2]

Orientadora:

Prof[a]. Dr[a]. Solange Oliveira Rezende[3]

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre na Área de Ciências de Computação e Matemática Computacional.

USP - São Carlos

Setembro/2002

---

[1]Trabalho realizado com o apoio da CAPES

[2]vfm@icmc.usp.br - valmirfmarques@yahoo.com.br - valmirfmarques@bol.com.br

[3]solange@icmc.usp.br

Este documento foi preparado com o formatador de textos LaTeX. O sistema de citações de referências bibliográficas utiliza o padrão *Chicago* do sistema bibTeX.

Todo homem, por natureza, deseja o Conhecimento.

Aristóteles(384-322 A.C.)

Dedicatória

À minha família.

Agradecimentos

São tantas pessoas a agradecer que vou evitar citar muitos nomes para não correr o risco de esquecer alguém. Vou apenas citar os nomes das pessoas que mais estiveram envolvidas com meu trabalho.

Quero agradecer primeiramente a DEUS, pela vida e por ter me dado forças pra conseguir romper barreiras.

À minha família, pelo apoio dado em todas as fases de minha vida, especialmente na graduação e agora no mestrado. Sem eles eu nunca teria conseguido chegar até aqui.

À Prof^{a}. Solange pela oportunidade, amizade, apoio e orientação.

Aos amigos do Labic. Não só do Labic, como do ICMC e da USP.

Aos funcionários e professores do ICMC e de outros órgãos da USP.

Ao pessoal do PMGRN, em especial ao Prof^{o}. Raysildo.

Ao Bruno Freitas, pelo apoio inicial a mim dado quando cheguei à São Carlos.

Às donas Antônia, Ana e Guiomar pelo suporte doméstico.

À USP, pela estrutura e pela qualidade do ensino e da pesquisa.

À CAPES, pelo apoio financeiro, sem o qual este trabalho não poderia ser viabilizado.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

A base de dados do Programa de Melhoramento Genético da Raça Nelore está crescendo consideravelmente, com isso, a criação de um ambiente que dê apoio à análise dos dados do Programa é de fundamental importância. As tecnologias que são utilizadas para a criação de um ambiente analítico são os processos de *Data Warehousing* e de *Data Mining*. Neste trabalho, foram construídos um *Data Warehouse* e consultas OLAP para fornecer visões multidimensionais dos dados. Além das análises realizadas com as consultas, também foi utilizada uma ferramenta de *Data Mining* Visual. O ambiente analítico desenvolvido proporciona aos pesquisadores e criadores do Programa um maior poder de análise de seus dados. Todo o processo de desenvolvimento desse ambiente é aqui apresentado.

Abstract

The Program of Genetic Improvement of Nelore Breed database have been growing considerably. Therefore, the creation of an environment to support the data analysis of Program is very important. The technologies that are used for the creation of an analytical environment are the Data Warehousing and the Data Mining processes. In this work, a Data Warehouse and OLAP consultations had been constructed to supply multidimensional views of the data. Beyond the analyses carried through with the consultations, a tool of Visual Data Mining also was used. The developed analytical environment provides to the researchers and cattlemen of the Program a greater power of data analysis. The whole process of development of this environment is presented here.

CAPÍTULO

1

Introdução

Muitas organizações automatizam seus negócios com sistemas baseados em interfaces sofisticadas que acessam bancos de dados[1] poderosos responsáveis pelo armazenamento de todas as informações relacionadas à área de atuação da mesma. A quantidade de dados armazenada nesses bancos geralmente está relacionada ao porte da organização.

Essas organizações normalmente investem na construção de bancos de dados que são utilizados exclusivamente para o armazenamento, atualização e consulta aos dados. Esses tipos de bancos são normalmente chamados de sistemas convencionais, operacionais ou de produção, orientados ao processamento de transações (OLTP[2]).

Esse tipo de automação auxilia sobremaneira as atividades diárias de uma organização. Porém, os dados armazenados podem trazer mais benefícios se forem sabiamente explorados. Atualmente, já é possível fazer análises de dados sofisticadas que possam dar apoio ao processo de tomada de decisão dessas organizações, tornando-as mais eficientes, produtivas e competitivas. Essa análise envolve o uso de visões multidimensionais das informações e a descoberta de relacionamentos implícitos nos dados.

Os avanços ocorridos na área de tecnologia da informação e o fato das bases de dados estarem armazenando grandes volumes de dados têm se tornado uma motivação maior para que as organizações com bases de tal ordem de grandeza, conheçam e entendam a fundo esse bem tão valioso que lhes pertence Decker &

[1] Os termos bancos de dados e bases de dados serão utilizados indistintamente neste trabalho.
[2] *On-Line Transactional Processing*

Focardi (1995); Li (1996). Contrastando com o passado, onde as bases de dados consistiam de pequenos volumes de informações e não existiam ferramentas adequadas que auxiliassem na análise das mesmas, devido, entre outros motivos, ao limite da tecnologia da época.

Dentre os principais avanços ocorridos em tecnologia da informação estão os processos de *Data Warehousing* e de *Data Mining*. O processo de *Data Warehousing* objetiva satisfazer as necessidades dos usuários quanto ao armazenamento dos dados que servirão para extrair e exibir de uma forma multidimensional, através das ferramentas OLAP[3], as informações necessárias aos usuários responsáveis pelas tomadas de decisões de uma organização Inmon (1997); Kimball (1997); Poe et al. (1998).

O processo de *Data Mining* objetiva automatizar o processo de extração de conhecimento à partir dos dados armazenados, auxiliando na descoberta de relações embutidas nos dados, sendo seu objetivo principal encontrar padrões válidos e úteis nos mesmos Holsheimer et al. (1995); Fayyad et al. (1996).

Uma base de dados para a qual se faz necessária a aplicação dos processos de *Data Warehousing* e de *Data Mining*, é a do Programa de Melhoramento Genético da Raça Nelore (PMGRN), uma vez que a mesma está aumentando consideravelmente de tamanho, devido à aderência de muitos criadores ao Programa. Essa base de dados contém informações sobre as fazendas participantes do Programa, bem como, de seus respectivos animais. Sobre os animais são armazenadas, entre outras, informações sobre medidas de peso e perímetro escrotal, informações relacionadas à reprodução dos animais e informações sobre DEPs e medidas ponderais.

A existência do PMGRN se deve ao fato do Brasil ser um país com uma população bovina estimada em 148,1 milhões de cabeças, e possuir um dos menores índices de produtividade do mundo no setor, considerando-se as baixas taxas de natalidade, altas taxas de mortalidade, longos intervalos entre partos e idades elevadas para o abate e primeiro parto FNP (1995). As causas dessa baixa produtividade são inúmeras e vão desde uma desinformação dos criadores, até a falta de planejamento estratégico para o setor agropecuário. Por esses motivos, a pecuária de corte brasileira precisa utilizar novas tecnologias que produzam animais geneticamente superiores e que transmitam precocidade, maior eficiência reprodutiva e velocidade de ganho de peso à sua progênie.

Em junho de 1988 teve início o PMGRN a partir da parceria entre criadores e

[3] *On-Line Analytical Processing*.

pesquisadores do Departamento de Genética da Faculdade de Medicina (DGFM), da Universidade de São Paulo (USP), campus de Ribeirão Preto-SP, em busca de tecnologias modernas e de fácil aplicação na pecuária e definição de metas para viabilizar o aumento da produtividade do rebanho de corte nacional.

Por outro lado, em 1997, o DGFM em parceria com o Laboratório de Inteligência Computacional (LABIC) do Instituto de Ciências Matemáticas e de Computação (ICMC) da USP, campus de São Carlos-SP, começaram a trabalhar em um projeto de pesquisa para analisar as informações do PMGRN, com a finalidade de encontrar situações interessantes, na base de dados do Programa, que possam vir a auxiliar nas pesquisas relacionadas ao melhoramento genético da raça Nelore.

Tendo em vista as tecnologias apresentadas, o constante crescimento da base de dados do PMGRN e o fato dos criadores e pesquisadores não possuírem um ambiente analítico que ofereça os recursos que um ambiente desse tipo pode proporcionar, fica assim, evidente a necessidade de aplicação das tecnologias de *Data Warehousing* e de *Data Mining* sobre os dados do Programa.

Visando desenvolver um ambiente analítico para o PMGRN, este trabalho teve como objetivo a construção de um *Data Warehouse* para armazenar os dados em um formato que agilize a execução de consultas OLAP, com o intuito de fornecer uma visão multidimensional dos dados aos especialistas do Programa. Além da construção do *Data Warehouse* e da elaboração de consultas OLAP, este trabalho também teve como objetivo pré-processar os dados para que os especialistas do programa possam extrair conhecimento visualmente, utilizando uma ferramenta de *Data Mining* Visual.

A finalidade deste capítulo foi contextualizar, motivar e apresentar os objetivos atingidos. O Capítulo 2, por sua vez, apresenta uma contextualização sobre o PMGRN. O Capítulo 3 enfatiza as tecnologias utilizadas no processo de análise de dados, nele são apresentados conceitos relacionados a *Data Warehousing* e *Data Mining*, bem como, alguns elementos de apoio a esse processo e alguns problemas relacionados ao mesmo. No Capítulo 4 são apresentadas as ferramentas utilizadas neste trabalho. O Capítulo 5 apresenta todo o processo de desenvolvimento do *Data Warehouse*. O Capítulo 6 aborda como os dados poderão ser analisados, utilizando o *Data Warehouse* construído. No Capítulo 7 são apresentadas as conclusões deste trabalho e, por fim, é apresentada toda a referência bibliográfica utilizada para a elaboração desta dissertação de mestrado.

CAPÍTULO

2

O Programa de Melhoramento Genético da Raça Nelore

Este capítulo objetiva contextualizar o domínio do PMGRN. Estudar sobre o domínio do problema em questão é uma atividade necessária para a aplicação dos processo de *Data Warehousing* e de *Data Mining*.

O PMGRN avalia várias características genéticas dos animais cadastrados e as publica anualmente em um sumário. Por meio desse sumário os criadores podem selecionar os melhores animais para procriar.

Com o objetivo de apresentar o PMGRN, este capítulo foi estruturado em duas seções. A Seção 2.1 enfatiza os sumários produzidos pelo Programa, bem como as características avaliadas e a definição e interpretação da Diferença Esperada na Progênie (DEP). A Seção 2.2 apresenta os objetivos do Programa, os resultados atingidos, as avaliações realizadas e alguns outros assuntos relacionados à compreensão do domínio.

2.1 Sumários do PMGRN

Como parte das atividades de extensão de serviços à comunidade e retribuição ao esforço da sociedade para a manutenção das universidades e instituições de pesquisa, vem se desenvolvendo no Departamento de Genética da FMRP-USP, sob a coordenação do Prof. Dr. Raysildo B. Lôbo, o Programa de Melhoramento Genético da Raça Nelore. O mesmo teve início em junho de 1988, com uma primeira reunião entre pesquisadores da FMRP-USP e um grupo de criadores, em Ribeirão Preto-SP. Foi uma reunião histórica, onde os pesquisadores mostraram

o caminho a ser seguido com avanços e respaldos técnicos e os criadores falaram dos trabalhos desenvolvidos em suas propriedades. Hoje, com parâmetros definidos e um programa em execução, o Prof. Raysildo, juntamente com uma equipe qualificada, vem desenvolvendo pesquisas e formação de pessoal especializado de alto nível, bem como criando e conduzindo programas de melhoramento genético em diversas raças.

O PMGRN publica anualmente um sumário com informações do Programa, juntamente com alguns dados dos principais animais cadastrados. A edição do Sumário 2002 apresenta notável progresso no tocante ao crescimento e aperfeiçoamento desse trabalho de melhoramento genético. Atualmente, a base de dados conta com 1.277.850 pesagens, 183.346 medidas de perímetro escrotal e 429.806 animais cadastrados, distribuídos em 12 estados e 2 países, em um total de 180 rebanhos participantes da Avaliação Genética Lôbo et al. (2002).

Examinando a Figura 2.1, nota-se que o número de animais da Avaliação Genética de 1997 para a de 2002 aumentou consideravelmente, evidenciando a expansão e a abrangência cada vez maior do PMGRN.

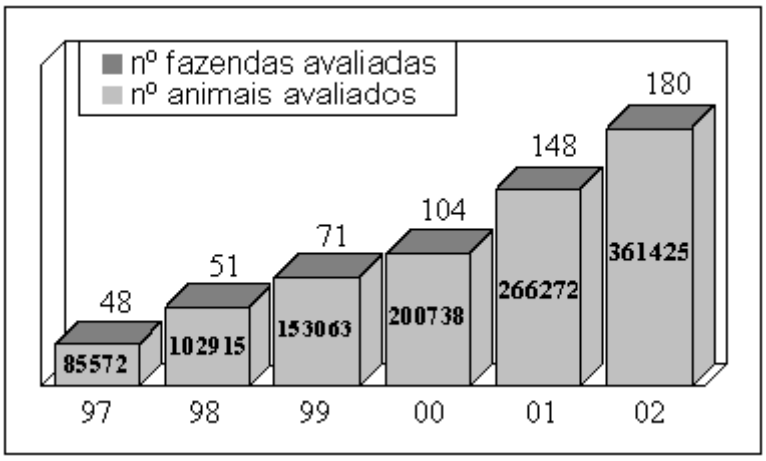


Figura 2.1: Evolução do número de animais e de fazendas analisadas no período.

No ano de 2000, o PMGRN expandiu de forma significativa sua base de dados, devido ao ingresso de muitas fazendas, resultante, principalmente, do apoio oficial da Associação de Criadores de Nelore do Brasil (ACNB). O número de animais avaliados atingiu o total de 200.738, representando um aumento de 31% em relação ao ano de 1999. No ano de 2001 houve um aumento de 31% no número de animais e 41% no número de fazendas em relação ao ano de 2000, ou seja, o Programa passou a ter 266.272 animais e 148 fazendas avaliados. Já no ano de 2002 houve um aumento de 35% no número de animais e 21% no

número de fazendas em relação ao ano de 2001. Dessa forma, no ano de 2002 o Programa avaliou 361.425 animais e 180 fazendas.

Esse crescimento representa um expressivo aumento no tamanho do banco de dados do Programa, o que possibilita uma significativa melhoria do nível de acurácia das informações apresentadas. O suporte material para viabilizar o incremento do banco de dados, inclusive projetando seu crescimento para os próximos anos, deu-se pela aquisição de um novo multi-processador com 4 Gbytes de memória RAM, multiplicando em 8 vezes a antiga capacidade de memória RAM, que era de 0,5 Gbytes. Novos e mais freqüentes relatórios poderão agora estar sendo fornecidos aos participantes do Programa, agilizando os diversos trabalhos de seleção.

Nos sumários publicados pelo Programa são avaliadas diferentes características. A seguir são apresentadas essas características, juntamente com uma breve descrição sobre as mesmas Lôbo et al. (2002).

Idade ao Primeiro Parto (IPP): é uma característica importante como indicadora da precocidade sexual, além de afetar a produtividade pela sua influência na produção de bezerros durante a vida útil da matriz e na eficiência reprodutiva do rebanho. Touros com DEPs negativas, expressando os dias a menos para o primeiro parto, devem ser utilizados.

Período de Gestação (PG): característica de pequena variação, tem reflexos econômicos na pecuária zebuína, uma vez que esse é extremamente longo se comparado com o dos taurinos. É também importante por estar relacionada ao peso ao nascer e com partos distócicos. Touros com DEPs negativas, expressando os dias a menos de duração da gestação, devem ser utilizados.

Peso ao Nascer (PN): é a primeira informação do animal, mostrando o seu vigor e desenvolvimento pré-natal, sendo um indicador da facilidade de parto. Touros com DEPs baixas ou negativas são desejáveis para essa característica.

Peso Adulto (PA): o peso corporal, considerado como indicador do tamanho adulto do animal, foi definido como o primeiro peso obtido (kg) dos 4 aos 12 anos de idade. O PA tem relação com os custos de manutenção em vacas e com a velocidade de crescimento do animal. Touros com DEPs baixas ou médias são desejáveis para essa característica.

Produtividade Acumulada (PAC): é um índice que indica a produtividade da fêmea, em kilogramas de bezerros desmamados por ano. Mede a capacidade do animal em se reproduzir regularmente e a uma menor idade e desmamar animais com maior peso. Touros com maiores DEPs devem ser utilizados, pois o criador estará selecionando para a habilidade maternal, fertilidade e precocidade sexual.

Peso aos 120 dias: o peso, em kilogramas, aos 120 dias, é importante, já que nesse período ocorre o pico de lactação na raça Nelore. O mesmo é relevante para os produtores de bezerros e foi usado como âncora nas análises bicaracteres para minimizar os efeitos de descartes realizados até a desmama. Touros com DEPs mais elevadas são os mais indicados, tomando-se o cuidado de examinar a DEP do PA.

Peso aos 365 (450) dias: é o peso, em kilogramas, aos 12 (15) meses de idade do animal e expressa o potencial de ganho em peso no período pós-desmama. Touros com DEPs mais elevadas são os mais indicados, tomando-se o cuidado de examinar a DEP do PA.

Perímetro Escrotal (PE) aos 365 (450) dias: essa medida é expressa em centímetros, sendo tomada trimestralmente dos 9 aos 18 meses de idade. É importante na seleção de bovinos de corte, pela correlação favorável com a fertilidade e a precocidade sexual. Touros com DEPs mais elevadas são os mais indicados.

O criador selecionador, ao escolher animais para o acasalamento, deve em primeiro lugar levar em conta as DEPs que satisfaçam os seus objetivos ou critérios de seleção. O PMGRN apresenta um índice denominado Mérito Genético Total (MGT), com o intuito de fornecer ao criador a oportunidade de escolher animais geneticamente superiores, porém, harmonicamente balanceados para a habilidade maternal, fertilidade e crescimento pré e pós-desmame Lôbo et al. (2002).

A DEP é usada em todo o mundo para comparar o mérito genético de animais para várias características e prediz a habilidade de transmissão genética de um animal avaliado como progenitor. Ela é expressa na unidade da característica, por exemplo: kg para peso, cm para PE e meses para IPP, com sinal positivo ou negativo.

A DEP para efeito direto é um preditor da habilidade de um animal em transmitir genes para crescimento ou fertilidade à sua progênie. A DEP para efeito maternal na característica período de gestação, peso ao nascer e aos 120 dias prediz a diferença esperada em peso, duração da gestação, das progênies das filhas do animal avaliado, devido às diferenças na habilidade maternal apresentada por elas.

A DEP para efeito direto na produtividade acumulada prediz a habilidade do animal em transmitir à sua progênie genes para a capacidade de se manter a produção durante toda vida do animal. A DEP para efeito maternal total é obtida somando-se metade da DEP direta para peso aos 120 dias à DEP maternal da mesma característica. A mesma expressa, em kilogramas, o potencial de desmama que um animal pode transmitir, incluindo a habilidade de um animal em transmitir genes para crescimento e produção de leite para as suas filhas.

Com o intuito de elucidar ainda mais o conceito de DEP, considere um exemplo com o peso ao sobreano em gado de corte. Se a DEP (P450) para o touro A for de 10 kg e a DEP para o touro B de -5 kg, a diferença média entre as progênies de A e B será de 15 kg. Isso significa que podemos esperar que a progênie do touro A produza, em média, 15 kg a mais em peso aos 450 dias que a do touro B, sob as mesmas condições de criação. Esse valor reflete a diferença no valor genético médio dos gametas produzidos pelos touros, pois o material genético dos pais é transmitido à sua descendência por meio dos seus gametas. O valor genético médio dos gametas produzidos pelos reprodutores é que determina a habilidade de transmissão genética dos mesmos Lôbo et al. (2002). Convém ressaltar que experimentos realizados em Rezende et al. (2000) confirmam que o alto poder preditivo das DEPs dos pais afetam o desempenho reprodutivo de sua progênie.

2.2 Objetivos, Resultados e Avaliações

Os principais objetivos do PMGRN são aumentar a eficiência reprodutiva e a taxa de crescimento nos rebanhos de corte, assim como, estabelecer critérios de seleção mediante a aplicação de técnicas clássicas de melhoramento genético animal e de modernas biotecnologias que possibilitem um aumento significativo da produtividade nacional PMGRN (2002).

Com relação aos resultados atingidos pelo programa, a Tabela 2.1 sumariza o número de informações e médias de desempenho reprodutivo das matrizes do PMGRN. Os resultados são satisfatórios, cabendo destacar o número de serviços por concepção (1,4), o intervalo entre partos (404 dias) e a fertilidade real (175

kilogramas de bezerro desmamado por ano).

CARACTERÍSTICAS	NÚMERO	MÉDIA
Peso ao parto (kg)	10.649	463,0
Peso à desmama (kg)	5.740	462,0
Variação do peso no aleitamento (%)	3.486	-0,9
Peso em abril (kg)	10.993	477,0
Peso em outubro (kg)	8.926	448,0
Intervalo entre partos (dias)	17.180	404,0
Número de serviços/concepção	13.368	1,4
Fertilidade real (kg)	16.494	175,0
Relação de desmama (%)	7.926	43,4
Idade ao primeiro parto (meses)	6.285	36,0
Idade média atual (meses)	35.676	85,0
Referência: Novembro de 1995		

Tabela 2.1: Comportamento reprodutivo médio das vacas do PMGRN PMGRN (2002).

Na Figura 2.2 é mostrado o número de pesagens realizadas no PMGRN, por ano, cabendo destacar um aumento acentuado a partir de 1993. O período de 1988 a 1993 engloba pesagens mensais na maioria dos rebanhos, enquanto que, a partir de 1994 e, com maior ênfase, em 1995, as pesagens de animais jovens, até 21 meses de idade, passaram a ser realizadas a cada três meses. Na Tabela 2.2 são apresentadas médias de pesos e perímetros escrotais (PE) às idades-padrão de animais participantes do PMGRN mantidos a pasto.

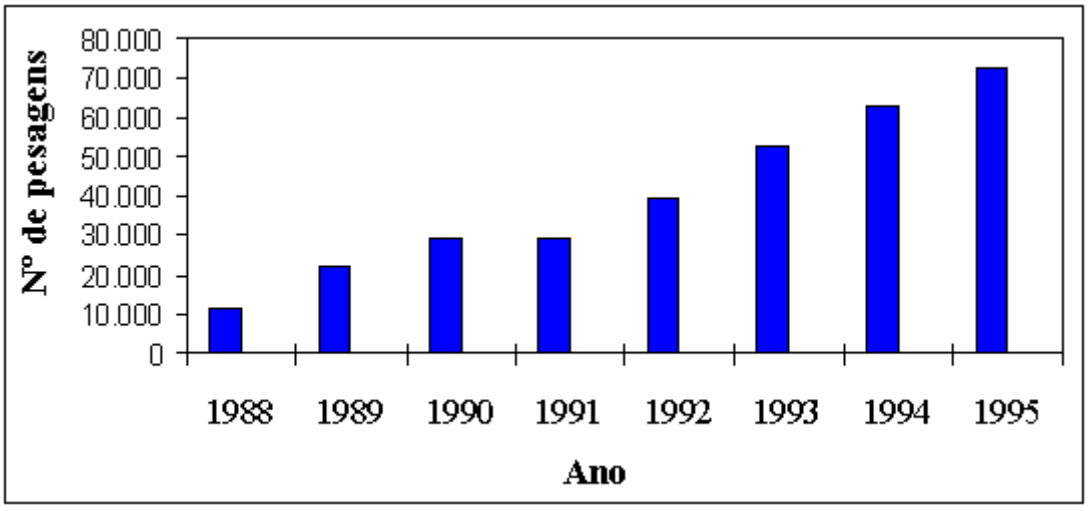


Figura 2.2: Número de pesagens realizadas por ano PMGRN (2002).

Analisando-se o desempenho ponderal dos animais participantes do PMGRN nos últimos anos, foi elaborado um indicador capaz de expressar a precocidade

	MACHO		FÊMEA	
CARACTERÍSTICA	**N**	**MÉDIA**	**N**	**MÉDIA**
Peso aos 120 dias de idade (kg)	12.349	123,50	10.793	113,92
Peso aos 240 dias de idade (kg)	11.003	201,83	9.819	182,84
Peso aos 365 dias de idade (kg)	9.871	244,82	9.432	216,42
Peso aos 550 dias de idade (kg)	6.642	330,91	6.683	284,05
PE aos 365 dias de idade (cm)	3.772	19,62	-	-
PE aos 550 dias de idade (cm)	3.603	25,92	-	-
Referência: 16/04/96				

Tabela 2.2: Médias de pesos e perímetros escrotais PMGRN (2002).

de peso. Considerando-se que, atualmente, os animais jovens estão mais pesados que os de gerações anteriores a uma mesma idade, foi adequado propor a idade como função do peso. Tomando-se pesos de particular importância, 350 kg para machos e 300 kg para fêmeas, pode-se evidenciar na Figura 2.3 a redução da idade para os animais atingirem os pesos propostos, considerando-se os anos de atuação do PMGRN.

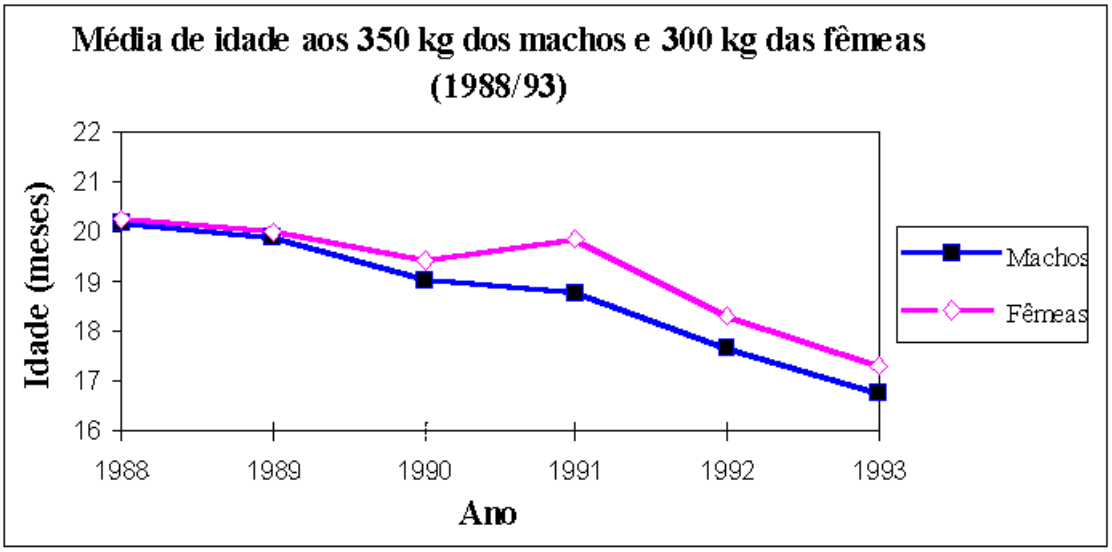


Figura 2.3: Precocidade de peso dos animais do PMGRN PMGRN (2002).

Examinando-se a Figura 2.4, nota-se que houve uma resposta à seleção no decorrer do período estudado. Cabe ressaltar que as DEPs para efeito direto para peso aos 240 dias de idade (DEPDP240) e peso aos 365 dias de idade (DEPDP365) mostraram a mesma tendência de crescimento. Contudo, verifica-se que não houve mudança genética para habilidade materna, expressa pela DEP para efeito materno para peso aos 240 dias de idade (DEPMP240). Isto indica que o Programa deverá intensificar a seleção genética para essa característica

nos próximos anos.

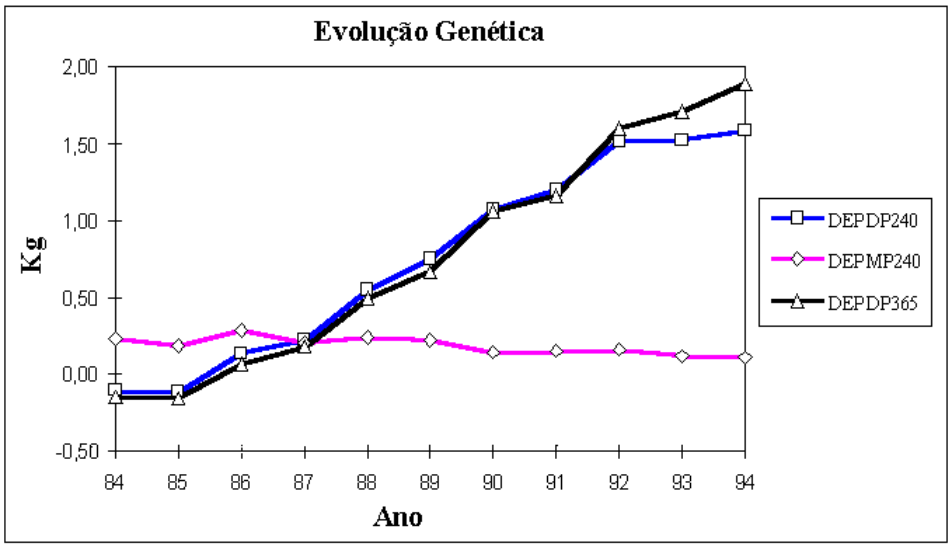


Figura 2.4: Evolução genética no período de 1984 a 1994 PMGRN (2002).

O PMGRN tem procurado, desde o início, aplicar a metodologia mais adequada para a avaliação genética dos animais dos rebanhos participantes, e não ignorou a possibilidade e as vantagens que poderia alcançar com a realização de provas de desempenho individual de machos jovens integrando animais dos rebanhos do PMGRN e outros, a fim de obter uma amostra mais representativa da raça e propiciar a difusão mais efetiva de material genético superior. Com essa finalidade foi criado o Centro de Avaliação de Touros Jovens (CAT), cujo objetivo é submeter animais pré-selecionados, com idade entre 200 e 290 dias, a um mesmo manejo durante 448 dias para avaliação de características de crescimento e reprodução, sendo que a cada ano, cerca de 100 animais pré-selecionados com base na DEP para o peso aos 120 dias de idade são submetidos a essa avaliação PMGRN (2002).

Além do CAT, em 1988 foi criado o Laboratório de Micromanipulação de Embriões (LME) com o objetivo de dar suporte ao PMGRN com a produção *in vitro* de embriões geneticamente superiores, a seleção precoce de touros jovens quanto à fertilidade e o estabelecimento de um método de criopreservação de oócitos e embriões para a formação do banco de gametas e embriões do PMGRN.

Na avaliação genética dos animais, para a determinação do ganho genético e para a estimação dos efeitos direto e materno das características de peso e perímetro escrotal a determinadas idades, são utilizados os procedimentos BLUP (*Best Linear Unbiased Predictor*) sob o Modelo Animal. Embora as metodologias

BLUP sejam de grande utilidade para a obtenção de DEPs confiáveis, sozinhas elas não garantem um aumento considerável no ganho genético. É preciso, em primeiro lugar, que os dados de genealogia e de produção sejam confiáveis. Em segundo lugar, é necessária uma correta definição dos modelos de análise por parte do melhorador e, finalmente, a aplicação adequada dos resultados da avaliação genética por parte do criador também é vital para produzir maior ganho genético anual.

A avaliação visual por escore objetiva identificar animais que reúnam maior número de características de importância econômica e melhorar alguns aspectos relacionados à composição de peso do animal. Para isso, propõe-se que as seguintes características sejam avaliadas PMGRN (2002): musculosidade, estrutura física, aspectos raciais e sexuais, conformação e ônfalo, reunidas na sigla MERCO. Na musculosidade deve ser considerada a distribuição muscular no corpo do animal, bem como o seu desenvolvimento e deve-se valorizar animais com precocidade de desenvolvimento muscular. Na estrutura física deve-se analisar a sustentação do animal. Com relação aos aspectos raciais e sexuais, as características produtivas e reprodutivas são pouco ou nada influenciadas por características raciais, sendo observadas apenas aquelas que podem influenciar de uma forma negativa o desempenho do animal. Na conformação o peso do animal está relacionado ao tamanho do esqueleto e a forma do corpo com o maior ou menor teor de gordura. Os animais devem ter não só grande peso ou tamanho, mas também baixo teor de gordura e alto rendimento de carcaça. Quanto ao ônfalo, sabe-se que machos de umbigo excessivamente comprido, criados extensivamente, podem ferir o prepúcio em talos de gramíneas, podendo comprometer o órgão reprodutor. Dessa forma, são valorizados os futuros reprodutores, machos e fêmeas, que possuem umbigo com forma ideal e tamanho reduzido.

Este escore tem como principal característica a facilidade de aplicação por técnicos que possuam conhecimento básico no assunto e que passem por um treinamento prático. Para cada característica avaliada o animal pode obter de 1 a 5 pontos, sendo que a maior pontuação representa o grau mais favorável. Um animal que, comparado ao seu grupo de contemporâneos (GC), for considerado intermediário (3 pontos) para determinadas características, servirá de referência para a classificação dos demais abaixo (1 ou 2 pontos) ou acima da média (4 ou 5 pontos). Em resumo, a avaliação é comparativa, onde a pontuação dada a um animal é sempre relativa aos demais. Outro aspecto importante é que os pontos não devem ser totalizados, evitando-se, desta forma, que defeitos sejam

mascarados PMGRN (2002).

2.3 Considerações Finais

Com as informações apresentadas neste capítulo pôde-se observar que o PMGRN tem conseguido alcançar resultados satisfatórios com suas pesquisas relacionadas ao melhoramento animal da raça Nelore. Conseguiu-se uma melhora significativa de algumas características, tais como, a diminuição do tempo para o animal atingir um determinado peso, ou o crescimento de determinadas DEPs, sendo esse resultado alcançado principalmente pelo processo de seleção aplicado.

Além do mais, foi apresentado que os dados do PMGRN tem aumentado significadamente à cada ano. São mais fazendas e animais cadastrados. Como conseqüência, tem havido um aumento significativo na quantidade de pesagens e medidas de perímetro escrotal, tornando assim, mais evidente a necessidade de construção de um ambiente que dê apoio à análise desses dados. O próximo capítulo se focará em apresentar as tecnologias e metodologias que são utilizadas para o desenvolvimento e análise dos dados desse tipo de ambiente.

CAPÍTULO

3

Uma Visão Geral de *Data Warehousing* e *Data Mining*

Atualmente, uma importante questão estratégica para o sucesso de uma organização está relacionada à sua capacidade de analisar e reagir rapidamente às mudanças nas condições de seus empreendimentos. Para que isso ocorra, torna-se necessário que a organização disponha de mais e melhores informações. Os avanços na área de tecnologia da informação estão possibilitando que essas organizações possam manipular um grande volume de dados.

Diariamente, dados sobre os diversos negócios de uma organização são gerados e armazenados, passando a fazer parte do patrimônio de informações da mesma. Porém, essas informações encontram-se, muitas vezes, espalhadas por diferentes sistemas e exigem um esforço considerável para serem integradas, para então, poderem dar apoio efetivo à tomada de decisão. A partir dessas considerações, pode-se verificar que, embora tenham ocorridos avanços tecnológicos de armazenamento e manipulação dos dados, ainda se observa uma enorme deficiência na obtenção eficiente de informações estratégicas que possam auxiliar o processo decisório.

Em vista disso, um novo conjunto de tecnologias vem ganhando um certo destaque na atualidade. Uma delas é o processo de *Data Warehousing*, que oferece às organizações uma maneira flexível e eficiente de se obter informações, à partir dos dados, que apóiem seus processos de tomada de decisão. A outra é o processo de *Data Mining*, sendo o mesmo definido como um processo de

extração de conhecimento válido e previamente desconhecido, à partir de uma base de dados, sendo que os dados que compõem a base de dados podem ser originários de várias fontes. Na Figura 3.1 pode ser observada a diferença entre os resultados obtidos por essas tecnologias.

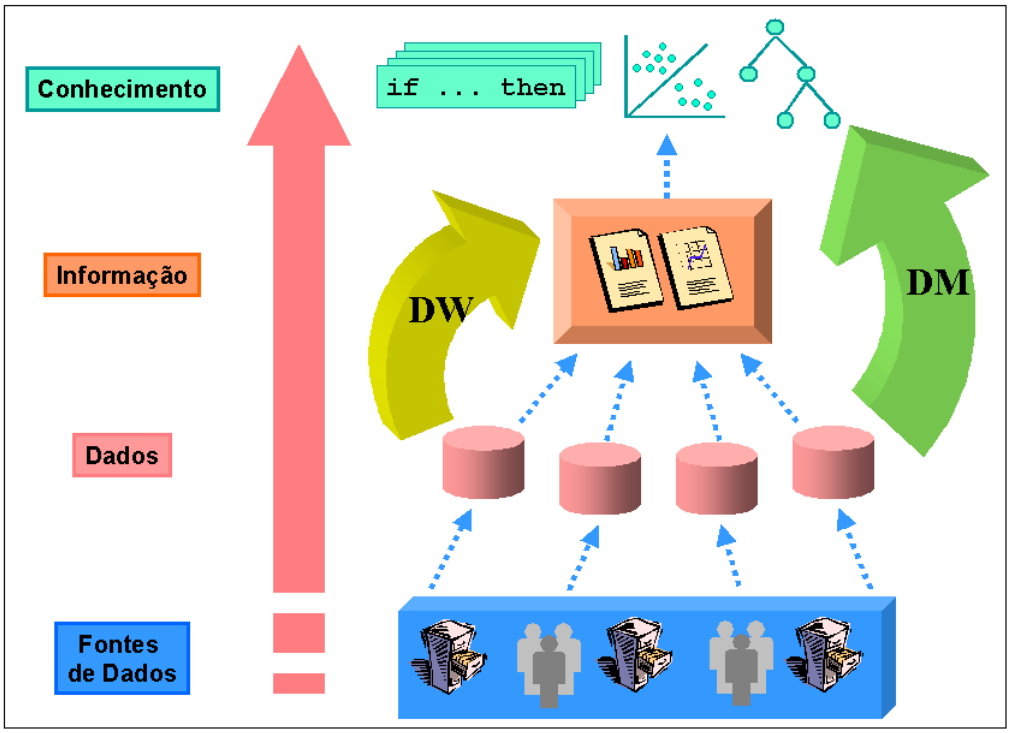


Figura 3.1: Resultados obtidos com DW e DM Rezende & Moreira (2000).

Vale ressaltar neste momento a diferença entre os resultados obtidos com *Data Warehousing* e *Data Mining*, ou seja, a diferença entre informação e conhecimento. A informação é obtida quando atribui-se algum significado aos dados, já o conhecimento é gerado quando consegue-se elaborar uma regra ou relação sobre os dados, sendo a mesma confrontada com uma informação.

Com o objetivo de apresentar uma visão geral desses conceitos, este capítulo está estruturado da seguinte forma: na Seção 3.1 são enfatizados os principais conceitos relacionados a *Data Warehousing*; na Seção 3.2 é apresentado o processo de *Data Mining*; na Seção 3.3 são apresentados alguns elementos de apoio ao processo de análise de dados; na Seção 3.4 são mostrados alguns problemas relacionados a esse processo e, por fim, na Seção 3.5 são realizadas algumas considerações finais sobre este capítulo.

3.1 Data Warehousing

O processo de construção, acesso e manutenção de um *Data Warehouse* (DW) é denominado de *Data Warehousing*. Esse processo objetiva integrar e gerenciar dados extraídos de diversas fontes, com o propósito de ganhar uma visão detalhada de parte ou do todo de um negócio.

Um DW é um banco de dados cuja função é proporcionar aos seus usuários uma única fonte de informação a respeito dos seus negócios, servindo também como ferramenta de apoio ao processo de extração de conhecimento. Além disso, é responsável pelo agrupamento dos dados históricos de uma organização, sejam eles provenientes de qualquer tipo de banco de dados, planilhas eletrônicas, documentos textuais, entre outros. Assim, um DW é um grande repositório de dados, obtidos a partir de várias fontes, que tem diferenças fundamentais em relação aos bancos de dados convencionais Inmon (1997).

Um *Data Warehouse* é uma coleção de dados orientada por assuntos, integrada, variante no tempo e não volátil, que tem por objetivo dar apoio ao processo de tomada de decisão. A seguir será apresentada uma descrição de cada uma dessas características Inmon (1997); Poe et al. (1998):

Orientado por Assuntos: um DW sempre armazena dados importantes sobre temas específicos da organização e conforme o interesse das pessoas que irão utilizá-lo.

Integrado: um DW deve ser capaz de integrar dados provenientes de fontes de dados distintas para obter uma representação única.

Variante no Tempo: os dados são dependentes do tempo. A cada mudança ocorrida na base de dados operacional, uma nova entrada deve ser criada no DW, a fim de representar essa mudança. Dessa forma garante-se o histórico das alterações ocorridas nos dados. Essa característica pode ser exemplificada com as mudanças de endereço de um cliente. Em uma época as vendas para o cliente foram realizadas enquanto ela morava em um dado endereço, e em outra época, enquanto ele morava em outro endereço.

Não Volátil: uma vez que um dado é inserido no DW, ele não pode ser modificado ou excluído. Sempre que houver uma atualização no mesmo, um novo item de dado é criado para representar essa mudança.

As bases de dados operacionais são utilizadas para realizar tarefas básicas, ou seja, aquelas tarefas que constituem o dia a dia das operações de uma organização. Por outro lado, o *Data Warehouse* é utilizado para apoiar o processo de tomada de decisão, logo, dados históricos e resumidos são mais importantes do que registros detalhados. Na Tabela 3.1 são apresentadas algumas diferenças entre BDs operacionais e DWs Inmon (1997); Barquini (1996); Kimball (1997); Poe et al. (1998).

Característica	**BD Operacional**	**Data Warehouse**
Objetivo	Operacional	Informativo
Processamento	OLTP	OLAP
Operação	Transações Simples	Consultas Complexas
Número de Usuários	Milhares	Centenas
Usuário	Operadores Projetistas de Sistema Admin. de Sistema	Analistas Executivos Usuários do Conhecimento
Condições dos Dados	Dados Operacionais	Dados Analíticos
Volume	MB e GB	GB e TB
Histórico	60 a 90 dias	5 a 10 anos
Granularidade	Detalhados	Detalhados e Agregados
Acesso a Registros	Dezenas	Milhares
Atualização	Contínua (tempo real)	Periódica
Modelagem	Entidade-Relacionamento	Dimensional
Integridade	Transação	A cada atualização
Número de Índices	Poucos/Simples	Muitos/Complexos
Intenção dos Índices	Localizar um registro	Aperfeiçoar consultas
Junções	Muitas	Poucas

Tabela 3.1: Diferenças entre Base de Dados Operacional e *Data Warehouse*.

De acordo com as diferenças apresentadas, os implementadores de sistemas de *Data Warehousing* bem sucedidos, descobriram ser necessário criar um banco de dados separado logicamente e, muitas vezes, fisicamente das fontes de dados. Essa separação se deve às diferenças encontradas nos dados manipulados por cada sistema, na tecnologia envolvida, nos usuários e nas características de processamento.

Nesta seção são abordados os principais tópicos relacionados a *Data Warehouse*, tais como: topologias, arquitetura e ferramentas, metadados, metodologia de desenvolvimento, modelagem multidimensional, OLAP, povoamento e apoio a extração de conhecimento.

3.1.1 Topologias

Um DW pode ser implementado utilizando-se diferentes topologias, sendo as principais: Centralizada, *Data Marts* e Distribuída Kimball (1997); Gardner (1998); Samos et al. (1998).

Na topologia Centralizada, Figura 3.2, um único *Data Warehouse* concentra todas as informações disponíveis da organização, ou seja, os dados históricos e operacionais são extraídos e integrados em um grande repositório. Esse tipo de DW possui uma topologia simples, pois estão inseridos nele todas as informações disponíveis da organização, ou seja, ele contém os dados de todas as áreas e processos da mesma.

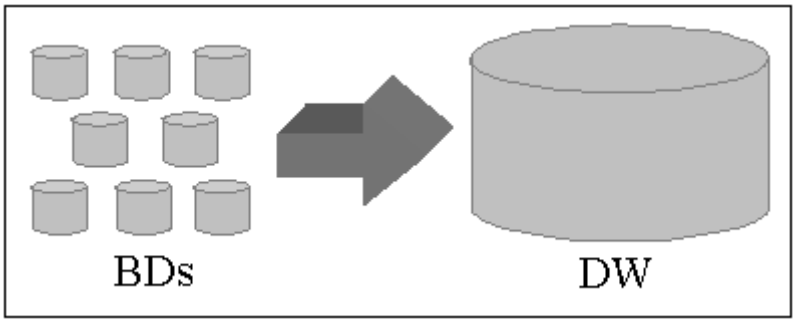


Figura 3.2: Topologia Centralizada.

A topologia *Data Mart* objetiva organizar cada departamento de uma organização, sendo que cada um possui seu próprio repositório de informações. Podem ser independentes de um DW, Figura 3.3, possuindo dados de uma determinada seção de uma organização, ou podem ser dependentes, Figura 3.4, onde vários *Data Marts* são criados à partir de um *Data Warehouse*.

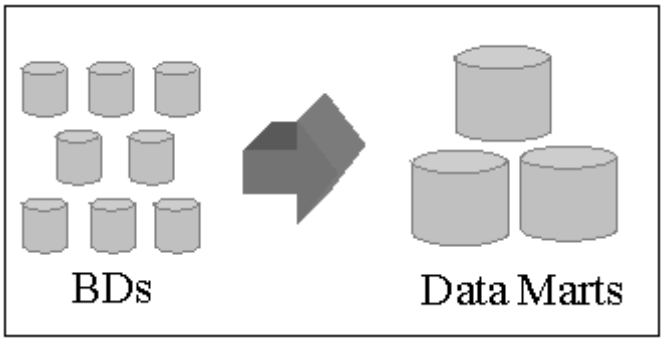


Figura 3.3: Topologia *Data Marts* independentes.

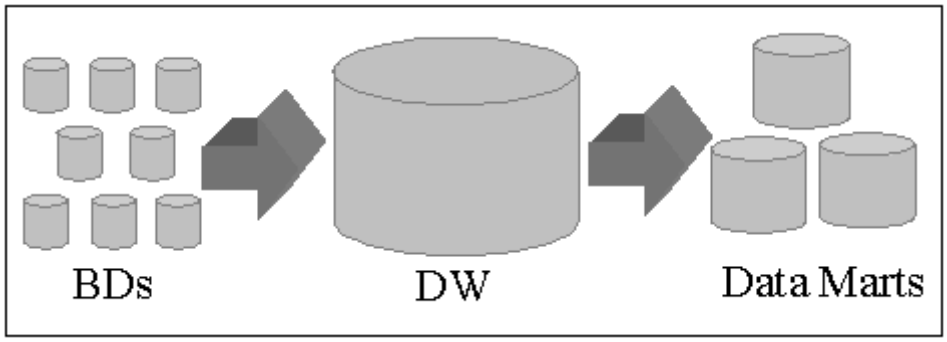


Figura 3.4: Topologia *Data Marts* dependentes.

Esse tipo de topologia procura otimizar análises para obter melhores resultados nas tomadas de decisões. Às vezes, torna-se mais interessante montar *Data Marts* independentes, uma vez que são mais simples e rápidos de serem implementados. Por outro lado, caso uma organização queira criar um DW formado a partir de *Data Marts* individuais para realizar análises sobre um escopo mais geral, pode descobrir que esta será um tarefa desgastante ou mesmo inviável, caso a especificação do projeto dos *Data Marts* não tenha considerado esse fato.

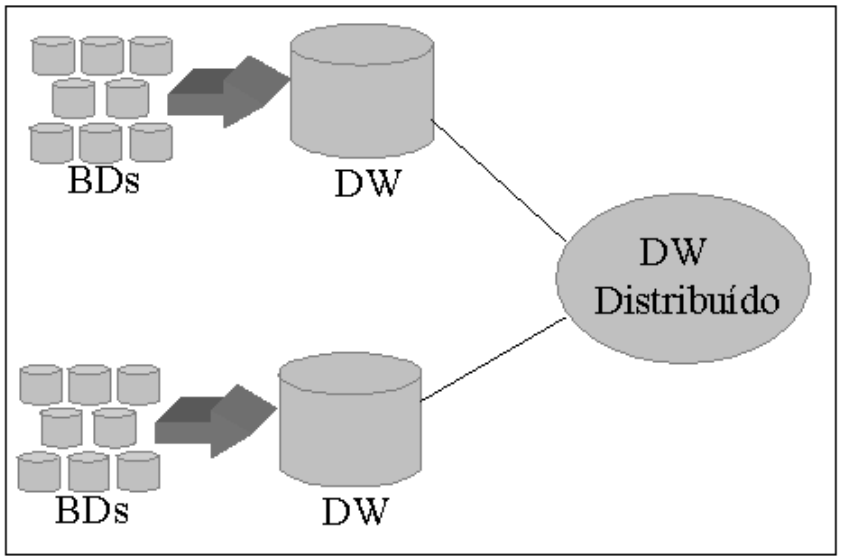


Figura 3.5: Topologia Distribuída.

E, por fim, a topologia Distribuída, Figura 3.5, consiste de vários repositórios de dados conectados por uma rede de computadores com apoio a processamento distribuído. Os usuários desse tipo de topologia podem acessar qualquer um dos repositórios de forma transparente, ou seja, todos eles são vistos como um único DW. Dependendo da tecnologia de comunicação de dados empregada, essa

topologia pode reduzir em muito o desempenho de um sistema. Dessa forma, muitas vezes a sua implementação só é viável quando os requisitos do sistema não exijam que um grande volume de operações sejam realizados remotamente e as cargas de dados não sejam muito pesadas e freqüentes.

As topologias apresentadas estão de alguma forma englobadas em uma arquitetura de DW.

3.1.2 Arquitetura e Ferramentas

O conceito de *Data Warehousing* evoluiu para uma arquitetura voltada para a extração de informação especializada à partir dos dados operacionais de uma organização e exibição desses dados utilizando ferramentas de visualização multidimensionais. A seguir será descrita uma arquitetura ideal para projeto de sistemas de *Data Warehousing*, sendo a mesma ilustrada na Figura 3.6 Corey et al. (2001):

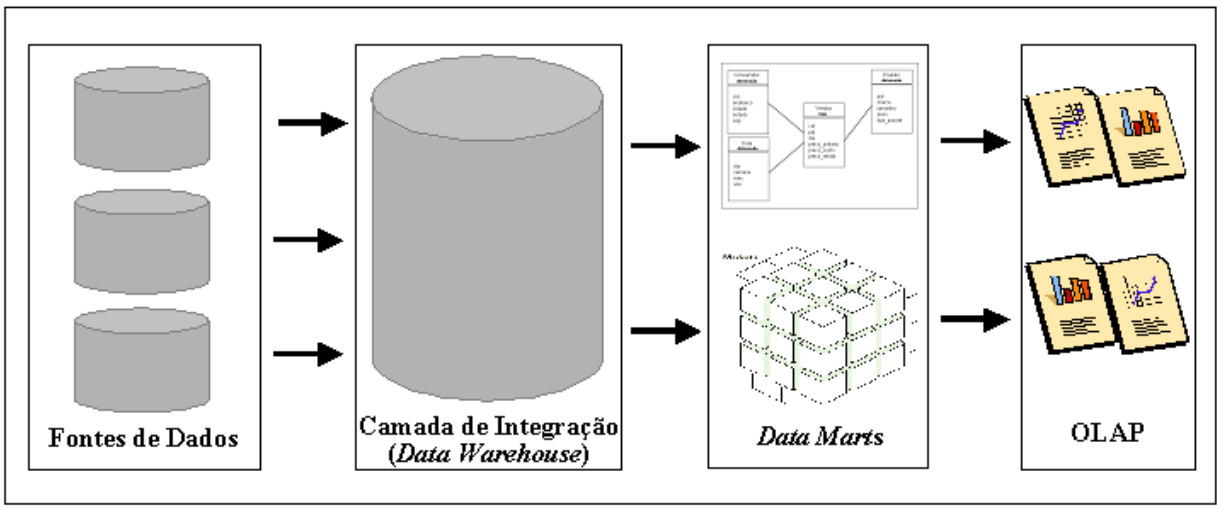


Figura 3.6: Uma arquitetura para *Data Warehousing*.

Nessa arquitetura o fluxo dos dados inicia-se com a extração dos mesmos das fontes e carga no *Data Warehouse* ou camada de integração, sendo que, antes de serem carregados no DW, os dados deverão passar por um processo de transformação e limpeza.

Uma área de adaptação é um depósito de dados na qual os dados dos sistemas de origem podem ser integrados, transformados, limpos e preparados para carregamento no *Data Warehouse*. Dependendo do trabalho, a área de adaptação pode ser física ou virtual. No caso de grandes trabalhos, que possuem várias áreas de assuntos e bancos de dados de destino, o ideal é utilizar uma área de

adaptação física. As áreas de adaptação virtuais são o mesmo que as áreas de adaptação físicas, no sentido do que elas fazem, mas o tamanho de seu trabalho permite que as mesmas funções sejam executadas dinamicamente na memória Corey et al. (2001). Na arquitetura apresentada, ela está localizada entre as fontes de dados e a camada de integração.

Após os dados estarem armazenados no DW, eles já podem ser transferidos para os *Data Marts* ou estruturas de consulta de alto desempenho. Os *Data Marts* são acessados pelos analistas, gerentes e executivos, utilizando as ferramentas OLAP, para buscarem as informações, que são exibidas num formato multidimensional, a qual darão um amplo apoio ao processo de tomada de decisão.

Nessa arquitetura o *Data Warehouse* é uma camada onde os dados provenientes de diversas fontes são integrados, logo, os dados estão num formato normalizado. Já o *Data Mart* é uma estrutura de consulta de alto desempenho, pois os dados estão representados em uma forma desnormalizada, para que as consultas sejam executadas mais eficientemente, pois são necessários menos junções de tabelas quando os dados forem recuperados.

Apesar das consultas serem realizadas sobre o *Data Mart*, o fato de se ter uma camada de integração evita a repetição da extração, pois é provável que vários *Data Marts* exijam dados das mesmas fontes. Se os dados não forem trazidos dessas fontes, através de um repositório comum, então cada *Data Mart* terá que acessar cada fonte. Além do mais, um DW garante uma interpretação padronizada dos dados e fornece um repositório que é bem mais flexível do que as estruturas desnormalizadas dos *Data Mart*.

Existem outras estratégias alternativas que podem ser adotadas para o construção do DW, como construir somente a Camada de Integração sem *Data Marts*, ou construir apenas os *Data Marts* sem a Camada de Integração, ou ainda, não construir nem a Camada de Integração nem os *Data Marts*, e as consultas serem realizadas diretamente nas fontes. Claro que a escolha vai depender das necessidades do usuários quanto ao desempenho das consultas e outros aspectos Corey et al. (2001).

Existem vários tipos de ferramentas utilizadas sob um *Data Warehouse* Orli (2001): ferramentas para armazenamento, extração, transformação e limpeza de dados; repositórios de metadados; transferência de dados e replicação; gerenciamento e administração; e gerenciamento de consultas e de relatórios.

Além dessas, as ferramentas OLAP e as ferramentas utilizadas no processo de

Data Mining, são outros tipos de ferramentas que se beneficiam das características de um DW, pois os resultados obtidos com as mesmas auxiliam efetivamente aos tomadores de decisão.

As ferramentas de *Data Mining* são necessárias quando deseja-se extrair conhecimento de um repositório de dados. Já todas as outras ferramentas são necessárias em um ambiente de *Data Warehouse*.

3.1.3 Metadados

Metadados são normalmente definidos como dados sobre os dados, ou uma abstração dos dados, ou ainda, dados de mais alto nível que descrevem dados de um nível inferior Gupta (1997); Sherman (1997).

Os metadados constituem-se no principal recurso para a administração dos dados e assumem uma maior importância no ambiente de *Data Warehouse*. Em um ambiente operacional, os metadados são importantes para os desenvolvedores e administradores do banco de dados. Por outro lado, o ambiente de apoio à tomada de decisão é bastante distinto, sendo que nele os analistas de dados procuram por fatos não usuais. Seus usuários precisam examinar seus dados e, para isso, devem conhecer sua estrutura e significado Inmon et al. (1999).

Exemplos de metadados são informações sobre os modelos lógicos utilizados na especificação da dimensionalidade e no processamento analítico de um DW. Os metadados são responsáveis também pela gerência do sistema como um todo, indicando de onde os dados vêm, como são transformados, quando são atualizados, o que significam, quem os vê, e assim por diante.

3.1.4 Metodologia de Desenvolvimento

A escolha correta da estratégia a ser adotada é fundamental para se obter sucesso no desenvolvimento de um *Data Warehouse*, sendo que a mesma deve ser adequada às características e necessidades específicas do ambiente onde ele será instalado. Existem várias abordagens para o desenvolvimento de um DW, devendo-se fazer uma escolha fundamentada em pelo menos três dimensões: o escopo do DW, o grau de redundância dos dados e o tipo de usuário alvo Weldon (1997); Orr (2000).

A especificação dos requisitos do ambiente de apoio à decisão é diferente da especificação dos sistemas do ambiente operacional de uma organização. Por exemplo, os requisitos dos sistemas do ambiente operacional são identificáveis a partir das funções a serem executadas pelo sistema. Os requisitos dos siste-

mas de suporte à decisão são, por sua vez, indeterminados. O principal objetivo de um *Data Warehouse* é prover dados com qualidade mas os requisitos dependem das necessidades de informações individuais de seus usuários. Ao mesmo tempo, os requisitos dos sistemas do ambiente operacional são relativamente estáveis ao longo do tempo, enquanto que os dos sistemas de suporte à decisão são instáveis, ou seja, dependem das variações das necessidades de informações dos responsáveis pelas tomadas de decisões.

Na realidade, é difícil apontar, no momento, uma metodologia consolidada e amplamente aceita para o desenvolvimento de um *Data Warehouse*. O que se encontra na literatura e nos exemplos de sucesso de implementações são propostas no sentido de se construir um modelo dimensional a partir do modelo de dados operacional da organização de forma incremental. De qualquer forma, a metodologia a ser adotada é ainda bastante dependente da abordagem escolhida, em termos de ambiente, distribuição, etc. Barquini (1996); Kimball (1997); Inmon (1997); Poe et al. (1998).

A Figura 3.7 apresenta uma seqüência de passos, que pode servir de guia para o projeto de repositórios de dados Barquini (1996); Kimball (1997); Inmon (1997); Poe et al. (1998). A metodologia é composta por seis fases, onde as cinco últimas devem se repetir para cada nova área de negócio a ser considerada no projeto.

A primeira fase corresponde a justificativa de um projeto do DW. O objetivo é procurar identificar quais são as vantagens que um sistema de DW trará à organização. Caso o projeto se justifique, deve-se então tomar lugar a fase de planejamento. Essa fase objetiva a definição de uma visão corporativa do *Data Warehouse*, definição da arquitetura e topologia do sistema, e definição da área de negócio a ser enfatizada. Em seguida vem a fase de análise da área de negócio, cuja tarefa principal é a modelagem conceitual dos dados para a área de negócio selecionada. A próxima fase corresponde ao projeto do sistema, sendo que a mesma enfatiza o detalhamento dos resultados adquiridos na fase de análise, bem como, o projeto das consultas OLAP. A quinta fase é a implementação, onde são criados os objetos físicos, é feito o povoamento do *Data Warehouse* e a implementação das consultas OLAP. A última fase corresponde à revisão, onde são verificados os resultados obtidos com a implementação do sistema. Um documento final deve registrar todo conhecimento obtido durante o projeto da área de negócio selecionada, servindo como modelo para as próximas iterações do projeto.

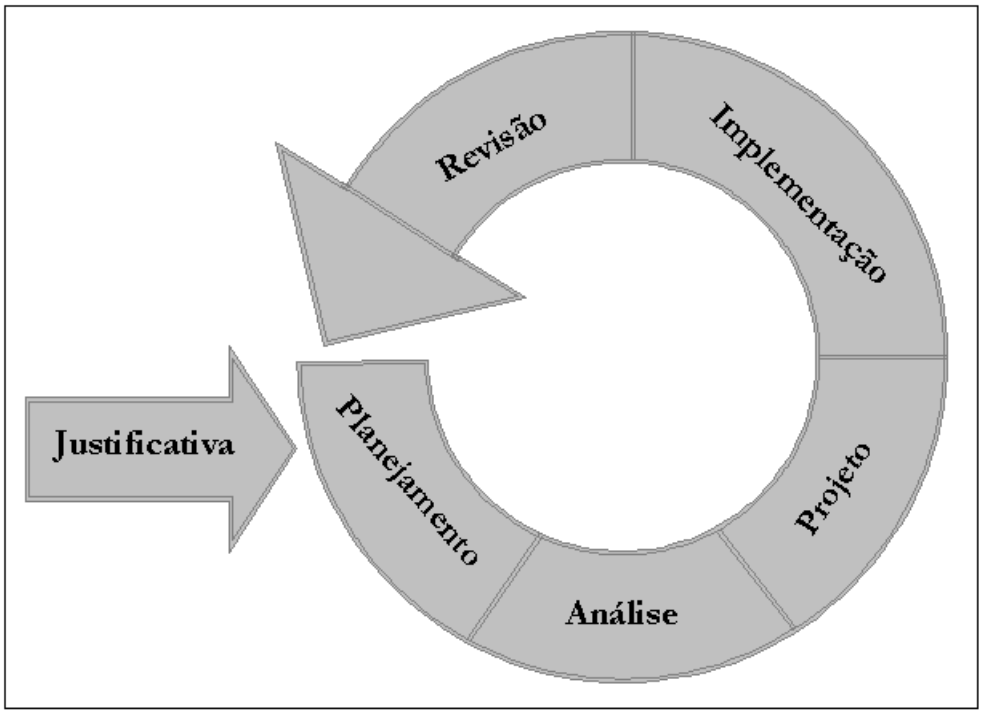


Figura 3.7: Metodologia para o desenvolvimento de DWs.

Um outro tópico que deve ser considerado no projeto de um *Data Warehouse* é a sua granularidade, a qual se refere ao nível de detalhe em que as unidades de dados são mantidas. Por exemplo, para uma informação que envolva quantidade de tempo, o nível de detalhe pode ser dias, meses ou anos. Quanto maior o nível de detalhes, menor o nível de granularidade. Essa é uma questão fundamental no projeto de um DW, pois afeta diretamente a quantidade de dados armazenada e, ao mesmo tempo, o tipo de consulta que pode ser realizada Barquini (1996); Labio et al. (1997).

A equipe de desenvolvimento de sistemas de apoio à tomada de decisão não é muito diferente dos outros tipos de equipes de tecnologia da informação. De uma forma geral, precisa-se de gerentes, pessoal técnico, pessoal de infra-estrutura e usuários. Mais especificamente, para o desenvolvimento de sistemas de DW pode ser necessário os seguintes cargos: diretor de projeto, arquiteto de DW, administrador de banco de dados, administrador de sistema, especialista em migração de dados, especialista em sistemas legados, especialista em transformação/organização dos dados, especialista em fornecimento de dados, líder de desenvolvimento de *Data Mart*, gerente ou administrador de operações/centro de dados, gerente de configuração, consultor de empresa, consultor de gerenciamento de mudança, especialista em controle de qualidade/teste, especialista em

infra-estrutura, analista de controle de produção, usuário avançado, instrutor, redator técnico, profissional de relações públicas, administrador de metadados, patrocinadores corporativos, profissional de *help desk* (suporte), executivo empresarial de usuário final, especialista em ferramentas, pessoa de relações com o fornecedor, *webmaster*, gerente de repositório de metadados, analista de novas tecnologias, gerente de novas tecnologias, usuários finais e consultores.

A aplicação em particular determinará como esses cargos serão alocados. Em um projeto de *Data Warehouse* muito grande, esses papéis seriam alocados entre 7 a 30 indivíduos. Já em um projeto menor, uma única pessoa poderia desempenhar dois, três ou até todos esses papéis Corey et al. (2001).

O desenvolvimento de sistemas de *Data Warehousing* acarreta um certo nível de risco. Esses riscos podem ser divididos em três categorias: risco de tecnologia, risco de gerenciamento de projeto e risco comercial.

No risco de tecnologia, a equipe de desenvolvimento pode não conseguir fazer com que as tecnologias funcionem corretamente. Talvez a equipe não consiga que as ferramentas de movimentação de dados funcionem, ou que o banco de dados seja carregado, ou que forneça dados com rapidez suficiente. No risco de gerenciamento de projeto, embora consiga-se fazer com que as tecnologias funcionem, simplesmente não consegue-se apresentar o projeto a tempo ou dentro do prazo. Já o risco comercial é o mais desprezado nos projetos de DW, sendo também o mais provável de causar danos a esses projetos, pois nele, o sistema, após terminado, não é utilizado. O maior problema desse risco é que ele não é identificado até que o sistema esteja finalizado e todo o dinheiro do projeto tenha sido gasto Corey et al. (2001).

Além dos riscos apresentados, o *Data Warehouse Institute*[1] aponta os dez erros mais comuns no desenvolvimento de um DW:

1. Começar o projeto com o tipo errado de patrocínio;
2. Gerar expectativas que não podem ser satisfeitas, frustrando os usuários quando forem utilizar o DW;
3. Dizer: Isto vai ajudar os gerentes a tomarem decisão melhores e outras afirmações politicamente ingênuas;
4. Carregar o DW com informações só porque estavam disponíveis;
5. Falhar no objetivo de acrescentar valor aos dados através de mecanismos de desnormalização, categorização e navegação assistida;

[1] `www.dw-institute.com`

6. Escolher um gerente para o DW que seja voltado para a tecnologia ao invés de voltado para o usuário;
7. Focalizar o DW em dados tradicionais internos orientados a registro e ignorar o valor potencial de dados textuais, imagens, som, vídeo e dados externos;
8. Fornecer dados com definições confusas e sobrepostas;
9. Acreditar nas promessas de desempenho, capacidade e escalabilidade dos vendedores de produtos para DW;
10. Usar o DW como uma justificativa para modelagem de dados e uso de ferramentas CASE.

Na próxima subseção serão apresentadas as técnicas utilizadas para modelar os dados em sistemas de apoio à tomada de decisão.

3.1.5 Modelagem Multidimensional

A modelagem multidimensional é uma técnica utilizada para a conceitualização de modelos de negócio como um conjunto de medidas descritas por aspectos comuns Barquini (1996). Para entender melhor essa definição, é importante descrever alguns conceitos relacionados com a modelagem multidimensional, independentemente da tecnologia de banco de dados utilizada para implementá-la.

O esquema estrela é capaz de modelar as múltiplas dimensões de um repositório de dados através de tabelas de dimensão, que estão relacionadas com uma tabela central, também chamada de tabela de fatos. Os fatos são valores ou índices que podem ser medidos em um determinado processo de negócio, as dimensões são classes que descrevem as medidas numéricas, sendo que cada dimensão é descrita por um conjunto de atributos que muitas das vezes formam uma hierarquia Gatziu & Vavouras (1999); Golfarelli et al. (1998); Kimball (1997). Como exemplo, podemos tomar uma tabela de fatos como sendo as vendas realizadas por uma empresa. Nesse exemplo, as tabelas de dimensão devem armazenar informações como o tipo do produto envolvido na venda, a data em que a venda foi efetuada e o consumidor envolvido. Os atributos ano, mês, semana e dia da dimensão Data formam um hierarquia. O esquema estrela desse exemplo é mostrado na Figura 3.8.

O esquema estrela objetiva a desnormalização dos dados, para se obter um melhor desempenho, no ambiente de apoio à tomada de decisão, em relação às estruturas altamente normalizadas das bases de dados operacionais. O segredo

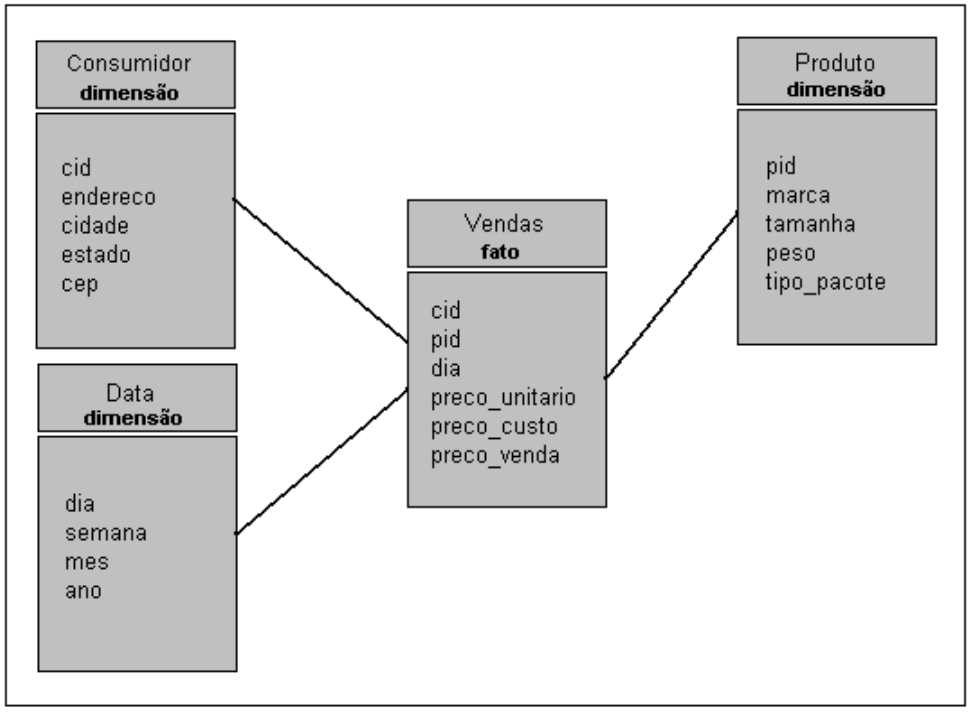


Figura 3.8: Exemplo de esquema estrela sobre vendas Marques et al. (2000).

para se obter esse desempenho é limitar o número de uniões que terão de ser realizadas e a complexidade de cada união. Esse esquema objetiva também a criação de um modelo de dados que seja mais compreensível ao usuário final, procurando representar a maneira natural de como ele enxerga o seu negócio, uma vez que, os esquemas E/R Entidades/Relacionamentos) são de difícil interpretação por parte dos usuários finais, além de não representar a maneira natural de como eles visualizam seu negócio Kimball (1997); Todman (2001).

Uma variação do esquema estrela, chamado de esquema floco de neve, é utilizado para representar as hierarquias das dimensões através da normalização das tabelas de dimensão. No esquema floco de neve, as tabelas de dimensão podem se tornar tabela de fatos de outras tabelas obtidas. A vantagem desse tipo de esquema é que torna-se mais fácil a manutenção das tabelas de dimensão, já que há uma diminuição na redundância dos dados. Entretanto, uma estrutura não normalizada é mais eficiente no momento de execução das consultas, um requisito indispensável em sistemas de DWs Gatziu & Vavouras (1999); Golfarelli et al. (1998); Kimball (1997).

Em grandes projetos, o DW normalmente contém entre 10 e 25 esquemas estrela, sendo cada um formado por 4 a 12 dimensões. Muitas dessas dimensões

poderão ser compartilhadas por cada esquema estrela. Quando o esquema de um DW é composto por mais de uma estrutura do tipo estrela, este pode ser chamado de esquema constelação Barquini (1996).

As tabelas de dimensão são caracterizadas por vários aspectos gerais. Normalmente elas são altamente desnormalizadas. Embora, freqüentemente fala-se que o esquema estrela é desnormalizado, na verdade somente as tabelas de dimensão são desnormalizadas. As tabelas de dimensão possuem mais colunas do que as tabelas do banco de dados operacional e geralmente possuem menos registros do que as tabelas de fatos.

Em certas situações é necessário utilizar uma chave substituta para a dimensão, pois ela permite a captura do histórico da dimensão e em outros casos fornecem um desempenho de união melhor do que as chaves operacionais. As dimensões também possuem referências aos registros correspondentes nas tabelas de origem, além de campos de data adicionais e flags indicando se um determinado registro da dimensão está ativo ou não Corey et al. (2001).

Há uma outra maneira de considerar a multidimensionalidade de um repositório, onde as múltiplas dimensões do *Data Warehouse* são representadas por meio de cubos de dados. Cada eixo do cubo corresponde a uma dimensão. Considerando o esquema estrela apresentado anteriormente, as dimensões produto, data e consumidor vão constituir os eixos do cubo de dados. Os pontos de interseção entre todas as dimensões são chamados células e representam uma visão do cubo. A maior parte das visões podem ser computadas em função de outras. Diz-se que tais visões são dependentes. A Figura 3.9 ilustra um cubo de dados.

O conceito de banco de dados multidimensional (BDM) é bem mais simples do que o de banco de dados relacional. Ao invés de armazenar informações como registros em tabelas, BDMs armazenam os dados em *arrays* ou matrizes. Existe no mercado uma classe de SGBDs (Sistema Gerenciador de Banco de Dados) que incorporam a tecnologia de banco de dados multidimensional, são os chamados Sistemas Gerenciadores de Banco de Dados Multidimensionais (SGBDMs). O grande problema desse tipo de SGBD é a sua capacidade de armazenamento ainda limitada para as necessidades de um DW. Dessa forma, esses produtos são mais utilizados como gerenciadores de *Data Marts*, trabalhando com apenas um subconjunto de dados do DW Bauer & Lehner (1997); Colliat (1996).

Projetistas de bancos de dados podem e devem separar o conceito de visão multidimensional dos dados, obtida através da modelagem multidimensional, do conceito de armazenar os dados de forma multidimensional. Os resultados da

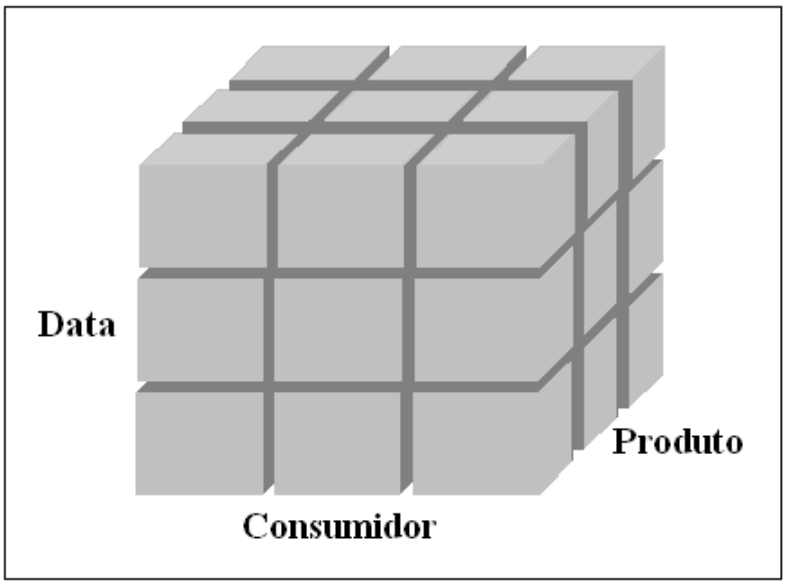


Figura 3.9: Um exemplo de cubo de dados sobre vendas.

modelagem multidimensional podem ser implementados diretamente utilizando a tecnologia de banco de dados multidimensional ou através do esquema estrela, em um banco de dados relacional. A falta de um modelo de dados multidimensional convencional, tal como o esquema estrela, para bancos de dados relacionais e a falta de um método de acesso padrão, tal como o SQL, acabaram influenciando a utilização da tecnologia de bancos de dados relacional para representar e armazenar dados multidimensionais Kimball (1997).

3.1.6 OLAP

A característica principal dos sistemas OLAP (*On-line Analytical Processing*) é permitir uma visão conceitual multidimensional dos dados armazenados. A visão multidimensional é mais útil para os analistas do que a visão tabular tradicional utilizada nos sistemas de processamento de transação. Ela é mais natural, fácil e intuitiva, permitindo uma visão dos negócios da organização em diferentes perspectivas e, dessa maneira, torna o analista um explorador de informações Wu & Buchmann (1997); Shoshani (1997); Campos & Rocha (1997).

As ferramentas OLAP são projetadas para apoiar análises e consultas *ad hoc* em um *Data Warehouse*, além de ajudarem analistas e executivos a sintetizarem informações sobre a organização, através de comparações, visões personalizadas, análise histórica e projeção de dados em vários cenários. Ferramentas OLAP são implementadas para ambientes multi-usuário, arquitetura cliente-

servidor e oferecem respostas rápidas e consistentes às consultas interativas executadas pelos analistas, independente do tamanho e complexidade do DW Codd (1993); Chaudhuri & Dayal (1997); Inmon et al. (1999).

Originalmente, o OLAP era um conceito simples, utilizado para descrever toda a análise realizada em dados agregados. Porém, surgiram novas variações sobre o tema, como ROLAP (OLAP Relacional), MOLAP (OLAP Multidimensional), HOLAP (OLAP Híbrido), DOLAP (OLAP de *Desktop*) e WOLAP (OLAP para Web). Nessas arquiteturas de OLAP, a interface para a camada analítica normalmente é a mesma, o que diferencia algumas arquiteturas é o modo como os dados são fisicamente armazenados.

No ROLAP os dados são armazenados em tabelas de um SGBD Relacional, em uma forma desnormalizada, sendo os mesmos modelados com o esquema estrela. No MOLAP os dados são armazenados nos *arrays* de dados de um SGBD Multidimensional, sendo que não existe um modelo conceitual para representar os dados para esse tipo de SGBD. O HOLAP é um híbrido entre ROLAP e MOLAP. O DOLAP, por outro lado, é uma variação que existe para portabilidade. Ele cria conjuntos de dados multidimensionais que podem ser transferidos do servidor para o desktop. Isso proporciona certas vantagens para os usuários de computador portátil, como os vendedores que estão sempre na rua e não têm acesso ao dados em seus escritórios. Já o WOLAP, é OLAP voltado para Internet Corey et al. (2001).

Uma decisão de projeto importante a ser tomada é sobre a estratégia OLAP a ser adotada, ou seja, decidir se deverá ser utilizado ROLAP ou MOLAP. Cada uma tem suas vantagens e desvantagens. Os bancos de dados MOLAP têm um limite quanto ao tamanho físico do conjunto de dados que pode ser manipulado. Por exemplo, o banco de dados multidimensional Oracle Express pode, teoricamente, manipular o equivalente a 2^{63} células. Porém, as restrições de armazenamento e desempenho limitarão o tamanho do banco de dados Express bem antes que a capacidade física seja alcançada. Também existe um limite para a quantidade de dimensões que ele pode manipular e ainda oferecer um desempenho razoável. A estratégia MOLAP é ideal em situações nas quais os dados podem ser divididos em partes menores. Quanto menores os conjuntos, mais rápidos serão os tempos de compilação. Já o ROLAP possui a vantagem de poder ser executado em grandes conjuntos de dados.

Outras características que também devem ser levadas em conta durante a escolha das estratégias MOLAP ou ROLAP, são o desempenho de consulta, de-

sempenho de carregamento, capacidade analítica, tamanho dos conjuntos de dados, tratamento de dimensão e esforço de manutenção Corey et al. (2001). Essas características são descritas a seguir:

Desempenho de consulta : os sistemas ROLAP respondem as consultas exatamente como qualquer outro aplicativo de banco de dados relacional. Às vezes, as respostas voltam rapidamente e, às vezes, demoram. O administrador pode trabalhar para melhorar o tempo de resposta, construindo tabelas de resumo e índices. Por outro lado, a estratégia MOLAP fornece uma resposta bastante previsível e rápida para praticamente qualquer consulta. Em parte, isso se deve ao fato de que os bancos de dados multidimensionais calculam previamente muitos, e às vezes todos, os valores possíveis em seus hipercubos.

Desempenho de carregamento : A maioria dos bancos de dados multidimensionais não são atualizados diariamente. Na verdade, o ciclo de atualização mais comum é o mensal. Infelizmente, um dos custos do desempenho que se pode obter de um banco de dados multidimensional são longos tempos de carregamento. Por outro lado, os bancos de dados relacionais, freqüentemente podem ser carregados mais rapidamente. Existem várias etapas nesse processo, incluindo o próprio carregamento, indexação e construção de tabelas de resumo. Além disso, é comum que os *Data Warehouses* e *Data Marts* relacionais seja atualizados diariamente.

Capacidade analítica : os bancos de dados MOLAP tendem a ter um suporte melhor para análises de série temporais e estatísticas. Os bancos de dados ROLAP, por outro lado, às vezes são atrapalhados pelas limitações da SQL.

Tamanho do conjunto de dados : os bancos de dados multidimensionais tendem a crescer muito rapidamente, particularmente quando mais dimensões são modeladas nele. Outro motivo pelo qual os bancos de dados multidimensionais podem ficar muito grandes deve-se ao grande número de valores de resumo previamente calculados que eles possuem. Reunindo tudo, esses bancos de dados podem ficar muito grandes em pouco tempo. Por outro lado, existem limitações físicas para o quanto tais bancos de dados podem crescer. Já os bancos de dados relacionais oferecem suporte para um crescimento praticamente ilimitado.

Tratamento de dimensão : Os bancos de dados ROLAP normalmente são construídos como esquemas estrela. As tabelas de dimensão em um esquema estrela podem ser bastante grandes. Já os bancos de dados MOLAP, não fornecem tal flexibilidade com as dimensões. Esses sistemas são limitados pela quantidade de diferentes níveis de dimensão que podem conter. Essa limitação está relacionada ao problema das explosões do tamanho do banco de dados, quando dimensões são incluídas.

Esforço de manutenção : A estratégia MOLAP é muito eficiente na manutenção. Uma vez estabelecida, ela é mantida de forma bastante automática. Já a estratégia ROLAP exige mais trabalho para o preenchimento e manutenção. O preenchimento é mais complexo porque não apenas uma, mas várias estruturas devem ser preenchidas. Além disso, índices e restrições talvez precisem ser ativados ou desativados durante esse processo. Uma vez carregado o banco de dados, se o desempenho for ruim, mais índices talvez precisem ser incluídos ou novas tabelas de resumo criadas. O administrador deve analisar regularmente o banco de dados para mantê-lo em perfeito funcionamento.

Independente de usar ROLAP ou MOLAP, no caso das soluções da Oracle, haverá o limite de 32 dimensões (o mesmo que as 32 colunas-chave no pensamento relacional tradicional). Esse limite nunca é um problema, pois a maior parte dos aplicativos OLAP usa de 5 a 7 dimensões e quase todos usam menos de 10 a 12 dimensões Corey et al. (2001).

A escolha da estratégia ROLAP ou MOLAP depende de vários fatores, mas principalmente da abrangência da aplicação. Se o DW a ser construído é grande e serão consideradas várias funções, então, provavelmente, uma banco de dados relacional deve ser utilizado. Caso contrário, se o que estiver sendo construído for um *Data Mart* bem definido, altamente centrado em análise, com dimensionalidade limitada e pouca necessidade de dados detalhados de nível atômico, então a estratégia multidimensional é a ideal.

No entanto, o mais importante é desenvolver um ambiente de DW consistente, antes de escolher uma estratégia OLAP. Essa é uma questão bem mais importante a ser resolvida do que a escolha de ROLAP em detrimento de MOLAP.

A fim de permitir uma visualização e manipulação multidimensional dos dados, as ferramentas OLAP oferecem diferentes funções Codd (1993); Inmon & Hackarthorn (1997):

Pivot: muda a orientação dimensional de uma pesquisa. Por exemplo, *pivot* pode consistir na troca de linhas e colunas, ou mover uma das dimensões da linha, para a dimensão da coluna, como ilustrado na Figura 3.10.

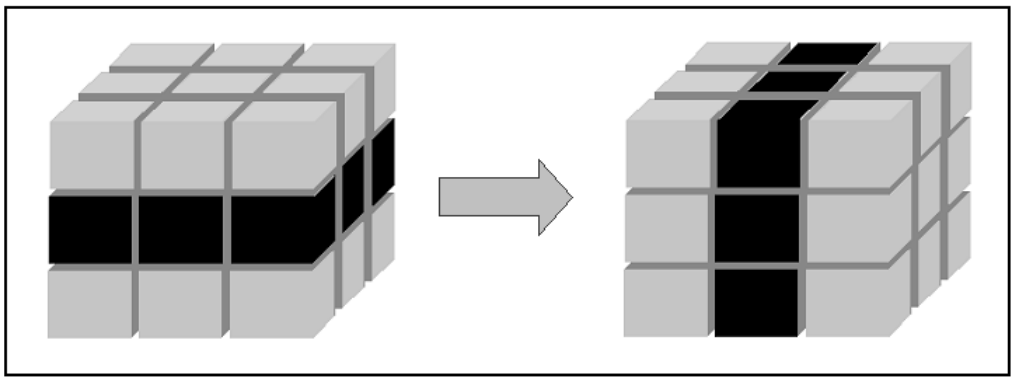


Figura 3.10: Operação de *Pivot*.

Roll-up: as bases de dados multidimensionais geralmente têm hierarquias ou relações de dados baseadas em fórmula dentro de cada dimensão. Então, a execução do *roll-up* computa todas essas relações para uma ou mais dimensões.

Slice: um *slice* é um subconjunto da estrutura multidimensional que corresponde a um valor simples em lugar de um ou mais atributos das dimensões. É como fixar um valor de uma das dimensões de um cubo e considerar para pesquisa o subcubo formado por esse valor e pelas outras dimensões do cubo inicial, como mostrado na Figura 3.11.

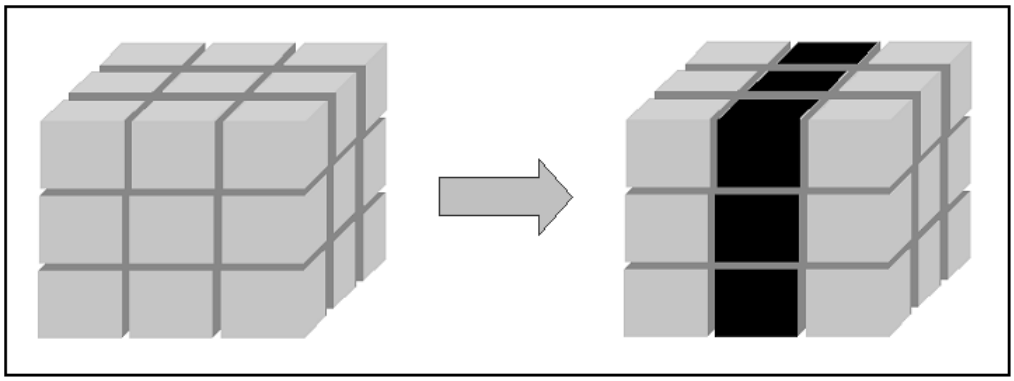


Figura 3.11: Operação de *Slice*.

Drill-down/up: consiste em fazer uma exploração em diferentes níveis de detalhe das informações, como por exemplo, analisar uma informação por

continente, país ou estado, partindo da mesma base de dados. Essa função é ilustrada na Figura 3.12.

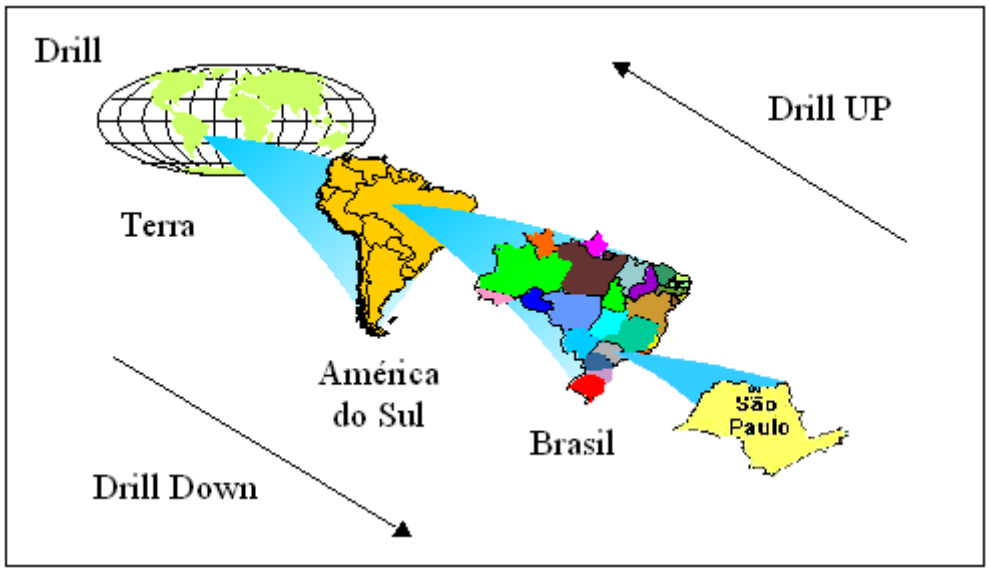


Figura 3.12: Operação de *Drill-down/up*.

Drill-across: é o processo de unir duas ou mais tabelas fatos de mesmo nível de detalhes, ou seja, tabelas com o mesmo conjunto de colunas e restrições dimensionais.

Essas funções podem ser utilizadas a vontade pelos usuários de um ambiente de *Data Warehouse*, conforme as suas necessidades de informações. Após a apresentação dos diversos conceitos relacionados a OLAP, o foco da próxima subseção é o povoamento do DW.

3.1.7 Povoamento

A primeira etapa a ser realizada no povoamento de um *Data Warehouse* consiste na extração dos dados das fontes operacionais. Os dados são então copiados para uma área de trabalho temporária, onde recebem algum tratamento especial. Essa área de trabalho temporária corresponde à área de adaptação apresentada na Subseção 3.1.2.

A forma como essas informações são extraídas varia de acordo com os recursos oferecidos pela fonte de dados. Durante o processo de extração dos dados, deve-se isolar os dados que foram inseridos e atualizados desde a última extração, processo este conhecido como captura das mudanças ocorridas nas fontes.

As regras de captura dos dados devem ser impostas pelo administrador do DW, dependendo das necessidades dos usuários, tráfego na rede e período de menor sobrecarga, tanto das fontes de dados quanto do DW. Essas regras podem variar para cada tipo de origem Chaudhuri & Dayal (1997); Williams (1997).

Existem algumas técnicas que podem ser utilizadas para identificar alterações nas tabelas do banco de dados de origem. A primeira técnica corresponde a identificação de indicadores de tempo nos registros. Se a fonte de dados possuir informações de tempo indicando a data em que um registro foi inserido, atualizado ou removido, o processo de extração dos dados é mais facilmente executado. Uma outra técnica interessante para capturar alterações em registros das fontes de dados é a utilização de *triggers* sobre as tabelas de origem. Com esta técnica, sempre que um registro for inserido, atualizado ou excluído, os *triggers* gravam uma mensagem em um arquivo de log. Estas mensagens serão utilizadas posteriormente para atualizar o DW. Uma outra técnica consiste na aplicação de ferramentas AIS (*Application Integration Software*), sendo as mesmas utilizadas para passar informações entre aplicativos. Uma última técnica envolve a identificação de alterações nos dados de origem comparando o seu estado atual com o estado no qual estavam quando foi realizada a carga no DW Corey et al. (2001).

Uma vez que os dados já se encontram na área temporária de trabalho, eles devem passar por uma fase de limpeza ou filtragem, onde o objetivo é garantir a integridade dos dados através de programas ou rotinas especiais que tentam identificar anomalias e resolvê-las, deixando os dados consistentes antes de serem carregados no *Data Warehouse*. A correção de erros de digitação, a descoberta de violações de integridade, a substituição de caracteres desconhecidos e a padronização de abreviações, podem ser exemplos de limpeza de dados Bohn (1997). Um outro tratamento que deve ser aplicado aos dados são os registros duplicados, representados com diferentes identificadores. Estes registros duplicados devem ser integrados em um mesmo identificador antes de serem carregados no *Data Warehouse*. Após os dados estarem limpos e transformados, os mesmos já podem ser carregados no DW Corey et al. (2001).

O povoamento de um DW não ocorre uma única vez. Após a carga inicial outras cargas poderão ser necessárias para manter os dados armazenados no DW consistentes com os dados contidos nas fontes externas e nos bancos de dados de produção. A definição do intervalo de tempo entre uma carga e outra deve ser baseada na estabilidade dos dados das fontes. Se os dados estão sofrendo

alterações constantemente, as cargas do DW deverão ser constantes também, porém, se os dados sofrem alterações raramente, essa carga pode ser realizada em um intervalo de tempo maior.

O empecilho que há na alimentação dos dados das fontes para o DW não é técnico, mas gerencial. Muitos dos processos envolvidos, como mapeamento, integração e avaliação de qualidade, ocorrem durante a fase de análise, projeto e implementação Moriarty & Greenwood (1996). Especialistas afirmam que identificar fontes, definir regras de transformação e detectar e resolver questões de qualidade e integridade, consomem cerca de 80% do tempo de projeto. Infelizmente, não é nada fácil automatizar essas tarefas Miley (1997). Após o povoamento do DW, já é possível analisar os dados contidos no mesmo com consultas OLAP e extrair conhecimento com ferramentas de *Data Mining*, sendo esse o foco da próxima subseção.

3.1.8 Apoio a Extração de Conhecimento

Mesmo sabendo que a informação e/ou conhecimento relacionado ao produto de sucesso de uma organização esteja de alguma forma entre o grande volume de dados armazenados, pode existir ainda um grande caminho a ser percorrido até que esta informação e/ou conhecimento esteja de fato disponível. Sua extração eficaz, de forma que possa dar apoio ao processo de tomada de decisão, depende da existência de ferramentas especializadas que permitam tanto a captura dos dados relevantes de uma forma mais eficaz quanto a visualização dos mesmos Inmon & Hackarthorn (1997). Vale ressaltar que o termo extração, neste contexto, não deve ser confundido com a extração dos dados das fontes para o povoamento do *Data Warehouse*.

As ferramentas não devem apenas permitir o acesso aos dados, mas também permitir análises de dados significativas, de tal forma que possa transformar dados brutos em informação e que possa também dar suporte a extração de conhecimento para os processos estratégicos de uma organização. O sucesso na implantação de um *Data Warehouse* pode depender da disponibilidade da ferramenta adequada para as necessidades de seus usuários.

Dois tipos de ferramentas são muito utilizadas para analisar os dados de um *Data Warehouse*. As ferramentas OLAP servem para extrair informações e as ferramentas de *Data Mining* para extrair conhecimento. A diferença básica entre essas duas ferramentas está na maneira como a exploração dos dados é abordada. Com ferramentas OLAP a exploração é feita na base da verificação, isto é,

o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta para refutá-la ou confirmá-la. Com *Data Mining*, a questão é total ou parcialmente desconhecida e a ferramenta é utilizada para a busca do conhecimento.

Em *Data Mining*, fala-se freqüentemente em encontrar padrões, regras e fatos nos dados armazenados. Mas o que são padrões, regras e fatos? Em uma tabela contendo dados, um padrão é definido como um conjunto de linhas que compartilham os mesmos valores com duas ou mais colunas. Um fato é representado por um padrão com fator de confiança superior a 50%. A partir deste fato pode ser deduzida uma regra, por exemplo, *se item = carro então cor = vermelho* Azmy (1998). A próxima seção aborda o processo de *Data Mining* com mais detalhes.

3.2 Data Mining

Um dos principais avanços ocorridos na área de tecnologia da informação e que vem se consolidando cada vez mais, é o processo de *Data Mining*. Esse é composto por técnicas e ferramentas capazes de automatizar o processo de análise e compreensão dos dados, sendo seu objetivo principal encontrar padrões válidos e úteis nos mesmos Fayyad et al. (1996). Convém ressaltar, que no contexto deste trabalho, os termos *Data Mining* e Extração de Conhecimento de Bases de Dados[2] serão utilizados indistintamente.

O processo de *Data Mining* se apresenta como um conjunto de técnicas e ferramentas que contribuem de forma significativa para o problema de aquisição de conhecimento. Esse processo é responsável pela busca de padrões interessantes, obtendo vantagens como o refinamento de tipos específicos de conhecimento, redução de custos e de recursos humanos envolvidos Han (1995); Fayyad et al. (1996).

A Extração de Conhecimento a partir de Bases de Dados é um processo interativo e iterativo Mannila (1997b). Sua interatividade está relacionada à compreensão, por parte dos usuários desse processo, sobre o domínio da aplicação. Para uma melhor definição das funções dos usuários que utilizam o processo de *Data Mining*, eles são divididos em três classes Rocha (1999): o Especialista do Domínio, que possui amplo entendimento do domínio da aplicação; o Analista, que executa o processo de *Data Mining* e o Usuário Final, que utiliza o conhecimento extraído para auxiliá-lo em sua tomada de decisão.

Esses usuários podem até não possuir funções separadas no processo de *Data Mining*. Dessa forma, pode haver situações, em que o especialista também é o

[2] *Knowledge Discovery in Databases - KDD*

usuário final ou que o especialista auxilie ou execute funções que pertencem ao analista. O êxito do processo depende da interação entre esses usuários Fayyad et al. (1996); Fayyad & Simoudis (1997).

O processo é dividido em três fases iterativas: pré-processamento, extração de padrões e pós-processamento, sendo que antes dessas, deve ocorrer a fase de conhecimento do domínio e após, a fase de utilização do conhecimento, conforme ilustrado na Figura 3.13.

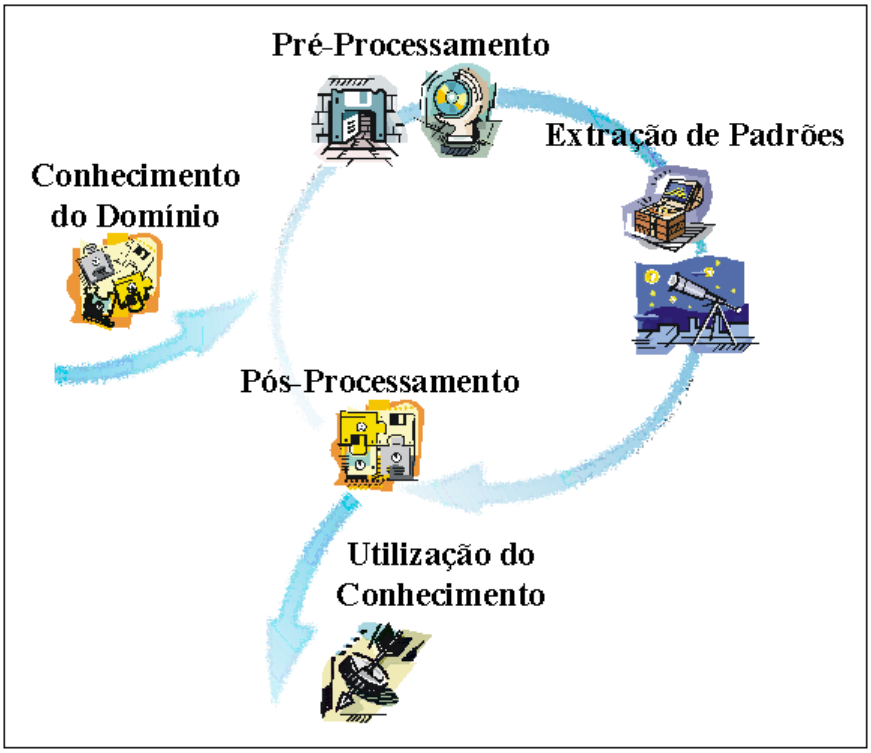


Figura 3.13: Fases do processo de *Data Mining*.

De uma forma geral, o processo inicia-se com a definição e compreensão do domínio da aplicação, na qual o analista toma conhecimento do domínio da aplicação, considerando aspectos tais como os objetivos dessa aplicação e as fontes de dados. Logo após, dá-se lugar à fase de pré-processamento, que consiste na realização de uma seleção de dados a partir das fontes, de acordo com os objetivos da aplicação do processo. Os conjuntos de dados resultantes dessa seleção precisam passar por um processo de limpeza e preparação, para então, serem submetidos à fase de extração de padrões, onde serão utilizados métodos e ferramentas para encontrar relacionamentos ocultos nos dados. Na fase de pós-processamento, os padrões encontrados são avaliados quanto à sua qualidade e utilidade para que, em caso positivo, sejam utilizados para apoiar algum

processo de tomada de decisão.

A Figura 3.14 ilustra o tempo normalmente necessário para a execução de cada uma das fases do processo de *Data Mining*. Como se pode ver, 80% de todo o processo geralmente é ocupado pela fase de pré-processamento Mannila (1997b). A fase de extração de padrões gasta 10% do tempo e a fase de pós-processamento ocupa os 10% restantes do tempo total Cabena et al. (1998).

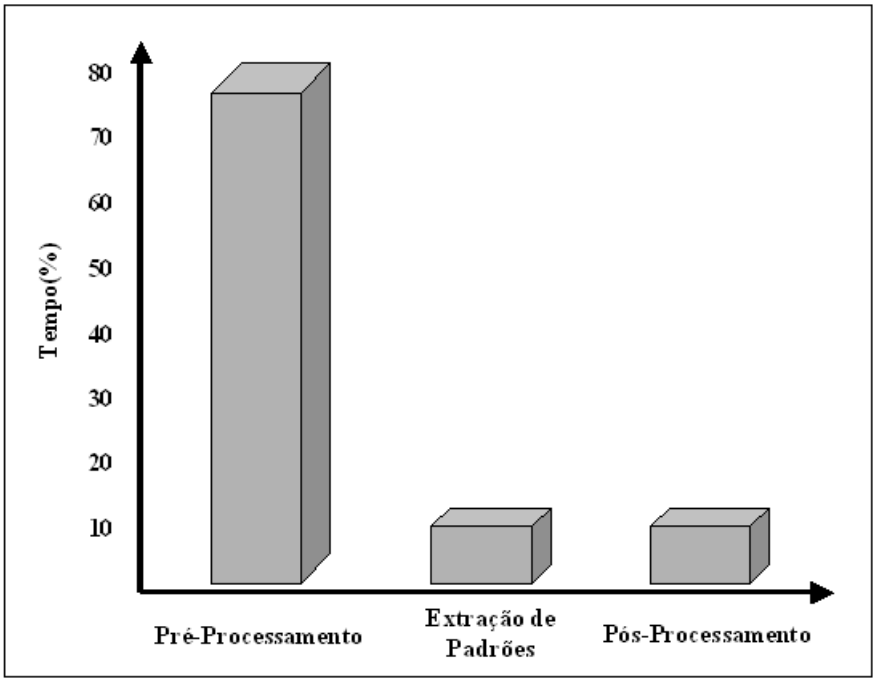


Figura 3.14: Tempo gasto por cada fase no processo de *Data Mining*.

Deve ser observado que as ferramentas de visualização são muito importantes no processo, pois, elas são de grande relevância para facilitar o entendimento e avaliação, principalmente pelo usuário final, dos resultados de cada fase Rezende et al. (1998).

Vale ressaltar que, por ser um processo iterativo, a ocorrência de mudanças em qualquer uma das fases afetará o sucesso de todo o processo. Dessa forma, os resultados de uma determinada fase podem acarretar o recomeço de todo o processo Fayyad et al. (1996); Netz et al. (2000).

Por outro lado, pelo fato de existir uma grande variedade de problemas e diferentes tipos de bases de dados, é praticamente impossível definir uma metodologia única para a extração de conhecimento. Dependendo do problema, existe em algumas ocasiões a necessidade de criar sub-tarefas dentro de cada tarefa Adriaans & Zantinge (1996). A seguir, será apresentado um detalhamento maior

da funcionalidade de cada fase.

Conhecimento do Domínio

A definição dos objetivos a serem atingidos são um dos aspectos fundamentais a serem considerados para o sucesso do processo de *Data Mining*. Essa definição é geralmente realizada pelo especialista do domínio com o apoio do analista do processo, com a finalidade de definir o domínio da aplicação, o conhecimento prévio relevante, a viabilidade e custos da aplicação, duração do projeto, resultados esperados, entre outros. Uma análise cuidadosa do problema é necessária nesta tarefa, para que uma melhor compreensão do domínio seja alcançada Félix (1998).

Pré-processamento

Extrair conhecimento de grandes bases de dados pode se tornar uma tarefa inviável, principalmente pela limitação do número de registros que os algoritmos de extração de padrões podem manipular. Grandes volumes de dados podem gerar um espaço de busca de padrões combinatoriamente imenso. Além disso, a busca de conhecimento em grandes bases de dados pode ocasionar, ainda, o aumento das possibilidades de se encontrar padrões pouco significativos. Possíveis soluções para esse problema envolvem a tentativa de selecionar amostras significativas ou selecionar conjuntos de dados a partir de interações com o especialista do domínio.

A escolha de uma amostra que reflita com a maior fidelidade possível os objetivos da aplicação do processo, além da própria base de dados, é de suma importância para as demais fases do processo de *Data Mining*. Seleções de amostras pouco significativas podem produzir resultados imprecisos ou sem valor. Além disso, pequenos conjuntos podem levar a conclusões incorretas.

Após a tarefa de seleção e amostragem, é necessário limpar e preparar os conjuntos de dados, a fim de atender às exigências e limitações dos formatos de entrada dos algoritmos para extração de padrões. Nessa tarefa devem ser observados alguns fatores, como os que se seguem: verificação das características da base de dados, como tipos de dados e padronização do conteúdo dos registros; eliminação dos registros duplicados e lixo nos dados produzidos pelas migrações; tratamento de ruídos nos dados; avaliação do grau de representatividade dos atributos; manipulação de valores de atributos ausentes e representação dos dados de acordo com os objetivos da tarefa Batista (2000).

A fase de pré-processamento deve ser executada, sempre que possível, com o

acompanhamento do especialista do domínio, pois este possui um conhecimento mais profundo do domínio em questão. Vale lembrar que, apesar de existirem algumas técnicas de pré-processamento de dados, a participação de um especialista do domínio é de fundamental importância.

É importante destacar que se a base de dados estiver em um *Data Warehouse*, problemas como padronização e limpeza nos dados podem ser em grande parte resolvidos, pois o *Data Warehouse* provê métodos que, entre outros, permitem a integração, padronização e sumarização de dados Inmon (1996); Kimball (1997). Com os dados selecionados, amostrados, limpos e preparados pode-se dar início à fase de extração de padrões.

Extração de Padrões

A fase de extração de padrões está relacionada à aplicação de algoritmos que, mediante limitações de eficiência computacional aceitáveis, são capazes de produzir uma relação particular de padrões a partir de grandes massas de dados Fayyad et al. (1996); Han (1999); Weiss & Indurkhya (1998). Por ser considerada uma das fases cruciais e a mais complexa do processo de *Data Mining*, a fase de extração de padrões, por sua vez, pode ser dividida em várias tarefas: escolha da atividade, escolha do algoritmo e extração de padrões propriamente dita, conforme ilustrado na Figura 3.15.

A seguir será apresentada uma descrição sucinta de cada tarefa:

Escolha da atividade: a escolha de uma atividade na fase de extração de padrões é realizada conforme algum objetivo especificado entre o analista e o especialista. Essas atividades estão divididas em dois grupos: a predição e a descrição. As atividades do grupo de predição envolvem o uso de atributos de um conjunto de dados para prever um valor futuro de um outro atributo. Esse grupo é composto pelas atividades de classificação e regressão Brand & Gerritsen (1998). As atividades do grupo de descrição procuram padrões interpretáveis pelos humanos e que descrevam os dados, sendo que esses dados não possuem um atributo-meta especificado, como acontece no grupo de atividades de predição. Esse grupo é composto pelas atividades de regras de associação, *clustering*, sumarização, caracterização, discriminação, evolução, desvio, modelo de dependência, análise de links e análises seqüenciais. Uma vez definida a atividade, deve-se escolher o algoritmo a ser aplicado.

Escolha do algoritmo: decidir sobre qual o melhor algoritmo a ser utilizado não

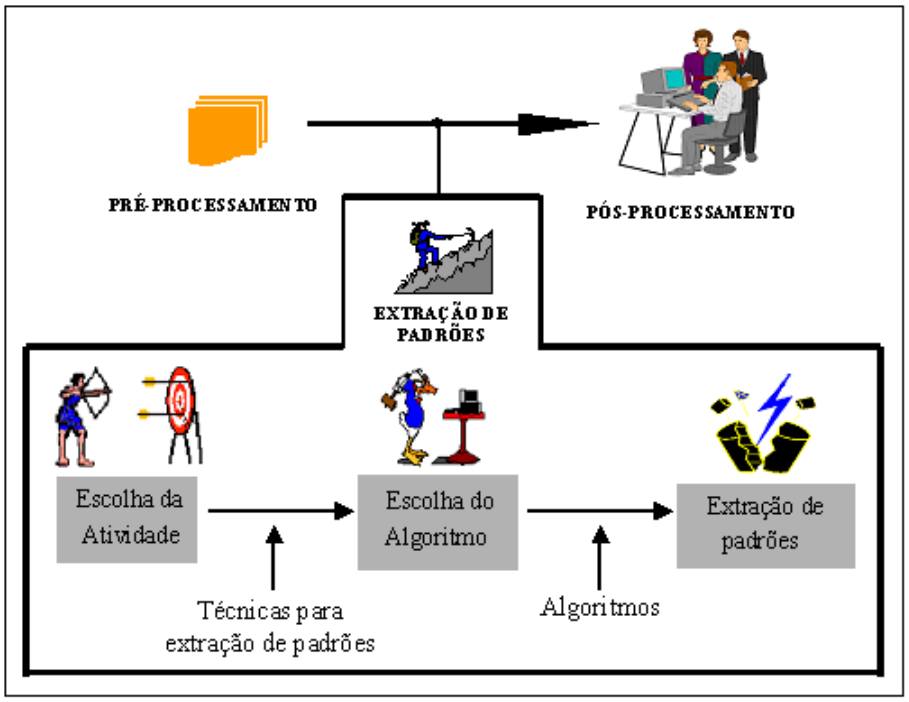


Figura 3.15: Tarefas da fase de extração de padrões.

é uma tarefa trivial, pois sabe-se que nenhum deles possui um bom desempenho em todos os domínios de aplicação Salzberg (1997); Batista (1997). A escolha do algoritmo é realizada pelo analista, sendo que o mesmo deve levar em consideração as restrições do domínio da aplicação e as preferências do usuário final. Com base nessas restrições, o analista pode escolher o algoritmo tomando como base alguns parâmetros, tais como: tipo de aprendizado, paradigmas de aprendizado, linguagens de descrição e como novos exemplos serão integrados Monard et al. (1997). Convém ressaltar que além da observação desses parâmetros, os testes experimentais possuem um papel importante na escolha do algoritmo, uma vez que não existem metodologias para a escolha do melhor algoritmo a ser utilizado em um dado domínio de aplicação Kliber et al. (1988).

Extração de padrões: uma vez escolhidos a atividade e o algoritmo, pode-se dar início à extração dos padrões. Essa tarefa a principal responsável em encontrar o conhecimento embutido no conjunto de dados e envolve a utilização de modelos adequados para a representação dos padrões encontrados. Os modelos resultantes, depois de avaliados, são utilizados para predizer os valores dos atributos definidos pelo usuário final Kerber et al. (1995);

Fayyad (1996). Os modelos gerados geralmente seguem os paradigmas de aprendizado estatístico Elder & Pregibon (1996); Glymour (1997); Padilha (1999), conexionista Haykin (1994); Braga et al. (2000) ou simbólico Mitchell (1997); Baranauskas & Monard (2000). Outra forma de se encontrar padrões nos dados é por meio de técnicas de visualização e exploração interativa. Vale ressaltar que nesse caso os padrões são identificados pelos especialistas do domínio visualmente.

Após a extração dos padrões, passa-se à fase de pós-processamento, onde o conhecimento adquirido deve ser avaliado pelo analista, com o apoio do especialista do domínio, quanto à sua importância e utilidade, para assim serem incorporados ao sistema da organização e serem utilizados em seus processos estratégicos.

Pós-Processamento

A descoberta de padrões nos dados de entrada não significa que o processo de *Data Mining* tenha sido finalizado, é necessário ainda que o usuário entenda e possa julgar a veracidade do conhecimento extraído comparando-o com o conhecimento do especialista do domínio.

Extraídos os padrões dos dados, torna-se necessária uma avaliação do conhecimento obtido e, para isso, são utilizados, entre outros, os critérios de precisão, compreensibilidade e interessabilidade. Esses critérios auxiliam na análise dos padrões encontrados, podendo ajudar também na filtragem do que foi aprendido e remoção dos padrões redundantes e irrelevantes Fayyad et al. (1996); Padmanabhan & Tuzhilin (1999); Horst (1999); Pugliesi (2001).

Utilização do Conhecimento

Após a validação do conhecimento, o mesmo deve ser consolidado, em outras palavras, deve ser incorporado a um sistema de aplicação. Após a consolidação, conflitos entre conhecimentos extraídos anteriormente poderão ser resolvidos. Além disso, todos os usuários que fazem parte do processo de tomada de decisão deverão receber todas as informações sobre as novas descobertas.

É importante destacar que as bases de dados são dinâmicas, ou seja, sofrem mudanças ao longo do tempo, conseqüentemente, o conhecimento embutido nessas bases também pode mudar. Portanto, o processo de *Data Mining* deve ser executado periodicamente.

A maioria das fases do processo envolvem a utilização de ferramentas de su-

porte à execução das tarefas pertinentes a essas fases.

3.3 Elementos de Apoio à Análise de Dados

Nesta seção serão abordados os principais elementos de apoio ao processo de análise de dados: Banco de Dados, *Data Warehouse*, Ferramentas de Visualização, Técnicas Estatísticas e Aprendizado de Máquina. Esses elementos auxiliam a análise de dados devido à otimização dos recursos e do tempo gastos nesse processo, bem como, o maior controle dos dados, no que se refere ao armazenamento e recuperação. Além disso, a compreensão do domínio é acentuadamente facilitada, uma vez que as técnicas estatísticas, em conjunto com as ferramentas de visualização, desempenham um papel fundamental em todo o processo. A seguir são apresentados esses elementos:

Banco de Dados: os dados a serem analisados deverão estar armazenados em algum Sistema de Gerenciamento de Banco de Dados (SGBD), sendo então eles a fonte dos dados. Porém, a importância do SGBD vai muito além da citada, ele auxilia também na tarefa de seleção e preparação dos dados, uma vez que nessa tarefa são selecionadas e criadas visões da base de dados em função da base operacional, determina-se o tamanho da amostra a ser utilizada e seleciona-se um conjunto de dados para ser utilizado pelo sistema.

***Data Warehouse*:** para se realizar o processo de *Data Mining* não é necessário que se tenha implementado um *Data Warehouse*, porém, se uma organização qualquer resolver realizar um processo de extração de conhecimento em um determinado domínio de aplicação e ela possuir um *Data Warehouse*, grande parte do tempo que deveria ser utilizado durante a fase de pré-processamento dos dados seria reduzido consideravelmente, permitindo, dessa forma, que o analista foque sua atenção nas outras etapas do processo.

Ferramentas de Visualização: as ferramentas de visualização de dados estão se tornando cada vez mais importantes no processo, pois permitem, entre outros aspectos, o aumento da capacidade de análise e de interpretação dos resultados obtidos Rezende et al. (1998). As técnicas de visualização de dados podem ser aplicadas como uma fase inicial e exploratória de *Data Mining* na identificação de domínios de interesses. Além disso, durante

todas as fases do processo de extração de conhecimento, pode-se utilizar essas técnicas para melhorar a comunicação entre os usuários envolvidos com o processo.

Técnicas Estatísticas: as técnicas estatísticas auxiliam diversas etapas do processo de *Data Mining*, como na seleção e amostragem dos dados, e na limpeza de ruídos. Ainda, os métodos estatísticos vêm sendo aplicados para extrair padrões dos dados, fazer reconhecimento de padrões e modelagem de dados, os quais são de grande valor para o processo Glymour (1997); Padilha (1999).

Aprendizado de Máquina: o Aprendizado de Máquina representa uma das partes centrais do processo de *Data Mining*, sendo muito utilizado na tarefa de extração de padrões. Pode ser definido como uma sub-área de Inteligência Artificial Rich & Knight (1993); Russell & Norvig (1995) que pesquisa métodos computacionais relacionados à aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente. Esses métodos podem ser entendidos como sistemas de aprendizado que tomam decisões baseadas em experiências acumuladas contidas em casos resolvidos com sucesso Carbonel & Langley (1987); Weiss & Kulikowski (1991); Mitchell (1997); Batista (1997).

3.4 Alguns Problemas Relacionados à Análise de Dados

Nesta seção são abordados alguns dos principais problemas relacionados ao processo de análise de dados e algumas das principais medidas que podem ser adotadas para evitá-los Mannila (1997a):

Definição dos Objetivos da Aplicação: a definição dos objetivos a serem atingidos influencia em muito o sucesso ou fracasso do processo de análise dos dados. Metas bem claras e discutidas entre os usuários do processo de *Data Mining* podem evitar que sejam especificadas muitas outras metas, não previstas na tarefa de compreensão do domínio, o que pode acarretar em um desperdício de tempo e recursos na tentativa de encontrar o que se pretende obter por intermédio do processo.

Ruído nos Dados: em se tratando de dados reais, não se deve assumir que todos os valores de um conjunto de dados estejam corretos. A maioria dos algoritmos utilizados para a extração de padrões trabalham com a suposição de que os registros de uma base de dados podem incluir valores de atributos com erros. Os erros em determinados valores dos atributos são normalmente chamados de ruído. O tratamento de ruído se dá geralmente pela consideração ou não da parte dos dados que apresentam ruídos. Alguns pesquisadores consideram que não vale a pena destinar esforços para eliminar ruídos do conjunto de dados, se é bastante provável que o conhecimento seja aplicado em outros conjuntos de dados com ruído Quinlan (1993).

Dados Incompletos: é muito freqüente, em bases de dados reais, a ausência de alguns valores de atributos. Esse problema ocorre por vários motivos: armazenamento de dados impuros; perda de dados; revisão dos dados armazenados; falta de observação de atributos preditivos e falhas na medição dos valores dos atributos Fayyad et al. (1996); Ramoni & Sebastiani (1997).

Tamanho das Bases de Dados: as bases de dados diferem em alguns aspectos dos conjuntos de treinamentos utilizados em Aprendizado de Máquina. As bases de dados utilizadas em *Data Mining* normalmente são grandes, enquanto que os conjuntos de treinamentos são pequenos. Conseqüentemente, examinar a base de dados por inteiro resulta em um processo caro e, principalmente, essa quantidade de dados não é suportada pela maioria dos algoritmos de Aprendizado de Máquina utilizados para a extração de padrões. Uma possível solução para esse problema consiste em fazer uma seleção e amostragem dos dados, sendo essa uma tarefa da fase de pré-processamento.

Atualizações Constantes das Bases de Dados: em geral, as bases de dados são atualizadas com freqüência, onde informações são adicionadas, modificadas ou removidas. Qualquer conhecimento previamente extraído dessa base de dados pode vir a ficar inconsistente, ou novos conhecimentos podem surgir e ficarem ocultos nos dados. Por isso, é importante que o processo de análise de dados seja realizado periodicamente.

Pré-processamento dos Dados: a fase de pré-processamento, consome a maior parte do tempo gasto no processo de *Data Mining*. Alguns dos principais

problemas que devem ser solucionados na fase de pré-processamento são: padronização do conteúdo dos registros, eliminação dos registros duplicados e avaliação do grau de representatividade dos atributos. Vale ressaltar que muitos desses problemas podem ser solucionados de uma forma mais automática quando utiliza-se *Data Warehouse* como ferramenta de apoio Batista (2000).

Pós-processamento do Conhecimento: o pós-processamento do conhecimento extraído é uma tarefa árdua que envolve a utilização de métodos para filtrar o conhecimento, removendo assim, padrões redundantes. Além disso, outro empecilho que impede a automatização dessa fase do processo de *Data Mining*, é a necessidade da presença do especialista para auxiliar na avaliação do conhecimento extraído dos dados.

3.5 Considerações Finais

Três conceitos importantes e fundamentais para o sucesso de qualquer organização que precise de soluções para apoiar o processo de tomada de decisão foram apresentados neste capítulo. Convém ressaltar, de uma forma geral, a diferença entre esses conceitos. De um lado tem-se o *Data Warehouse*, sendo esse definido como um banco de dados que objetiva proporcionar aos seus usuários uma única fonte de dados a respeito dos seus negócios, sendo responsável pelo agrupamento dos dados históricos.

De outro lado, tem-se dois tipos de ferramentas utilizadas para analisar os dados armazenados em um *Data Warehouse*: as ferramentas OLAP e de *Data Mining*. As ferramentas OLAP são utilizadas para extrair as informações que são apresentadas aos usuários em um formato multidimensional. As ferramentas de *Data Mining*, por outro lado, permitem que os usuários explorem e descubram conhecimento a partir dos dados, encontrando relacionamentos ocultos nos mesmos.

Existe uma outra maneira de se diferenciar OLAP e *Data Mining*. Quando se tem perguntas específicas e se sabe que os dados podem fornecer a resposta desejada, deve-se utilizar OLAP, mas quando não se sabe qual a pergunta, mas mesmo assim há a necessidade de extrair o conhecimento embutido nos dados, deve-se utilizar *Data Mining*.

Neste capítulo foi apresentada uma visão geral sobre *Data Warehousing* e *Data Mining*. Além disso, foram apresentados alguns dos principais elementos que

auxiliam o processo de análise de dados, bem como, alguns dos principais problemas encontrados neste processo. No próximo capítulo serão abordadas as ferramentas utilizadas neste trabalho.

CAPÍTULO

4

Ferramentas de Apoio

Para o desenvolvimento de um *Data Warehouse* é necessário um SGBD (Sistema Gerenciador de Banco de Dados) para armazenar e gerenciar os dados, ferramentas CASE (*Computer-Aided Software Engineering*) e ETL (*Extraction, Transformation e Load*) para o projeto e povoamento do DW. Para analisar os dados armazenados no DW são necessárias ferramentas OLAP e de *Data Mining*.

Neste trabalho foi utilizado o SGBD Oracle, a ferramenta CASE e ETL Oracle Warehouse Builder, a ferramenta OLAP Oracle Discoverer e a ferramenta de *Data Mining* Visual Spotfire.

Com o objetivo de apresentar as características e funcionalidades dessas ferramentas, este capítulo foi estruturado da seguinte forma: na Seção 4.1 é apresentado o SGBD Oracle, na Seção 4.2 é apresentado o Oracle Warehouse Builder, na Seção 4.3 é apresentado o Oracle Discoverer, na Seção 4.4 é apresentado o Spotfire e, por fim, na Seção 4.5 são feitas algumas considerações finais sobre essas ferramentas.

4.1 SGBD Oracle

Nesta seção é apresentada uma visão geral das características do SGBD Oracle relacionadas ao armazenamento e gerenciamento dos dados de um *Data Warehouse*. Esse SGBD possui muitos recursos que dão um amplo suporte à tarefa de gerenciamento dos dados de um ambiente operacional, como também na construção e gerenciamento do ambiente analítico. Ele possui utilitários para extrair os dados das origens, recursos para transformar estes dados e para

carregá-los no DW. Ele também oferece recursos para agilizar o acesso aos dados, fazer o *backup* e manter a segurança do DW.

Para garantir uma boa gerência dos dados armazenados, o Oracle possui arquivos, processos e estruturas lógicas mantidos em memória ou em disco. Esses elementos constituem a arquitetura do SGBD Oracle. Uma breve síntese dos elementos que compõem essa arquitetura é apresentada a seguir Luscher (2001):

Arquivos de Dados: nesses arquivos são armazenados todos os dados do SGBD. Existem dois tipos de dados nesses arquivos: dados de usuário e de sistema. Os dados de sistema referem-se às informações que o SGBD precisa para gerenciar os dados do usuário e a si mesmo.

***Tablespace*:** é uma estrutura que armazena os objetos do SGBD. Os tipos existentes de *tablespace* são: *tablespace* de sistema, de usuário, de dados, de índice e de reconstrução.

Registros de *Redo*: também chamados de registros de transação, são arquivos especiais onde são gravadas as operações realizadas pelas transações. São importantes para que o SGBD possa efetuar recuperação do banco de dados após uma falha.

***Buffer* de *Redo Log*:** é uma pequena região de memória reservada para registrar as transações atualmente processadas pelo SGBD. Essas informações são gravadas continuamente nos registros de *redo* em intervalos configuráveis de tempo.

Arquivos de Controle: são arquivos de tamanho reduzido que contém informações importantes sobre todos os arquivos associados a um banco de dados. Esses arquivos mantém a integridade do banco e ajudam a identificar quais registros de *redo* são necessários para o processo de recuperação.

SGA e PGA: o Oracle mantém uma região de memória onde processos servidor/cliente podem se comunicar. Existem dois tipos de memória: a Área Global de Sistema (SGA) e a Área Global de Programa (PGA). A SGA é uma área de memória compartilhada por todos os processos e armazena todas as informações que os mesmos necessitam. A PGA é uma área de memória usada por um único processo cujo conteúdo são dados e informações de controle do processo.

Processos de Suporte de Banco de Dados: no Oracle existem três tipos de processos: de servidor, de cliente e de fundo. O primeiro recebe os pedidos dos processos clientes e se comunica com o banco de dados. O segundo é responsável pela requisição das informações do cliente através dos processos de servidor. O terceiro tipo de processo atende à requisições das instâncias de banco de dados. Uma instância consiste na SGA e todos os processos de fundo.

Além das estruturas inerentes à arquitetura, o SGBD Oracle oferece aos desenvolvedores e administradores objetos de banco de dados. A interface entre o usuário e o SGBD é realizada através desses objetos. A seguir é apresentada uma descrição sucinta desses objetos Luscher (2001):

Tabela: é um objeto de banco de dados que contém os dados. A informação sobre cada tabela é fornecida pelo dicionário de dados.

Visão: uma visão é uma tabela que é derivada de outras tabelas, sendo que essas outras tabelas podem ser tabelas da base de dados ou outras visões previamente definidas. Uma visão também é chamada de tabela virtual, pois não existe fisicamente. As visões que existem fisicamente são denominadas de visões materializadas.

Seqüência: é um objeto destinado a gerar valores para campos chaves nas tabelas. Esse tipo de objeto é flexível e fácil de manusear.

Índice: é uma estrutura de acesso aos dados que objetiva acelerar a busca em tabelas.

Sinônimo: é um nome alternativo para um objeto de banco de dados. É como se fosse um pseudônimo para uma tabela, por exemplo.

Concessões: são privilégios que usuários proprietários de objetos concedem a outros usuários para trabalharem com seus dados.

Personagem: é um grupo de privilégios reunidos e concedidos aos usuários. Uma vez que concedemos privilégios a um personagem, todos os usuários que se tornam membros deste personagem herdam seus privilégios.

Funções: são segmentos de código que aceitam vários parâmetros e retornam um valor para o programa que o chamou.

Procedimentos: são análogos às funções, exceto pelo fato de não retornarem uma saída para o programa que o chamou.

Gatilhos: são segmentos de código associados às tabelas e rodam de modo transparente quando ocorrem eventos predefinidos nos dados da tabela.

Pacotes: são unidades simples de programa que contem procedimentos e funções que se relacionam, podendo ser chamadas em conjunto ou uma após a outra.

Os quatro últimos objetos são utilizados para gravar no SGBD as regras de negócio da corporação e são utilizados também na tarefa de transformação dos dados durante a atividade de povoamento do DW. A linguagem utilizada para codificar esses objetos é conhecida como PL/SQL (ou SQL Estendida), que é a base de toda a programação realizada no SGBD Oracle.

O Oracle implementa segurança em vários níveis: O primeiro nível fica por conta do Sistema Operacional, através de privilégios de sistema. A segurança do segundo nível é controlada pelo próprio SGBD, com validação de usuário e senha do banco de dados. Num nível de granularidade maior, tem-se os privilégios de objetos do BD. Esses privilégios dizem respeito ao tipo de acesso ao objeto e podem ser concedidos a usuários e personagens. Pode-se conceder privilégios diferentes a usuários distintos de um mesmo objeto através de um comando SQL do tipo *GRANT*.

O SGBD Oracle possui vários recursos para auxiliar na atividade de povoamento do DW. A ferramenta que será apresentada na Seção 4.2 é utilizada para fazer o projeto da atividade de povoamento do DW. A partir do projeto desenvolvido, essa ferramenta gera código para os utilitários do Oracle que são responsáveis pela extração, transformação e carga dos dados no DW. A seguir, será feita uma síntese dos utilitários pertencentes ao SGBD que são utilizados na atividade de povoamento Lane (2001):

SQL Loader: esse utilitário tem a função de mover dados de arquivos do sistema operacional para objetos do SGBD Oracle. Uma arquivo de sistema operacional contém dados de texto alfanuméricos, tais como, letras maiúsculas e minúsculas, dígitos de 0 a 9 e uma variedade de caracteres especiais.

Import e Export: o utilitário Export permite que os dados do SGBD sejam copiados para um arquivo binário compactado e o utilitário Import permite que

os dados copiados para esses arquivos possam ser copiados de volta para os objetos do SGBD. Ambos podem ser utilizados como parte da estratégia de migração, quando se está movendo dados dos sistemas transacionais para o DW. Além disso, eles representam uma parte da estratégia de *backup* do SGBD.

PL/SQL: é a linguagem procedural do SGBD. É uma ferramenta bastante flexível, e sempre deve ser considerada como parte da solução de transformação dos dados a serem carregados no DW.

Oracle Transparente Gateway: esse recurso tem a função de acessar dados em outros SGBDs que não sejam o Oracle. O Oracle possui um Gateway para cada um dos mais variados SGBDs existentes no mercado. Uma alternativa à utilização dos Gateways são os serviços de conexão heterogênia. Com esse recurso é possível estabelecer uma conexão genérica utilizando ODBC ou OLE com os mais variados SGBDs.

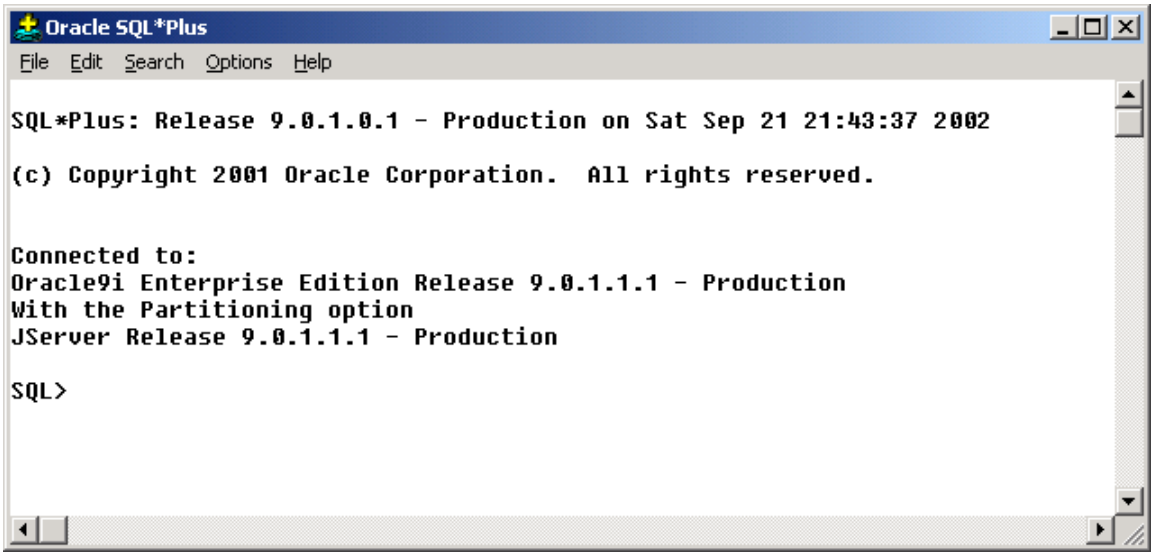


Figura 4.1: Oracle SQL*Plus.

Dentre as várias ferramentas pertencentes ao SGBD, as duas principais que são voltadas para administradores e que permitem gerenciar os objetos do banco de dados são o SQL*Plus e o Oracle Enterprise Manager (OEM). O SQL*Plus, ilustrado na Figura 4.1, é um utilitário de gerenciamento orientado por comandos. Já o OEM é um utilitário de gerenciamento visual, ou seja, tudo que é feito através de comandos no SQL*Plus pode ser feito visualmente no OEM. Na Figura

Figura 4.2: Oracle Enterprise Manager.

4.2 é apresentada a interface principal do OEM, através dele é possível executar muitos outros utilitários da Oracle.

Outro recurso do SGBD são as visões materializadas ou *snapshots*. Esse tipo de recurso é muito útil para agregar os dados de um DW, pois quando o usuário desejar visualizar esses dados resumidos, eles serão apresentados mais rapidamente, uma vez que os cálculos são previamente realizados no momento da criação da visão. Segundo Corey et al. (2001), deve-se criar visões materializadas quando o resultado dos dados agregados corresponder a no máximo 25% do tamanho total da tabela fato de detalhe correspondente. As visões materializadas podem ser utilizadas também para deixar previamente pronta a junção de uma tabela fato com suas dimensões, mas isso deve ser feito se essa operação for computacionalmente custosa para ser realizada durante a execução de uma consulta.

Quanto aos tipos de índices que devem ser utilizados, vale observar que, sobre atributos que possuem uma quantidade de valores distintos bem pequena, como é o caso de alguns atributos que fazem parte da hierarquia de uma dimensão e alguns atributos analíticos, é conveniente utilizar índices *bitmap*, uma vez que os dados no *Data Warehouse* são não voláteis. Esse tipo de índice é bem mais apropriado para esses atributos do que os convencionais índices *B-Tree*, devido

ao fato do acesso ser mais rápido e ocupar menos espaço. Para os demais atributos que possuem muitos valores distintos, como é o caso de um atributo chave, é mais conveniente utilizar índices *B-Tree* Lane (2001). Além desses, o Oracle possui mais as seguintes estratégias de indexação: índices baseados em função, índices de chave invertida e tabelas organizados por índices.

O Oracle possui também o recurso de particionamento dos dados. O particionamento envolve a divisão de tabelas grandes do DW em trechos menores e mais gerenciáveis. O uso de particionamento proporciona muitas vantagens e facilidades aos administradores de DW, pois cada partição em uma tabela particionada pode ser tratada logicamente como seu próprio objeto, quando colocada em seu próprio espaço de tabela. As linhas de cada partição podem ser excluídas ou atualizadas separadamente do conteúdo de outras partições. As partições podem ser eliminadas sem afetar os dados residentes em outras partições da tabela. Quando o volume aumenta em uma partição, ela pode ser dividida em duas partições, sem afetar o conteúdo das outras partições. Além do mais, as operações de manutenção podem ser realizadas em uma ou mais partições de uma tabela, sem afetar outras partições Corey et al. (2001).

Nesta seção foi feita uma breve apresentação das características dos recursos e utilitários do SGBD Oracle que dão um amplo suporte nas atividades de armazenamento e gerenciamento dos dados de um DW. A próxima seção, por sua vez, objetiva apresentar uma ferramenta voltada para o projeto tanto dos dados como da atividade de povoamento de um DW.

4.2 Oracle Warehouse Builder

O Oracle Warehouse Builder (OWB) é um aplicativo de múltiplas camadas: interface com o usuário, gerador de código, integradores, interface de aplicativo Java e o repositório de metadados. Juntas, essas camadas formam um produto ETL integrado que auxilia no carregamento e na manutenção do DW.

A integração do OWB com o SGBD Oracle permite que ele opere diferentemente de outras ferramentas ETL. Primeiramente, o OWB gera *scripts* que trazem os dados para o SGBD e, então, explora seus recursos gerando código PL/SQL para ajudá-lo a transformar e carregar os dados no DW. O SGBD Oracle se torna seu próprio mecanismo de transformação, eliminando a necessidade de um servidor de transformação adicional, com a vantagem de um carregamento otimizado.

A interface do OWB foi desenvolvida em Java, sendo a mesma composta pelos

seguintes componentes: conjunto de seleções de menu, barra de ferramentas, botão *wizard*, ícones de modo de operação, botão *utilities* e árvore de navegação de janela. Na Figura 4.3 é apresentado o console principal do OWB, onde são apresentados os projetos criados para as origens de dados e para a implementação do DW do PMGRN, que é apresentado no Capítulo 5.

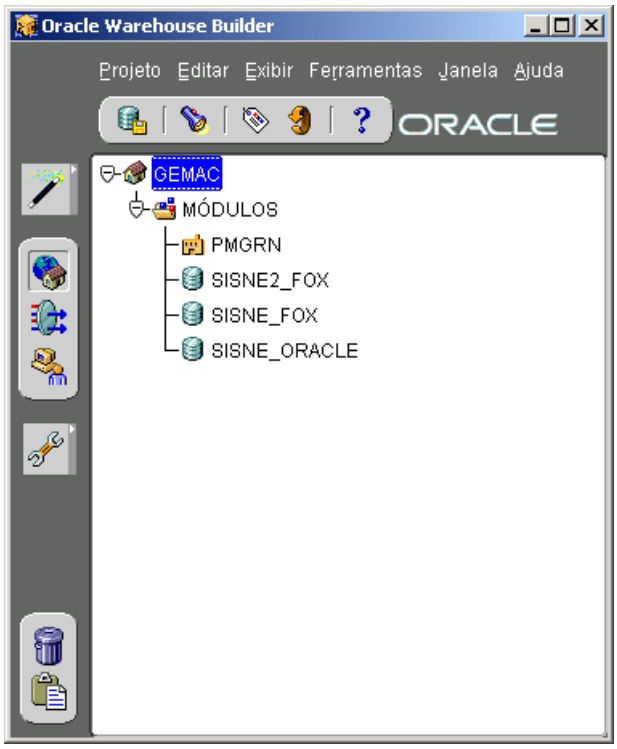


Figura 4.3: Console Principal do Oracle Warehouse Builder.

O console do OWB funciona em modos de operação. Esses modos ajudam a separar os diversos tipos de tarefas realizadas ao se construir um DW. Os três modos de operação são: Projetos, Transformações e Administração.

Na Figura 4.3 é apresentado o modo de operação Projetos. É nesse modo que os módulos de origem e de *Warehouse* (destino) são criados e onde o desenvolvedor projeta os esquemas do DW e as fontes de dados. No lado esquerdo da Figura 4.4 é apresentado o console de um módulo de origem e no lado direito da mesma figura é apresentado o console de um módulo de *Warehouse*. No modo de operação Projetos também são criados os processos de mapeamento das origens de dados para o DW e também são gerados o código e os *scripts* desses objetos e processos.

O modo Transformações tem a função de modificar transformações existentes ou criar outras personalizadas. No modo Administração são definidos os projetos

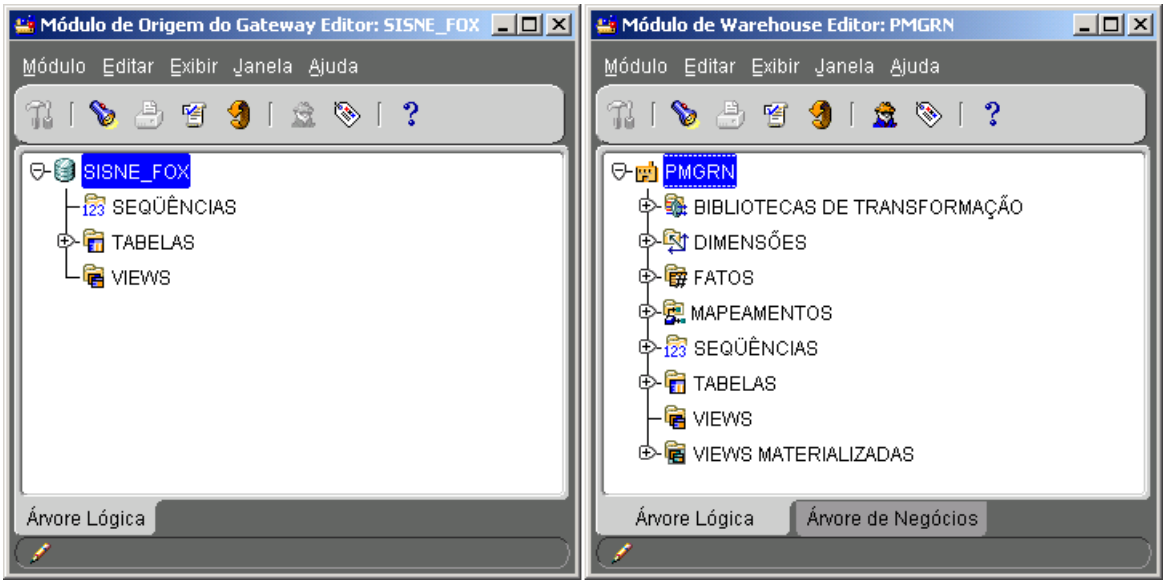


Figura 4.4: Os módulos de Origem e de *Warehouse* do OWB.

e os integradores são instalados.

Quando o desenvolvedor se conecta ao OWB para construir um DW, a primeira atividade que deve ser realizada é a criação de um projeto. É nesse projeto que todo o trabalho de construção do DW é realizado. Quando concluído, um projeto deve conter todos os objetos e definições de metadados necessários para se construir um DW.

O OWB mantém um repositório de metadados centralizado, que consiste em um conjunto de tabelas armazenadas no SGBD Oracle. As informações armazenadas no repositório de metadados podem ser utilizadas fora da ferramenta, permitindo assim, uma integração com ferramentas de *Business Intelligence* Alison et al. (2001).

Esta seção objetivou fazer uma breve apresentação dos principais componentes do Oracle Warehouse Builder. Na próxima seção será apresentada a ferramenta OLAP Oracle Discoverer.

4.3 Oracle Discoverer

O Oracle Discoverer é uma ferramenta OLAP que auxilia os usuários finais a analisarem seus dados para tomarem suas decisões, sem que eles tenham necessariamente que conhecer a estrutura do DW para acessá-lo. É uma ferramenta fácil de usar e que possibilita a criação de consultas *ad-hoc* e análise multidimensional.

Dentre as estratégias existentes de OLAP (ROLAP e MOLAP) que foram apresentadas no Capítulo 3, esta ferramenta se encaixa apenas na estratégia ROLAP, pois ela acessa somente os dados armazenados no SGBD Relacional da Oracle.

O Oracle Discoverer possui dois módulos principais: o Administration Edition e o User Edition. Com o objetivo de apresentar esses módulos, esta seção foi divida em duas subseções. Na Subseção 4.3.1 será apresentado o Administration Edition e na Subseção 4.3.2 será apresentado o User Edition.

4.3.1 Administration Edition

O Oracle Discoverer Administration Edition é voltado para os administradores de sistema. Ele importa automaticamente os metadados relacionais e permite renomear e esconder campos, definir perfis de usuários, hierarquias nas dimensões e tabelas sumarizadas. Todas essas metainformações são armazenadas na *End User Layer* (EUL), a qual objetiva esconder do usuário final a complexidade do modelo de dados e do código SQL. O console principal dessa ferramenta é apresentado na Figura 4.5.

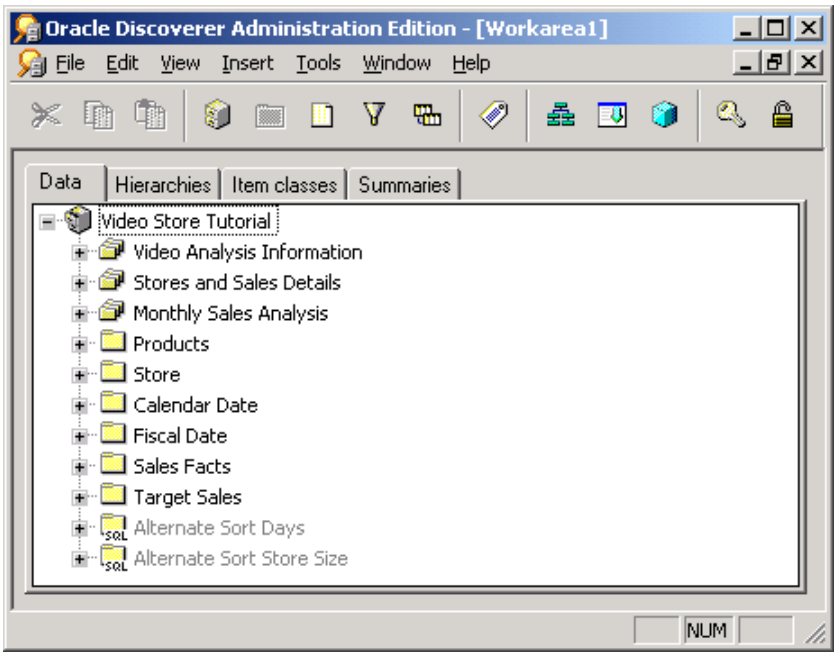


Figura 4.5: Console principal do Oracle Discoverer Administration Edition.

Além da EUL, outros elementos importantes que compõem o Administration Edition são: área de negócios, pastas, hierarquias, ítens, ítens de classe e sumários Oracle (2000).

Uma área de negócios consiste em um grupo de pastas, condições, junções,

cálculos, formatação, hierarquias e sumários que correspondem às informações de um assunto específico. Uma pasta, por sua vez, corresponde a um objeto do banco de dados, podendo ser uma tabela, visão ou o resultado de um consulta.

Uma hierarquia é criada para tornar possível a execução das operações de *drill-down* e *drill-up* sobre atributos hierárquicos. Elas não são definidas no banco de dados e devem ser criadas pelo administrador na área de negócios. Existem dois tipos de hierarquias no Discoverer: hierarquias de ítens e de data.

Um item é a representação de um atributo de uma tabela do banco de dados na EUL. Ao representar os atributos como ítens, o Discoverer habilita o administrador a fazer mudanças de formato, mudanças de nome e outros tipos de mudanças que fazem com que o usuário enxergue os dados com clareza. Os ítens são armazenados em pastas e podem ser criados, removidos e movidos entre diferentes pastas.

Um item de classe é um grupo de ítens que compartilham atributos similares. Com ítens de classe o administrador define as propriedades para ítens similares uma única vez e então atribui o ítem de classe aos ítens que possuem essas propriedades similares.

Os sumários objetivam fornecer um melhor desempenho para as consultas por usar dados pré-agregados. Com dados sumarizados as consultas poderão retornar os resultados em um tempo menor, ao contrário das consultas direcionadas para tabelas de detalhe, a qual requerem junções de várias tabelas e agregação de muitas tuplas.

Uma junção no Discoverer relaciona duas pastas usando um ou mais ítens comuns. Ela é utilizada para representar os relacionamentos existentes no banco de dados. No Discoverer também é possível a criação de cálculos sobre os ítens existentes e a criação de condições para filtrar os dados que serão apresentados.

Através do Administration Edition é possível também definir permissão de acesso e privilégios para as tarefas de cada usuário ou personagem. A permissão de acesso determina quem pode acessar os dados nas áreas de negócios e os privilégios determinam quais tarefas de cada usuário poderão ser executadas Oracle (2000).

4.3.2 User Edition

O Oracle Discoverer User Edition é voltado para os usuários do sistema. Ele permite aos usuários construir seus relatórios e gráficos de forma livre e flexível. Sua interface permite que os usuários acessem e analisem seus dados facilmente

com base na área de negócios criada no Administration Edition Brownbridge (2000).

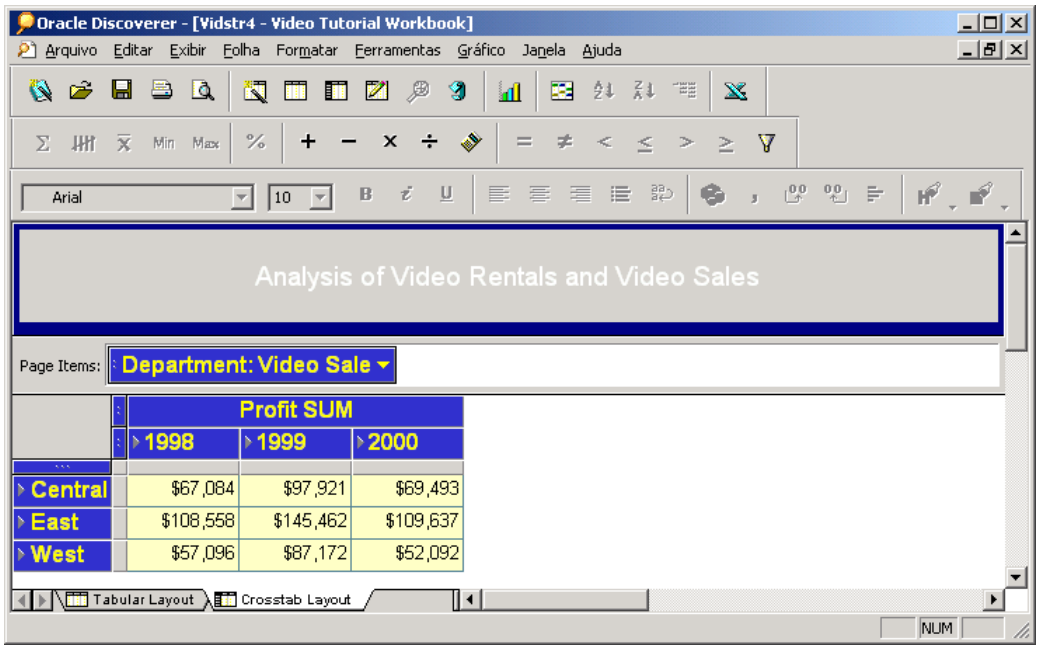


Figura 4.6: Console principal do Oracle Discoverer User Edition.

Os principais componentes do Oracle Discoverer User Edition são os cadernos de trabalho (*workbooks*), folhas (*sheets*) e gráficos. Na Figura 4.6 é apresentado um caderno de trabalho exemplo pertencente à instalação do Discoverer. Esse caderno de trabalho possui duas folhas: *Tabular Layout* e *Crosstab Layout*.

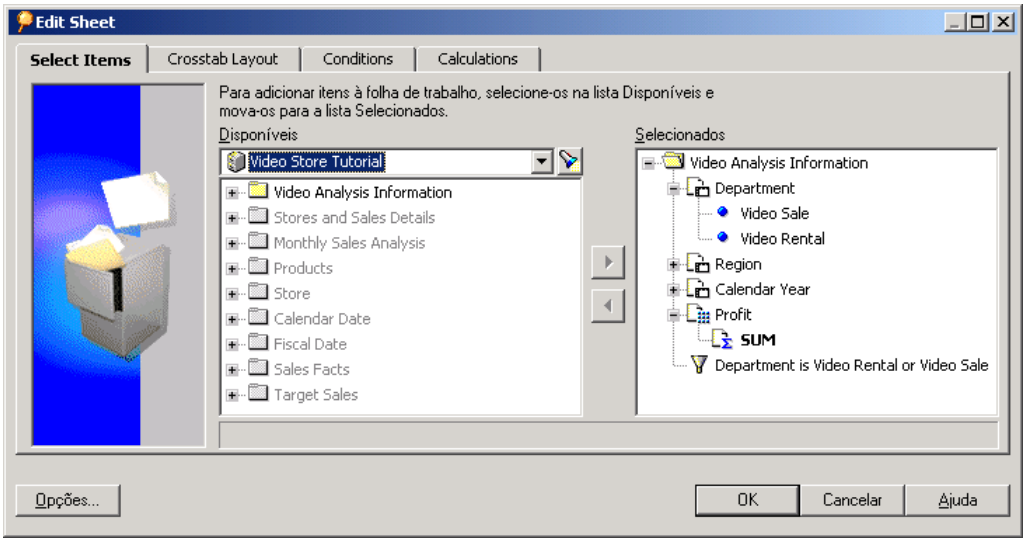


Figura 4.7: Console do Editor de Folhas.

As folhas fazem parte dos cadernos de trabalho e cada uma corresponde a uma consulta OLAP. Na Figura 4.7 é apresentado o editor de folhas. Ele é composto por quatro abas. Na aba *Select Items* é feita a seleção dos ítens que irão compor a consulta, na aba *Crosstab Layout* é montado o *layout* da consulta, na aba *Conditions* são especificados os filtros a serem aplicados sobre a consulta e na aba *Calculations* são criados ítens correspondentes a cálculos. Nesse editor o usuário poderá adicionar ou remover novos ítens, mudar o *layout* da consulta, criar ou remover condições e ítens calculados Brownbridge (2000).

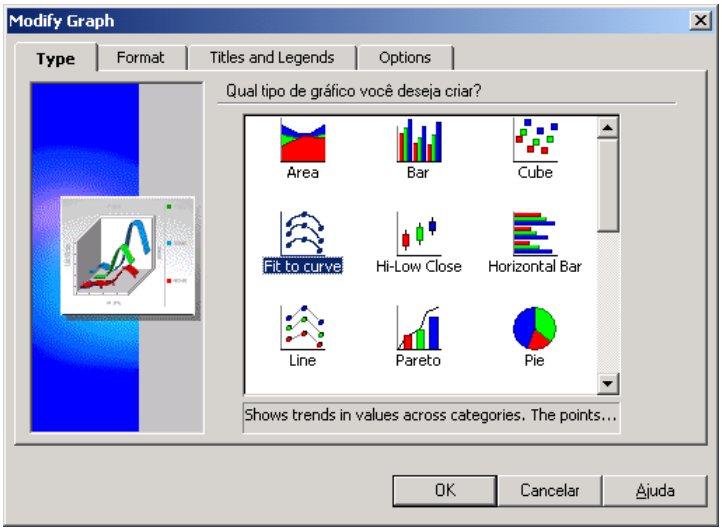


Figura 4.8: Console do Editor de Gráficos.

Para cada folha pode ser criado um gráfico para auxiliar na análise dos dados. Na Figura 4.8 é apresentado o editor de gráficos, sendo o mesmo composto por quatro abas. Na aba *Type* é especificado o tipo do gráfico, na aba *Format* são apresentados vários formatos que estão relacionados ao tipo do gráfico, na aba *Titles and Legends* é feita a especificação e formatação do título do gráfico, legendas e *labels* dos eixos e, por fim, na aba *Options* são escolhidas a escala do gráfico e linhas de grade. Neste editor os usuários poderão alterar o tipo e formato dos gráficos, título, legendas, *labels*, escala e etc Brownbridge (2000).

O Oracle Discoverer User Edition possui ainda ferramentas para ordenar, calcular totais e percentis sobre os dados resultantes de uma consulta e ainda permite que sejam passados parâmetros para a consulta.

Esta seção apresentou uma visão geral dos módulos do Oracle Discoverer, o Administration Edition e o User Edition. A próxima seção se focará em apresentar o Spotfire.

4.4 Spotfire

Enquanto a ferramenta OLAP Oracle Discoverer objetiva fornecer visões multidimensionais de dados resumidos, a ferramenta de *Data Mining* Visual Spotfire Ahlberg (1996) objetiva apresentar os dados em uma variedade de gráficos, onde o especialista pode extrair padrões por meio de análises visuais.

A visualização de dados pode ser usada em várias situações no processo de *Data Mining*. Pode ser usada como ferramenta de auxílio para a seleção dos dados na fase de pré-processamento, pode ser usada na fase de pós-processamento, para avaliar os resultados obtidos e, pode também ser utilizada para minerar os dados visualmente. A visualização de dados foi utilizada neste projeto para minerar os dados do PMGRN, os gráficos gerados com esses dados são apresentados no Capítulo 6.

As ferramentas modernas de visualização combinam a capacidade de construir cenas visuais complexas com controles de seleção interativa dos dados apresentados. Essa funcionalidade permite que um especialista possa explorar os dados. Em certas ferramentas a exploração pode ser feita tão facilmente que o próprio especialista pode detectar visualmente novos padrões de interesse. Esse tipo de exploração interativa de dados é chamada de *Data Mining* Visual ou mineração visual de dados. Uma boa ferramenta de *Data Mining* Visual tem as seguintes funcionalidades:

- Permite o uso de diversos atributos visuais, tais como, forma, cor, posicionamento, brilho, opacidade, tamanho, rotação e etc, para produzir gráficos multidimensionais de fácil interpretação.
- Permite navegação interativa, tais como, aproximação, rotação, reposicionamento e varreduras.
- Permite controle interativo dos formatos de apresentação e atributos visuais dos gráficos apresentados.
- Permite controle interativo dos dados, habilitando o especialista a enxergar os dados de uma perspectiva geral e permitindo também que o mesmo analise os dados em uma perspectiva mais detalhada.

O Spotfire possui uma interface amigável e de fácil utilização. Na Figura 4.9 é apresentado o console principal dessa ferramenta, onde estão presentes seus

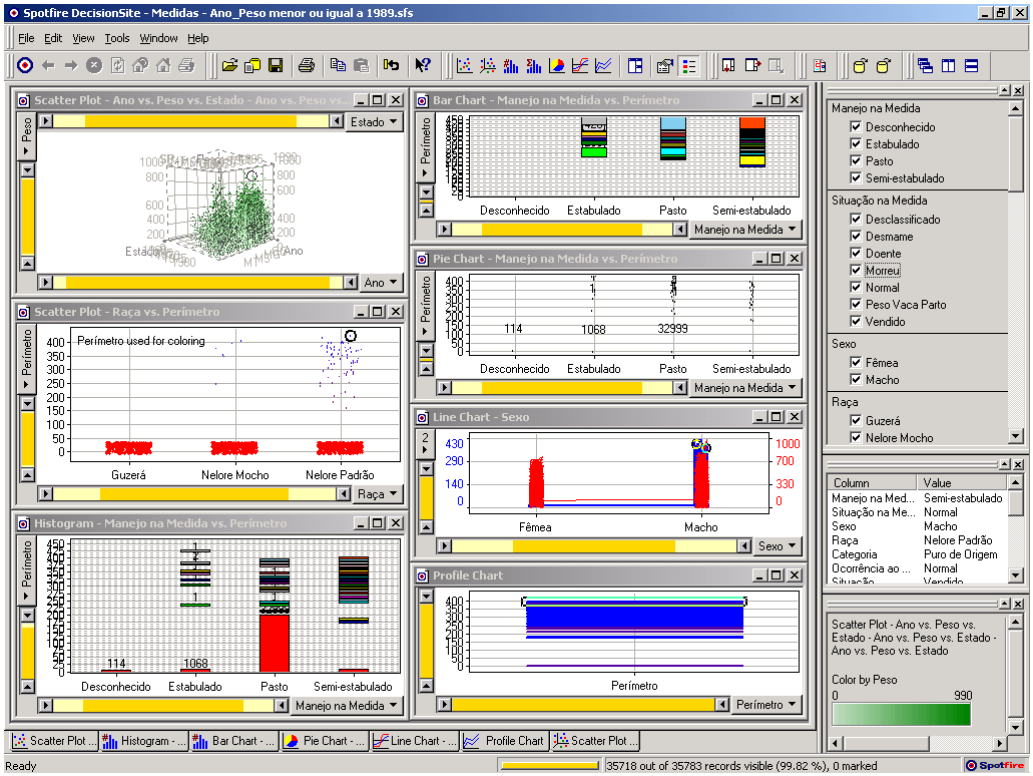


Figura 4.9: Console principal do Spotfire.

principais componentes. Conforme ilustrado na figura, é possível criar sete tipos de gráficos: *Scatter Plot 2D, Scatter Plot 3D, Histogram, Bar Chart, Pie Chart, Line Chart* e *Profile Chart*. Além dos gráficos, no lado direito da figura são apresentados também os componentes *Query Devices, Details on Demand* e *Legend*. O componente *Query Devices*, no topo, tem a função de apresentar os atributos e seus respectivos valores, logo, o usuário pode escolher os valores de um determinado atributo. O componente *Details on Demand*, no meio, tem a função de apresentar o valor de um ou mais pontos selecionados no gráfico e, por fim, o componente *Legend* apresenta a Legenda para os atributos selecionados.

Cada tipo de gráfico possui um editor de propriedades diferente. Na Figura 4.10 é apresentado o editor de propriedades para o gráfico *Scatter Plot*. Esse editor possui seis abas. Na aba *Annotations* são especificados o título do gráfico e alguns comentários. Na aba *Data and Background* são especificados os eixos visíveis e seus *labels*, bem como, a imagem e o texto do fundo. Na aba *Columns* é feita a formatação dos atributos. Na aba *Trellis* são especificados os atributos que formarão uma nova dimensão através da criação de um gráfico para cada valor de atributo. Nessa aba é possível também especificar o *layout* e formatação

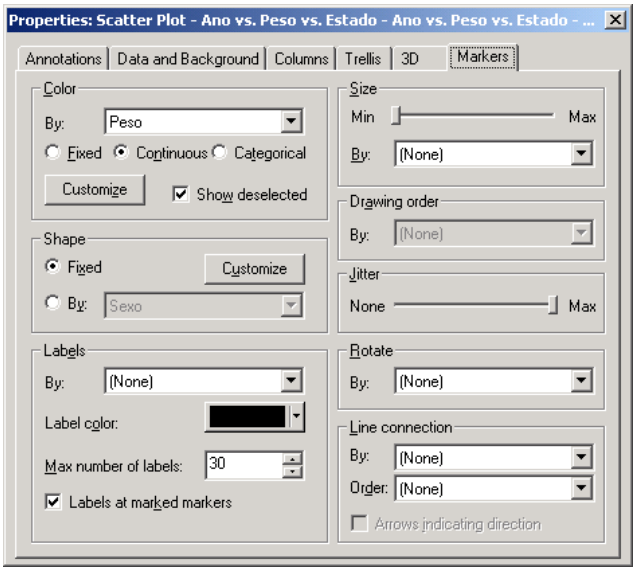


Figura 4.10: Editor de Propriedades do Gráfico *Scatter Plot*.

desses novos gráficos gerados. Na aba 3D é indicado se o gráfico será 2D ou 3D, se ele terá luzes e qual sua perspectiva e, por fim, na aba *Markers* é possível especificar outras dimensões para o gráfico através do uso de cores, formas e tamanhos. É possível também mudar a quantidade de *labels* visíveis, o eixo de rotação e o tamanho na qual as formas serão renderizadas (*Jitter*) Spotfire (2002).

4.5 Considerações Finais

Muitos fornecedores de soluções para DWs procuram oferecer todas as ferramentas de forma integrada. A vantagem é que o desenvolvimento com esse tipo de ferramenta é geralmente realizado em torno de um repositório de metadados, tornando o projeto e a manutenção do sistema muito mais simplificado. A desvantagem é que a organização fica a mercê de um único fornecedor, sendo que muitas vezes as soluções oferecidas por esse para determinada etapa do projeto podem não ser as melhores se comparadas com as existentes no mercado.

Das quatro ferramentas apresentadas nesta seção, as três primeiras pertencem a um mesmo fornecedor e são boas ferramentas para serem utilizadas no desenvolvimento de sistemas de *Data Warehousing*. A quarta ferramenta, que pertence a outro fornecedor, também é uma boa ferramenta para fazer *Data Mining* visual. Após uma breve apresentação das ferramentas que foram utilizadas, a próxima seção se focará em apresentar o processo de desenvolvimento do *Data Warehouse* do PMGRN.

CAPÍTULO

5

Desenvolvimento do *Data Warehouse*

Este capítulo objetiva apresentar o processo de desenvolvimento do *Data Warehouse* para o Programa de Melhoramento Genético da Raça Nelore seguindo a metodologia apresentada na Subseção 3.1.4. Para isso, o mesmo está estruturado da seguinte forma. Na Seção 5.1 são apresentadas as justificativas para o desenvolvimento do DW. Na Seção 5.2 é abordado o planejamento do sistema. Na Seção 5.3 são apresentados os resultados adquiridos na fase de análise. Na Seção 5.4 são relatados os resultados da fase de projeto. Na Seção 5.5 é mostrada a etapa de implementação. Na Seção 5.6 é apresentada a revisão do sistema e, por fim, na Seção 5.7 são apresentadas as considerações finais para este capítulo.

5.1 Justificativa

Pode-se destacar como justificativas para o desenvolvimento de um DW para o PMGRN, os seguintes ítens:

- A implementação do *Data Warehouse* utilizando um SBGD mais poderoso e que oferece mais recursos para a execução de consultas OLAP, uma vez que o banco de dados operacional atual está implementado em um SGBD com recursos e capacidade um tanto quanto limitados.

- Os dados em um *Data Warehouse* estão limpos e transformados, garantindo assim uma maior qualidade e confiabilidade nos mesmos, que são utilizados para o processo de tomada de decisão. Dados limpos também

são indispensáveis no processo de *Data Mining*, uma vez que parte desse processo corresponde ao tratamento dos dados.

- Com os dados armazenados em uma estrutura multidimensional relacional, como o esquema estrela, as análises multidimensionais realizadas com ferramentas OLAP tornam-se mais eficazes, uma vez que a quantidade de junções de tabelas são reduzidas. Dessa forma conseguir-se-á uma maior agilidade no acesso aos dados e respostas mais adequadas diante de situações inesperadas.

- Com as ferramentas OLAP, os usuários do PMGRN podem ter visões multidimensionais dos dados, ao contrário da tradicional visão tabular. Essas visões multidimensionais auxiliam efetivamente ao tomador de decisão, pois ele pode verificar tendências nos dados, utilizando dados resumidos, trocando as dimensões de lugar e navegando através de suas hierarquias. Dessa forma, os usuários podem testar suas hipóteses e pensar sobre questões que não haviam ainda sido levadas em consideração.

Apresentadas as justificativas para o desenvolvimento do *Data Warehouse*, já se pode passar para a fase de planejamento.

5.2 Planejamento

Esta fase objetiva desenvolver um plano estratégico de como um sistema de *Data Warehousing* será desenvolvido na organização. Algumas tarefas devem ser desenvolvidas durante esta fase, sendo elas, definição de uma visão corporativa do *Data Warehouse*, definição da arquitetura e topologia do sistema, e definição da área de negócio a ser trabalhada.

Para a aplicação em questão, uma visão corporativa do *Data Warehouse* é a seguinte. A corporação como um todo é o Departamento de Genética e os assuntos, ou áreas de negócio, são os laboratórios que compõem o Departamento: Imunogenética, Genética Bioquímica, Investigação de Paternidade, Abelhas, Melhoramento Animal, Molecular e Microrganismos, Citogenética, Genética Médica, Drosophila e Genética Molecular. Cada um desses laboratórios são *Data Marts* independentes entre si e dependentes unicamente do *Data Warehouse* corporativo do Departamento de Genética. Convém ressaltar, que o foco deste trabalho é a implementação de um *Data Mart* para o laboratório de Melhoramento Animal, no qual faz parte o PMGRN.

De acordo com a experiência adquirida por muitos especialistas em Sistemas de Suporte a Decisão, e conforme seus relatos, sabe-se que o ideal é implementar um assunto da organização, ou seja, um *Data Mart* por vez, sendo os outros *Data Marts* implementados aos poucos, de forma incremental Corey et al. (2001).

Para este trabalho foi definida e utilizada a arquitetura e topologia apresentada na Figura 5.1. Conforme pode-se observar, a topologia escolhida foi a de *Data Marts* dependentes de um *Data Warehouse*. Com essa topologia novos assuntos poderão ser integrados ao *Data Warehouse* e, a partir do DW, o *Data Mart* correspondente ao assunto adicionado será construído.

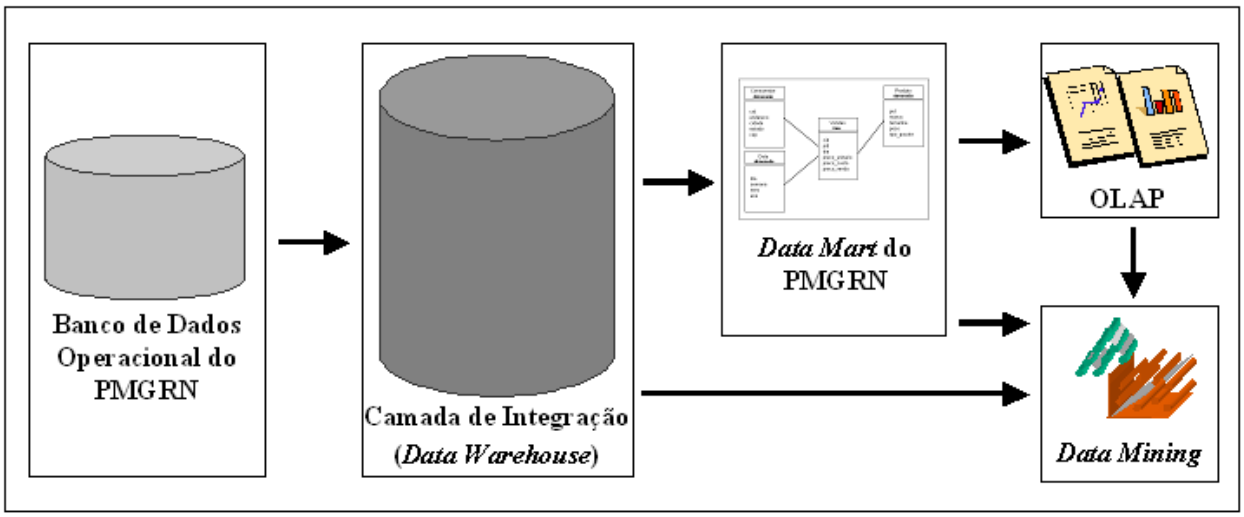


Figura 5.1: Arquitetura do *Data Warehouse*.

Nessa arquitetura, os dados são extraídos do banco de dados operacional do PMGRN, passam por um processo de limpeza e transformação e são carregados na camada de integração, ou o *Data Warehouse* propriamente dito. O modelo de dados conceitual utilizado no projeto da camada de integração foi o Diagrama Entidades-Relacionamentos (DER). Após os dados estarem carregados no DW, o próximo passo foi colocá-los em uma estrutura de consulta de alto desempenho, os *Data Marts*. Para o projeto do *Data Mart* do PMGRN foi utilizado o esquema estrela, uma vez que os dados foram armazenados em um SGBD relacional.

Com os dados armazenados no *Data Mart*, eles já podem ser acessados pelas ferramentas OLAP e tanto os dados contidos no *Data Mart*, quanto os dados contidos no *Data Warehouse* e as resultados das consultas OLAP, podem ser minerados por meio do processo de *Data Mining*.

Na arquitetura escolhida para o desenvolvimento deste DW foi utilizada uma área de adaptação virtual. Pois o trabalho necessário para realizar as trans-

formações nos dados do PMGRN permitiu que as funções fossem executadas dinamicamente na memória. Essa área está localizada entre o banco de dados operacional e a camada de integração. Finalizada a fase de planejamento, já se pôde passar para a fase de análise.

5.3 Análise

Com a especificação dos requisitos verificou-se quais informações são importantes para as necessidades de análises dos usuários. Detectou-se quais as entidades que compõem a camada de integração, sendo elas: Animal, Categoria (categ), Criador, DEP, Fazenda, Manejo, Ocorrência ao Parto (ocoparto), Peso Observado (peso_obs), Ponderal, Raça, Reproduz, Sexo, Série, Situação, Situação ao Nascimento (sitnasc), Tipo de Cadastro (tipocad) e Tipo de Acasalamento (tipoacas). Detectou-se também seus respectivos relacionamentos e atributos, conforme apresentado no DER da Figura 5.2.

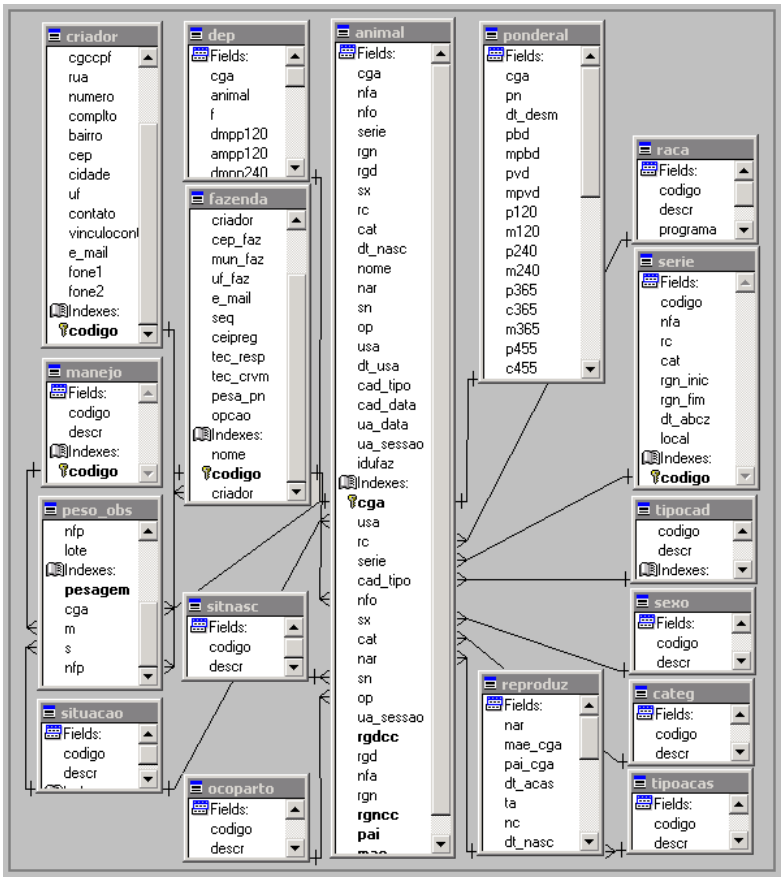


Figura 5.2: Diagrama Entidades-Relacionamentos da Camada de Integração.

Quanto ao *Data Mart* do PMGRN, detectou-se as dimensões que o compõe

à partir da desnormalização de várias entidades da camada de integração. Na Tabela 5.1 são apresentadas as dimensões e as respectivas entidades que foram desnormalizadas para formá-las. Em seguida, na Tabela 5.2, são apresentados os atributos que compõem as respectivas dimensões.

Dimensão	**Entidades**
Fazenda	Fazenda e Criador
Animal	Animal, Categoria, Sexo, Raça, Situação, Situação ao Nascimento e Ocorrência ao Parto
Reprodução	Reproduz, Tipo de Acasalamento e Manejo
Ponderal	Ponderal e Manejo
Medidas	Peso Observado, Manejo e Situação

Tabela 5.1: As dimensões e as entidades que foram desnormalizadas.

Fazenda	**Animal**	**Reprodução**	**Medidas**	**Ponderal**
Id_Fazenda	Id_Animal	Nar	Id_Animal	Id_Ponderal
Codigo	Cga	Tipo_Acas	Cga	M120
Nome_Fazenda	Nome	Manejo	Dt_Peso	M240
Nome_Criador	Sexo	Dt_Insercao	Manejo	M365
Municipio	Raca		Situacao	M455
Estado	Categoria		Dt_Insercao	M550
Pessoa	Sit_Nasc			M730
Pesa_ao_Nascer	Situacao			MPBD
Dt_Insercao	Oco_Parto			MPVD
Dt_Atualizacao	Dt_Insercao			Dt_Insercao
Flag	Dt_Atualizacao			
	Flag			

Tabela 5.2: As dimensões e seus atributos.

Os atributos Id_Fazenda e Id_Animal são chaves substitutas. Os atributos Codigo, Cga, Nar e {Cga, Dt_Peso} são as chaves originais da base de dados operacional. Os atributos Dt_Insercao, Dt_Atualizacao e Flag são utilizados, respectivamente, para indicar a data em que o registro foi inserido, atualizado e se o mesmo está ativo ou não na dimensão. Os demais são atributos analíticos.

Verificou-se também nesta fase de análise de requisitos, as hierarquias inerentes às dimensões, sendo as mesmas apresentadas na Tabela 5.3.

A hierarquia Fazenda indica que um Estado possui vários Municípios, que possuem vários Criadores, que por sua vez, podem possuir mais de uma Fazenda. O mesmo ocorre com as hierarquias Reprodução, Medidas e Tempo. Sobre essas hierarquias são realizadas as operações de *Drill-down* e *Drill-up* das consultas OLAP que foram implementadas e serão apresentadas no Capítulo 6.

Dimensão	Hierarquias
Fazenda	Estado ⇒ Município ⇒ Criador ⇒ Fazenda
Tempo	Ano ⇒ Mês
Reprodução	Tipo acasalamento ⇒ Manejo da vaca ao parto
Medidas	Manejo ⇒ Situação

Tabela 5.3: As dimensões e as suas hierarquias.

Os atributos que compõem uma hierarquia formam um agrupamento hierárquico natural entre si.

Detectou-se também os fatos e as operações agregadas que podem ser aplicadas sobre eles. Foram identificadas quatro tabelas de fatos que possuem vários atributos que podem ser quantificados e que possuem características comuns, sendo elas: DEPs, Medidas, Ponderal e Reprodução. Tem-se à seguir os esquemas estrela com as medidas que compõem as tabelas de fatos.

Os nomes dos atributos da tabela de fatos correspondem aos nomes utilizados no banco de dados operacional. Na tabela de fatos DEPs, Figura 5.3, os atributos correspondem a valores das DEPs dos animais (DMPPXXX, DDPPXXX, DDPEXXX), acurácia das DEPs (são os atributos que começam com a letra 'A'). O número anexado ao atributo indica a idade do animal em dias, na qual a DEP foi medida. O atributo MGT indica o Mérito Genético Total, os atributos NRXXX e NFXXX indicam número de rebanhos e número de filhos respectivamente e o atributo F indica o coeficiente de endogamia.

Na tabela de fatos Medidas, Figura 5.4, tem-se dois atributos: PESO (peso no animal em kg) e CE (perímetro escrotal do animal). Na tabela de fatos Ponderal, Figura 5.5, tem-se vários atributos, onde os que iniciam com a letra 'P' indicam valores ponderais de pesos e os que iniciam com a letra 'C' indicam valores ponderais de perímetro escrotal. Na tabela de fatos Reprodução, Figura 5.6, tem-se os atributos: NC (número de cobertura), PVP (peso da vaca ao parto), IVP (idade da vaca ao parto), IEP (intervalo entre partos), PG (período de gestação) e PS (período de serviço). Os atributos que terminam com 'FK', são as chaves estrangeiras para as respectivas dimensões.

Sobre os fatos foram aplicadas operações agregadas, tais como, média, mínimo, máximo, desvio padrão, funções de classificação e funções estatísticas. Como existem quatro tabela de fatos, o modelo conceitual possui quatro esquemas estrela, que formam um esquema constelação, pois alguns fatos compartilham certas dimensões com outro fatos.

Na Figura 5.7 é apresentado o esquema constelação para o *Data Mart* do PM-

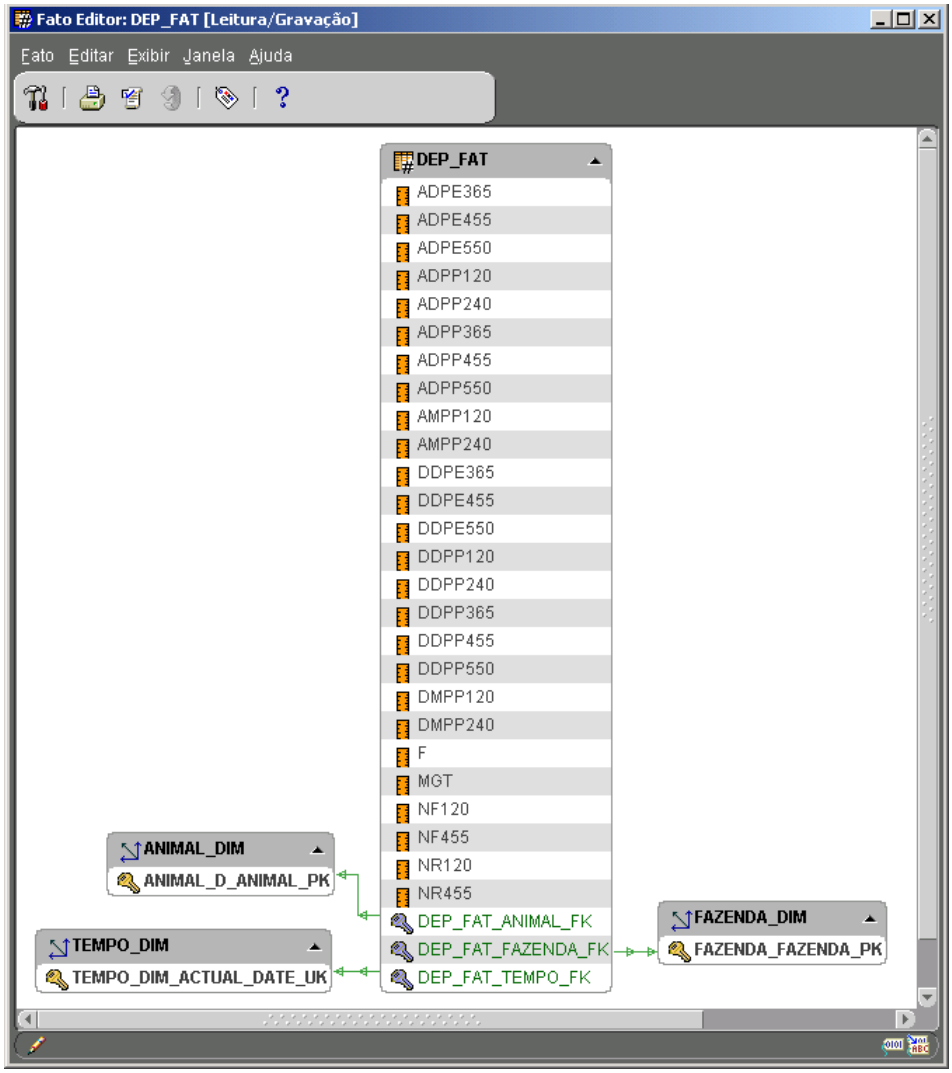


Figura 5.3: Esquema estrela para DEPs.

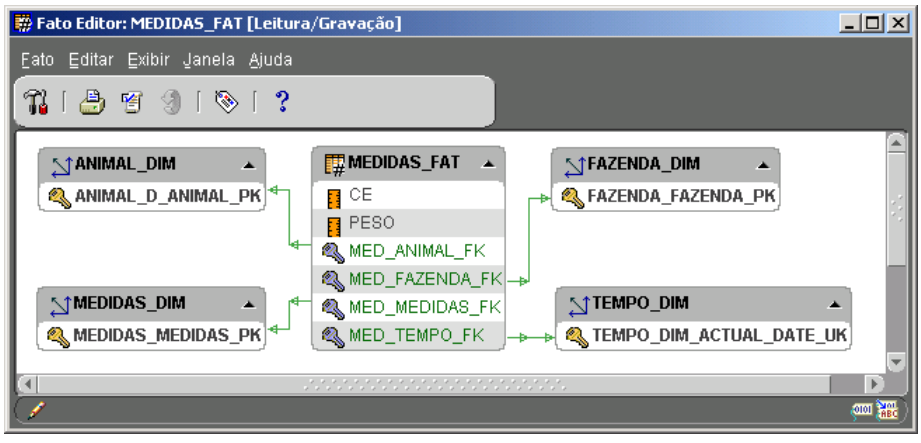


Figura 5.4: Esquema estrela para Medidas.

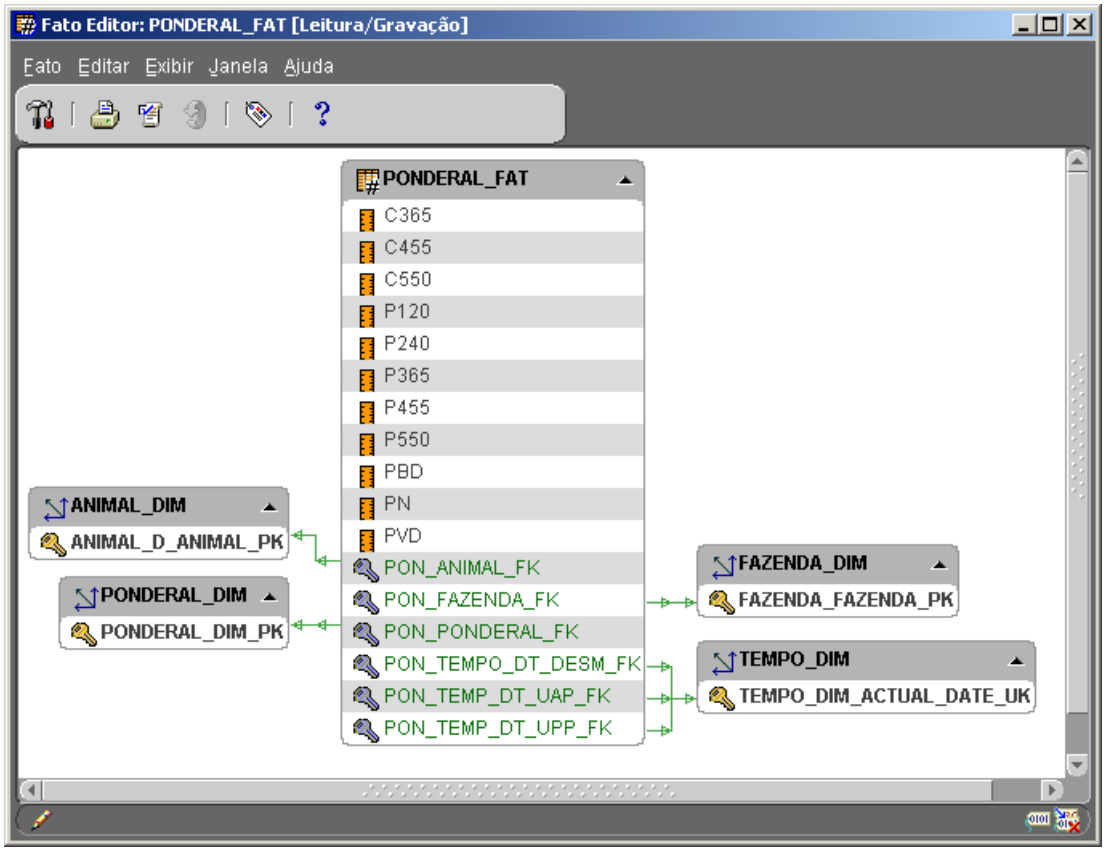


Figura 5.5: Esquema estrela para Ponderal.

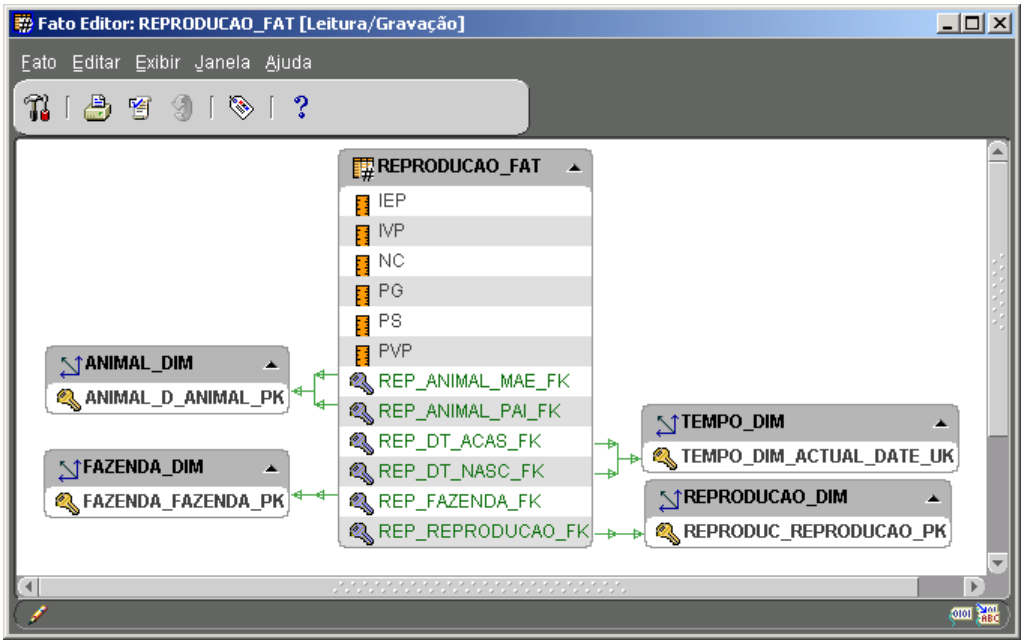


Figura 5.6: Esquema estrela para Reprodução.

GRN. Nela estão ilustrados os fatos, dimensões e seus respectivos relacionamentos. Finalizada a fase de análise, passou-se para a fase de projeto.

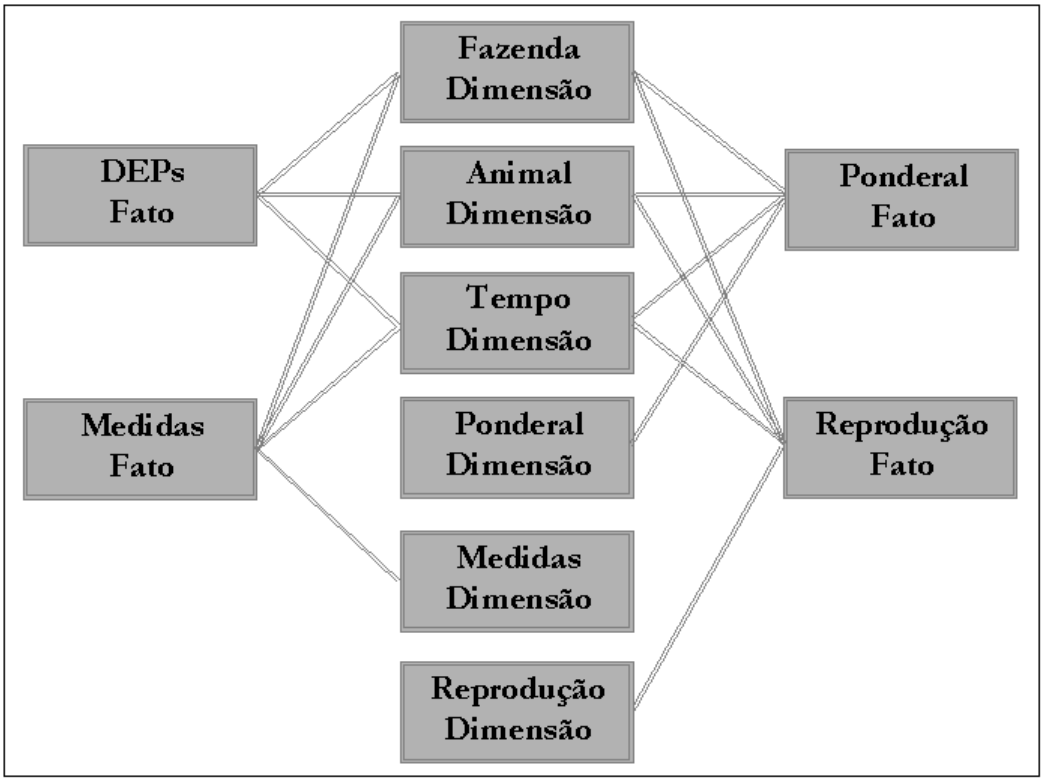


Figura 5.7: Esquema constelação do *Data Mart* do PMGRN.

5.4 Projeto

Enquanto na fase de análise procurou-se desenvolver uma visão bem alto nível dos modelos de dados, no projeto esses modelos foram refinados e detalhados, para que pudessem ser implementados. Para a camada de integração foram definidas as tabelas, os tipos de dados dos atributos, as restrições de integridade, as restrições referenciais e os índices.

Quanto ao *Data Mart* também foram adicionados mais detalhes de implementação aos fatos e às dimensões. Para as tabelas de fatos foram especificados os tipos dos atributos, restrições de integridade e referenciais, uma vez que a tabela de fatos possui chaves estrangeiras para as tabelas de dimensão. Também foi feita a especificação de índices sobre os atributos de chave primária e estrangeira. Para as dimensões foram especificados os tipos dos atributos, as restrições de integridade e os índices. Observa-se que no caso das dimensões não existe integridade referencial.

Quanto à necessidade de particionar as tabelas, tomou-se como base o relato feito no projeto desenvolvido pela IBM Bustamente & Sorenson (1994), no qual recomenda-se que sejam particionadas tabelas que possuem mais de dois milhões de tuplas. Como neste projeto a maior tabela possui aproximadamente 1.500.000 tuplas, não foi necessário utilizar nenhuma técnica de particionamento de dados. Quanto à granularidade dos dados, decidiu-se armazenar o máximo de detalhes possíveis para as tabelas de fatos, pois para a aplicação em questão, algumas operações agregadas não produzem resultados corretos se forem aplicadas sobre dados resumidos.

Nesta fase foram definidos também os mapeamentos dos dados dos sistemas de origem (o banco de dados operacional) para os sistemas de destino (o *Data Warehouse* e daí para o *Data Mart*). Inicialmente, para se fazer o mapeamento entre a origem e o destino, foi necessário criar, no Oracle Warehouse Builder, os módulos de origem e destino correspondentes. Na Figura 5.8 é apresentado, à esquerda, o módulo de origem chamado SISNE_FOX, que representa as tabelas do banco de dados FoxPro, sendo que esses metadados foram extraídos diretamente da fonte. À direita desse está o módulo de *Warehouse* de destino, o qual representa as tabelas da camada de integração e os fatos e dimensões do *Data Mart*.

Uma vez definidos a origem e o destino passou-se para a etapa de criação dos mapeamentos. Nessa etapa seguiu-se a estratégia de criar um mapeamento que envolve-se apenas uma tabela de origem e uma de destino. Os mapeamentos que foram criados estão ilustrados no lado esquerdo da Figura 5.9, esses mapeamentos envolvem rotinas que extraem, transformam e carregam os dados da base de dados operacional para a camada de integração e também da camada de integração para o *Data Mart*. As funções e procedimentos utilizados nos mapeamentos para fazer as devidas transformações nos dados são ilustradas no lado direito da Figura 5.9. Como existem muitos mapeamentos, fica inviável apresentar todos, portanto serão apresentados apenas três tipos diferentes de mapeamentos.

A Figura 5.10 ilustra um dos mapeamentos construídos, realizado entre a tabela do banco de dados operacional, peso_obs e a tabela do *Data Warehouse*, PESO_OBS. São utilizadas duas funções de transformação (FAZENDA_ID_FAZENDA e ANIMAL_ID_ANIMAL) e um gerador de dados (CONST).

A Figura 5.11 ilustra um outro mapeamento, realizado entre a tabela do DW, REPRODUZ e a tabela fato REPRODUCAO. Neste mapeamento é utilizada apenas uma função de transformação (ANIMAL_NAR_FAZENDA).

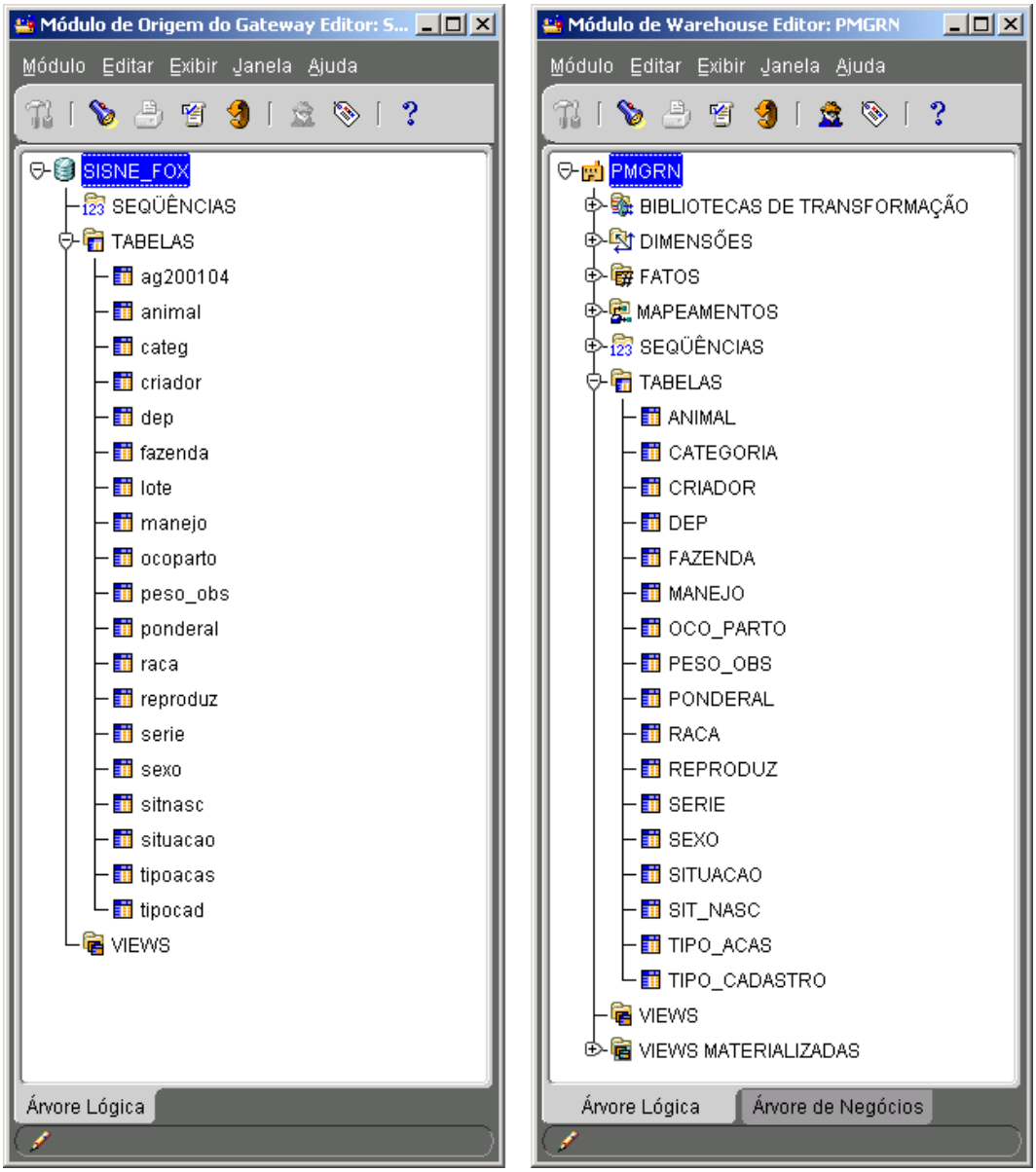


Figura 5.8: Os módulos de origem e destino.

A Figura 5.12 ilustra um outro mapeamento, realizado entre a tabela do DW, REPRODUZ e a dimensão REPRODUCAO. Neste mapeamento são utilizadas duas funções de transformação (TIPO_ACAS_DESCRICAO e MANEJO_DESCRICAO).

Outro elemento importante considerado na fase de projeto é a captura das mudanças ocorridas no banco de dados operacional. Em outras palavras, após a carga inicial dos dados no DW, que estratégia deve ser utilizada para realizar a extração, transformação e carga das mudanças ocorridas no banco operacional, para que as mesmas sejam refletidas no DW. Outro fator importante a ser considerado, é o intervalo de tempo no qual essa operação de atualização do DW deve

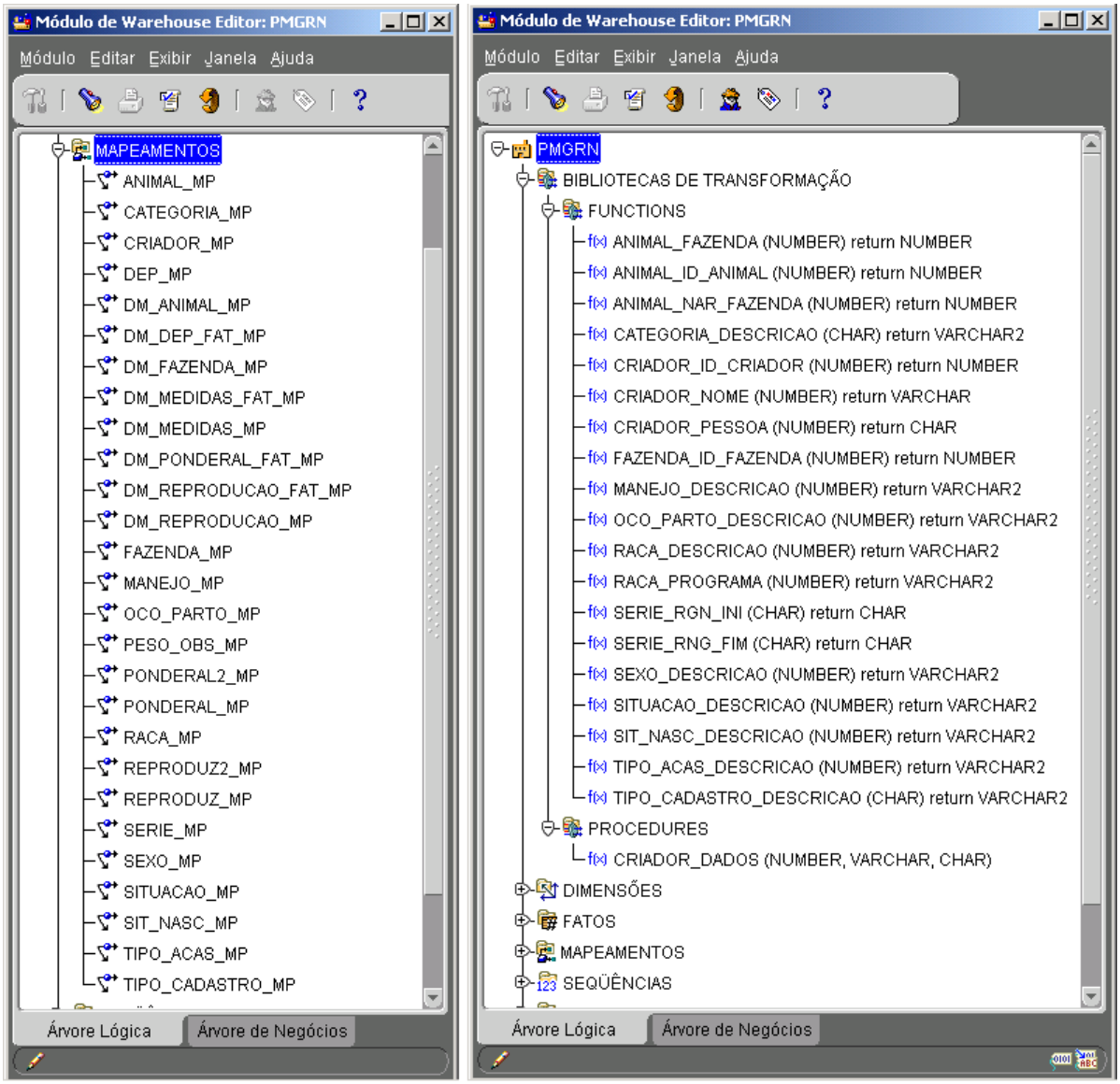


Figura 5.9: Os mapeamentos e suas funções de transformação.

ser realizada.

Como existe o elemento tempo associado a todas as tabelas do banco de dados operacional, tais como, a data em que uma medida foi realizada, ou a data em que um animal foi cadastrado ou alterado, esse será o ponto chave para identificar as alterações ocorridas no banco operacional, pois dessa forma, somente as tuplas novas e as que foram alteradas serão carregadas no *Data Warehouse*. Pelo fato da avaliação genética ser realizada duas ou três vezes por ano, esse é o intervalo de tempo ideal para fazer a atualização do DW e do Data Mart do PMGRN, ou seja, essa operação de atualização será realizada após a execução

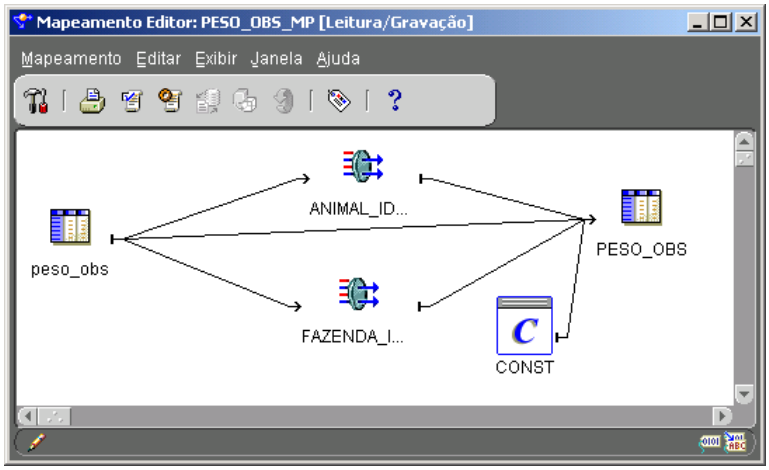


Figura 5.10: Um mapeamento da fonte de dados para o DW.

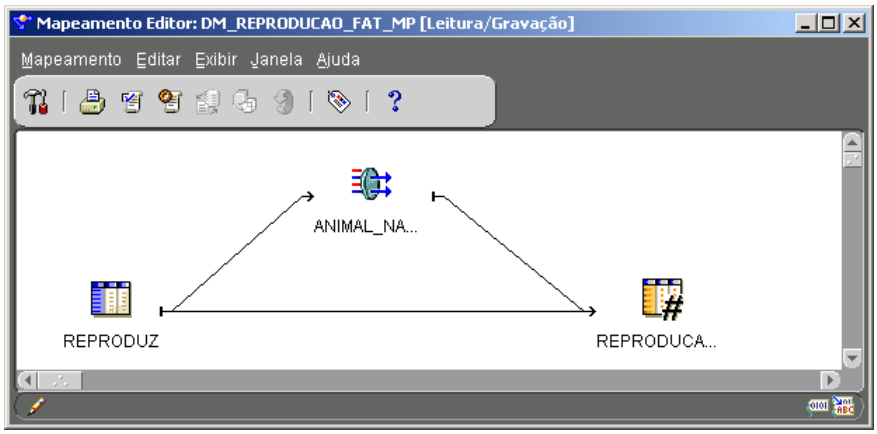


Figura 5.11: Um mapeamento de uma tabela do DW para uma tabela fato.

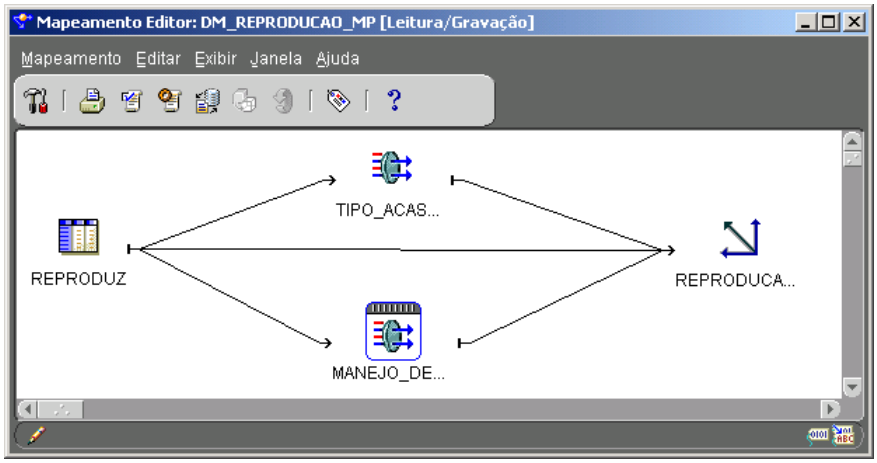


Figura 5.12: Um mapeamento de uma tabela do DW para uma dimensão.

de uma avaliação genética.

Definidos os detalhes da fase de projeto, passou-se para a implementação dos objetos do banco de dados e para o povoamento do *Data Warehouse*.

5.5 Implementação

Finalizada a modelagem do *Data Warehouse*, do *Data Mart* e dos mapeamentos, gerou-se os *scripts* SQL que definem os objetos de banco de dados, tais como tabelas, dimensões e índices. Gerou-se também os *scripts* referentes aos mapeamentos, sendo que esses consistem de *packages*, com código PL/SQL, que extraem os dados do banco operacional, via ODBC, fazem as devidas transformações e os carrega no destino.

Os *scripts* de criação dos objetos foram executados no SQL*Plus. Após a criação dos objetos, já se pôde povoar o *Data Warehouse*, através da execução das *packages* correspondentes a cada mapeamento. O processo de extração, transformação e carga dos dados do banco operacional para o DW levou aproximadamente quatro horas, considerando que os dados foram extraídos de um outro SGBD, o FoxPro. Já o processo de extração e carga do *Data Warehouse* para o *Data Mart* levou aproximadamente duas horas, considerando-se que os dados foram extraídos e carregados em objetos do mesmo SBGD. Essas atividades foram desenvolvidas em um Pentium III 600 MHz com 768 MB de memória e plataforma Windows 2000.

Após o encerramento da atividade de povoamento, constatou-se que os *tablespaces* destinados tanto para os objetos de dados, como para os índices do *Data Warehouse* e do *Data Mart*, ocupavam um total de aproximadamente 2 GB de espaço em disco rígido. Os objetos do *Data Mart* ficaram com as seguintes quantidades de tuplas: as tabelas fatos Dep_Fat, Medidas_Fat, Ponderal_Fat e Reproducao_Fat ficaram respectivamente com 429.717, 1.484.938, 417.617 e 479.315 tuplas e as dimensões Fazenda_Dim, Tempo_Dim, Animal_Dim, Medidas_Dim, Ponderal_Dim e Reprodução_Dim ficaram respectivamente com 241, 29.220, 485.445, 1.484.938, 417.617 e 481.207 tuplas.

Uma fase problemática do processo de povoamento ocorreu durante a tentativa de extração dos dados das tabelas FoxPro ponderal e reproduz. O processo de Conectividade Genérica usando ODBC (HSODBC) do Oracle começava a recuperar os dados e após recuperar algumas tuplas ele retornava uma mensagem de erro e parava. A solução para esse problema foi utilizar o Visual FoxPro para exportar as duas tabelas para um outro formato, o dBASE IV. Nesse outro formato

o HSODBC conseguiu extrair os dados das duas tabelas. Em compensação, o povoamento dos objetos do *Data Mart* à partir do DW, ocorreu sem problemas, pois os objetos de banco de dados de origem e destino pertencem ao mesmo SGBD. Em suma, extrair dados de SGBDs heterogênios não é uma tarefa trivial.

Uma observação importante a ser feita nesta etapa se refere ao prometido desempenho que se deveria obter com o projeto de esquemas estrela. Pois com a implementação das consultas OLAP, para a aplicação em questão, verificou-se que o desempenho é o mesmo se os dados forem acessados na camada de integração ou no *Data Mart*. Mas isso se deve ao fato das tabelas que foram desnormalizadas, no projeto dos esquemas estrela, possuírem poucas tuplas, o que não acarreta nenhum *overhead* em uma operação de junção que for executada sobre a camada de integração, onde as tabelas estão normalizadas.

Mas de qualquer forma, a implementação das consultas OLAP sobre os esquemas estrela ainda estava demorando vários minutos para serem executadas. A solução encontrada para se obter um melhor desempenho das consultas foi a criação de visões materializadas, nas quais as junções da tabela fato com suas dimensões foram previamente realizadas no momento de criação da visão. Dessa forma a única operação que é realizada no momento da execução da consulta são os cálculos das funções de agregação, sendo para isso necessário varrer a visão inteira. A desvantagem de se criar as visões materializadas se deve ao fato das mesmas duplicarem o espaço ocupado pelo *Data Mart*.

Neste projeto foram implementadas quatro visões materializadas, cada uma correspondente a uma estrela do esquema constelação da Figura 5.7. Na Figura 5.13 é apresentado o código SQL de criação da visão Dep_Mv. Nessa visão foi feita a junção da tabela fato Dep_Fat com as tabelas de dimensão Fazenda_Dim, Animal_Dim e Tempo_Dim.

Na Figura 5.14 é apresentado o código de criação da visão Medidas_Mv, onde foram juntadas a tabela fato Medidas_Fat e as tabelas de dimensão Fazenda_Dim, Animal_Dim, Tempo_Dim e Medidas_Dim. Na Figura 5.15 é apresentado o código da visão Ponderal_Mv, onde foram juntadas a tabela fato Ponderal_Fat e as tabelas de dimensão Fazenda_Dim, Animal_Dim, Ponderal_Dim e Tempo_Dim.

E, por fim, na Figura 5.16 é ilustrada a visão Reproducao_Mv, onde são juntadas a tabela fato Reproducao_Fat e as dimensões Fazenda_Dim, Animal_Dim, Tempo_Dim e Reproducao_Dim.

Outro aspecto desta etapa de implementação que deve ser observado é a utilização dos índices. Na etapa de povoamento do DW e do *Data Mart*, a criação

```
CREATE MATERIALIZED VIEW "DEP_MV"
  TABLESPACE "GEMAC_DATA"
  NOPARALLEL
  NOLOGGING
  BUILD IMMEDIATE REFRESH COMPLETE ENABLE QUERY REWRITE
  AS SELECT D.ID_DEP, A.ID_ANIMAL, A.CGA,
            A.SEXO, A.RACA, A.CATEGORIA, A.OCO_PARTO, A.SIT_NASC, A.SITUACAO,
            F.ID_FAZENDA, F.CODIGO CODIGO_FAZ, F.PESSOA, F.PESAGEM_AO_NASCER,
            F.ESTADO, F.MUNICIPIO, F.CRIADOR_NOME, F.FAZENDA_NOME, F.TEC_RESP,
            EXTRACT(YEAR FROM D.DA_ACTUAL_DATE) ANO,
            EXTRACT(MONTH FROM D.DA_ACTUAL_DATE) MES,
            EXTRACT(DAY FROM D.DA_ACTUAL_DATE) DIA,
            D.F, D.DMPP120, D.AMPP120, D.DMPP240, D.AMPP240, D.DDPP120, D.ADPP120,
            D.DDPP240, D.ADPP240, D.DDPP365, D.ADPP365, D.DDPP455, D.ADPP455,
            D.DDPP550, D.ADPP550, D.DDPE365, D.ADPE365, D.DDPE455, D.ADPE455,
            D.DDPE550, D.ADPE550, D.MGT, D.NR120, D.NF120, D.NR455, D.NF455
     FROM   FAZENDA_DIM F, ANIMAL_DIM A, DEP_FAT D
     WHERE  F.ID_FAZENDA = D.ID_FAZENDA AND A.ID_ANIMAL = D.ID_ANIMAL;
```

Figura 5.13: Script de criação da visão materializada Dep_Mv

```
CREATE MATERIALIZED VIEW "MEDIDAS_MV"
  TABLESPACE "GEMAC_DATA"
  NOPARALLEL
  NOLOGGING
  BUILD IMMEDIATE REFRESH COMPLETE ENABLE QUERY REWRITE
  AS SELECT MF.MEDIDAS_CGA, MF.MEDIDAS_DT_PESO,
            MD.MANEJO, MD.SITUACAO,
            A.ID_ANIMAL, A.SEXO, A.RACA, A.CATEGORIA, A.OCO_PARTO, A.SITUACAO, A.SIT_NASC,
            F.ID_FAZENDA, F.CODIGO CODIGO_FAZ, F.PESSOA, F.PESAGEM_AO_NASCER,
            F.ESTADO, F.MUNICIPIO, F.CRIADOR_NOME, F.FAZENDA_NOME, F.TEC_RESP,
            EXTRACT(YEAR FROM MF.MEDIDAS_DT_PESO) ANO_PESO,
            EXTRACT(MONTH FROM MF.MEDIDAS_DT_PESO) MES_PESO,
            EXTRACT(DAY FROM MF.MEDIDAS_DT_PESO) DIA_PESO,
            MF.PESO, MF.CE
     FROM   FAZENDA_DIM F, ANIMAL_DIM A, MEDIDAS_DIM MD, MEDIDAS_FAT MF
     WHERE  F.ID_FAZENDA = MF.ID_FAZENDA AND A.ID_ANIMAL = MF.ID_ANIMAL AND
            MD.MEDIDAS_CGA = MF.MEDIDAS_CGA AND MD.MEDIDAS_DT_PESO = MF.MEDIDAS_DT_PESO;
```

Figura 5.14: Script de criação da visão materializada Medidas_Mv

dos índices corretos nas tabelas do DW foi fundamental para se obter um bom desempenho no processo de extração, transformação e carga dos dados. O fato é que algumas funções de transformação fazem pesquisas em outras tabelas para cada tupla que é extraída de uma determinada tabela, ou seja, elas são executadas exaustivamente. Logo, alguns atributos utilizados por essas funções precisam ser corretamente indexados para que elas executem de forma a proporcionar um bom desempenho. Por outro lado, não houve a necessidade de se criar índices sobre as visões, uma vez que as mesmas são varridas completamente durante a execução de uma consulta.

```
CREATE MATERIALIZED VIEW "PONDERAL_MV"
  TABLESPACE "GEMAC_DATA"
  NOPARALLEL
  NOLOGGING
  BUILD IMMEDIATE REFRESH COMPLETE ENABLE QUERY REWRITE
  AS SELECT PF.ID_PONDERAL, A.ID_ANIMAL, A.CGA,
            A.SEXO, A.RACA, A.CATEGORIA, A.OCO_PARTO, A.SITUACAO, A.SIT_NASC,
            F.ID_FAZENDA, F.CODIGO CODIGO_FAZ, F.PESSOA, F.PESAGEM_AO_NASCER,
            F.ESTADO, F.MUNICIPIO, F.CRIADOR_NOME, F.FAZENDA_NOME, F.TEC_RESP,
            PD.MPBD, PD.MPVD, PD.M120, PD.M240, PD.M365, PD.M455, PD.M550, PD.M730,
            EXTRACT(YEAR FROM PF.DA_ACTUAL_DT_DESM) ANO_DESM,
            EXTRACT(MONTH FROM PF.DA_ACTUAL_DT_DESM) MES_DESM,
            EXTRACT(DAY FROM PF.DA_ACTUAL_DT_DESM) DIA_DESM,
            EXTRACT(YEAR FROM PF.DA_ACTUAL_DT_UAP) ANO_UAP,
            EXTRACT(MONTH FROM PF.DA_ACTUAL_DT_UAP) MES_UAP,
            EXTRACT(DAY FROM PF.DA_ACTUAL_DT_UAP) DIA_UAP,
            EXTRACT(YEAR FROM PF.DA_ACTUAL_DT_UPP) ANO_UPP,
            EXTRACT(MONTH FROM PF.DA_ACTUAL_DT_UPP) MES_UPP,
            EXTRACT(DIA FROM PF.DA_ACTUAL_DT_UPP) DIA_UPP,
            PF.PN, PF.PBD, PF.PVD, PF.P120, PF.P240, PF.P365, PF.C365,
            PF.P455, PF.C455, PF.P550, PF.C550, PF.P730, PF.C730
     FROM   FAZENDA_DIM F, ANIMAL_DIM A, PONDERAL_DIM PD, PONDERAL_FAT PF
     WHERE  PD.ID_PONDERAL = PF.ID_PONDERAL AND A.ID_ANIMAL = PF.ID_ANIMAL AND
            F.ID_FAZENDA = PF.ID_FAZENDA;
```

Figura 5.15: Script de criação da visão materializada Ponderal_Mv

```
CREATE MATERIALIZED VIEW "REPRODUCAO_MV"
  TABLESPACE "GEMAC_DATA"
  NOPARALLEL
  NOLOGGING
  BUILD IMMEDIATE REFRESH COMPLETE ENABLE QUERY REWRITE
  AS SELECT RF.REPRODUCAO_NAR, RD.TIPO_ACAS, RD.MANEJO,
            A1.ID_ANIMAL ID_ANIMAL_MAE, A1.SEXO SEXO_MAE, A1.RACA RACA_MAE, A1.CATEGORIA CAT_MAE,
            A1.OCO_PARTO OCO_PARTO_MAE, A1.SITUACAO SITUACAO_MAE, A1.SIT_NASC SIT_NASC_MAE,
            A2.ID_ANIMAL ID_ANIMAL_PAI, A2.SEXO SEXO_PAI, A2.RACA RACA_PAI, A2.CATEGORIA CAT_PAI,
            A2.OCO_PARTO OCO_PARTO_PAI, A2.SITUACAO SITUACAO_PAI, A2.SIT_NASC SIT_NASC_PAI,
            A3.ID_ANIMAL ID_ANIMAL_FILHO, A3.SEXO SEXO_FILHO,
            A3.RACA RACA_FILHO, A3.CATEGORIA CAT_FILHO,
            A3.OCO_PARTO OCO_PARTO_FILHO, A3.SITUACAO SITUACAO_FILHO, A3.SIT_NASC SIT_NASC_FILHO,
            F.ID_FAZENDA, F.CODIGO CODIGO_FAZ, F.PESSOA, F.PESAGEM_AO_NASCER,
            F.ESTADO, F.MUNICIPIO, F.CRIADOR_NOME, F.FAZENDA_NOME, F.TEC_RESP,
            EXTRACT(YEAR FROM RF.DA_ACTUAL_DATE_ACAS) ANO_ACAS,
            EXTRACT(MONTH FROM RF.DA_ACTUAL_DATE_ACAS) MES_ACAS,
            EXTRACT(DAY FROM RF.DA_ACTUAL_DATE_ACAS) DIA_ACAS,
            EXTRACT(YEAR FROM RF.DA_ACTUAL_DATE_NASC) ANO_NASC,
            EXTRACT(MONTH FROM RF.DA_ACTUAL_DATE_NASC) MES_NASC,
            EXTRACT(DAY FROM RF.DA_ACTUAL_DATE_NASC) DIA_NASC,
            RF.NC, RF.PVP, RF.IVP, RF.IEP, RF.PG, RF.PS
     FROM   FAZENDA_DIM F, ANIMAL_DIM A1, ANIMAL_DIM A2, ANIMAL_DIM A3,
            REPRODUCAO_DIM RD, REPRODUCAO_FAT RF
     WHERE  RD.REPRODUCAO_NAR = RF.REPRODUCAO_NAR AND A1.ID_ANIMAL = RF.ID_ANIMAL_MAE AND
            A2.ID_ANIMAL = RF.ID_ANIMAL_PAI AND A3.ID_ANIMAL = RF.ID_ANIMAL_FILHO AND
            F.ID_FAZENDA = RF.ID_FAZENDA;
```

Figura 5.16: Script de criação da visão materializada Reproducao_Mv

5.6 Revisão

Esta é a última fase da metodologia adotada e objetiva verificar os resultados obtidos com a implementação do sistema. Mas como este ambiente analítico ainda está em fase de implantação e avaliação, espera-se que ele alcance os objetivos propostos, de auxiliar nas pesquisas relacionadas ao melhoramento genético da raça Nelore.

5.7 Considerações Finais

A metodologia adotada neste trabalho para o desenvolvimento do DW é composta por seis etapas: Justificativa, Planejamento, Análise, Projeto, Implementação e Revisão. A mesma foi instanciada neste capítulo para solucionar o problema em questão. A primeira etapa é feita apenas uma vez e as outras são repetidas para cada *Data Mart* que é adicionado ao DW. Portanto, este capítulo pode servir como guia para novos *Data Marts* que forem implementados.

Pela prática adquirida com a utilização desta metodologia para a construção do *Data Mart* do PMGRN, pode-se afirmar que a mesma é adequada para a implementação de sistemas de *Data Warehousing*. O próximo capítulo apresentará a utilização do DW construído. Nele serão abordados a implementação das consultas OLAP e a aplicação do processo de *Data Mining* Visual, ou seja, serão apresentadas duas formas diferentes de se analisar os dados do DW.

CAPÍTULO

6

Analisando os Dados do *Data Warehouse*

No capítulo anterior foi apresentado o processo de desenvolvimento do *Data Warehouse* para o PMGRN. Esse repositório de dados, por si só, não tem utilidade alguma se o seu conteúdo não for analisado com ferramentas adequadas. O DW construído possui características direcionadas à oferecer dados com rapidez e qualidade para as análises de seus usuários.

Com o objetivo de apresentar as formas nas quais os especialistas do PMGRN poderão analisar os dados do DW, este capítulo foi estruturado da seguinte maneira: na Seção 6.1 são apresentadas as consultas implementadas com a ferramenta OLAP Oracle Discoverer, na Seção 6.2 são apresentadas as análises realizadas com a ferramenta de *Data Mining* Visual Spotfire e na Seção 6.3 são feitas algumas considerações finais sobre este capítulo.

6.1 OLAP

O Oracle Discoverer, apresentado na Seção 4.3, é uma ferramenta que proporciona aos usuários uma maneira fácil de manipular as consultas OLAP.

O primeiro passo realizado para a implementação das consultas OLAP foi a definição, no Oracle Discoverer Administration Edition, da Área de Negócios, Pastas, Itens e Hierarquias.

Na Figura 6.1 são apresentados os dados e hierarquias definidos na ferramenta. Foi definida a Área de Negócios PMGRN-DM, as Pastas correspondentes às visões materializadas, e os itens de cada Pasta. Foram definidas também as hierarquias

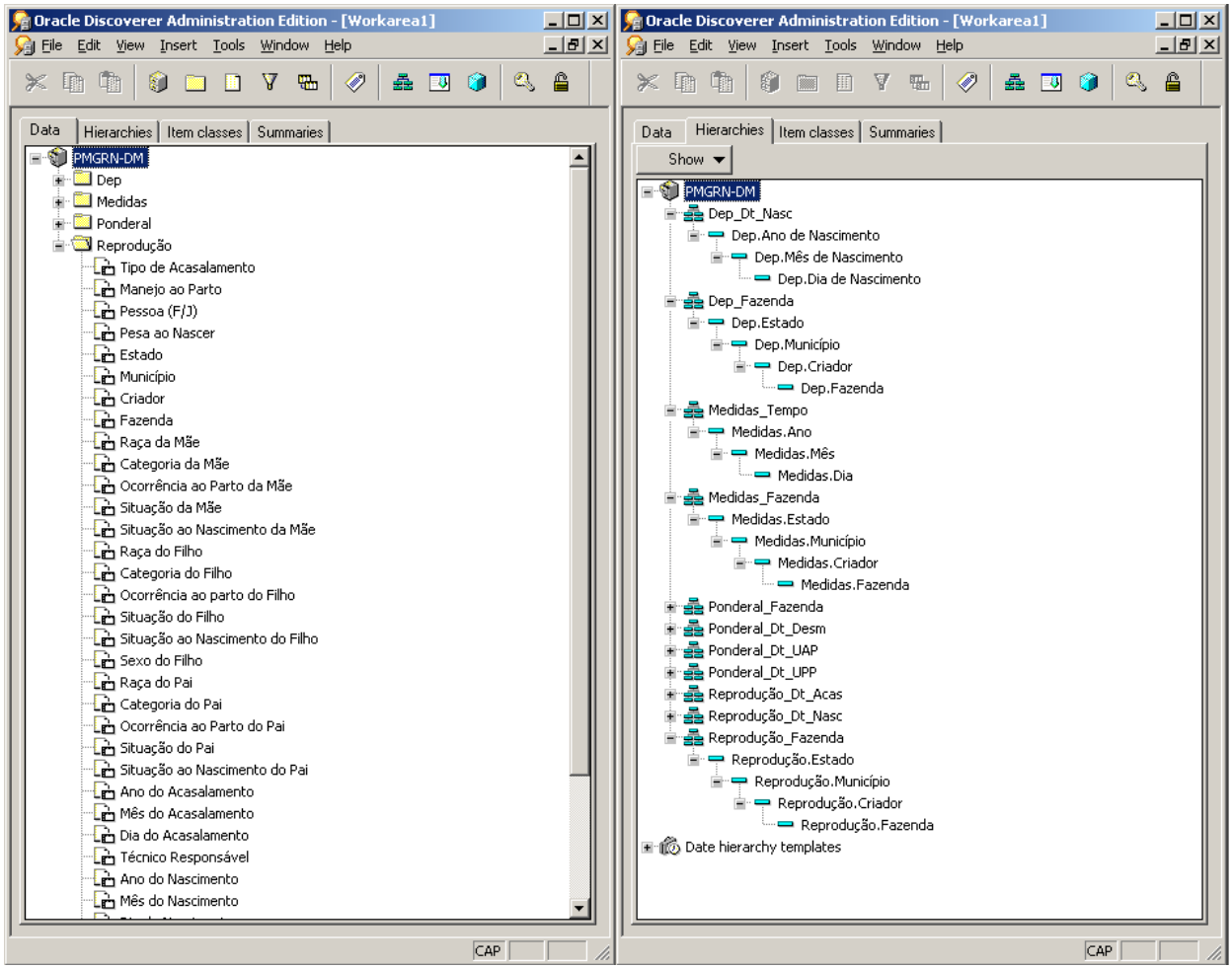


Figura 6.1: Definição dos dados e das hierarquias no Oracle Discoverer Administration Edition.

das visões materializadas. Todas essas definições foram feitas sob o *Data Mart* do PMGRN.

A implementação das consultas OLAP, definidas *a priori*, foram divididas em grupos. Como foram implementados quatro esquemas estrelas no *Data Mart* do PMGRN, cada grupo de consultas corresponde a um esquema estrela. Foram criados os grupos DEPs, Medidas, Ponderal e Reprodução.

Na Tabela 6.1 é apresentada a quantidade de cadernos e folhas que foram implementadas para cada grupo de consultas. No total foram criadas oitenta e sete folhas, sendo que cada folha possui operações agregadas aplicadas sobre um e no máximo dois atributos. Vale lembrar que essas operações agregadas foram aplicadas sobre as medidas que compõem as tabelas fatos.

As Tabelas 6.2 e 6.3 apresentam para cada caderno de DEPs e para cada

Grupos	Cadernos	Folhas
DEPs	11	34
Medidas	02	08
Ponderal	05	21
Reprodução	06	24

Tabela 6.1: Quantidade de Cadernos e Folhas dos Grupos de consultas OLAP.

Caderno	Folha	Atributos	Operações
1	01	MGT	N, Média, Mínimo, Máximo e Desvio Padrão
1	02	F	N, Média, Mínimo, Máximo e Desvio Padrão
1	03	NR455 e NF455	N, Média, Mínimo, Máximo e Desvio Padrão
2	04	DMPP120 e AMPP120	N, Média, Mínimo, Máximo e Desvio Padrão
2	05	DMPP240 e AMPP240	N, Média, Mínimo, Máximo e Desvio Padrão
3	06	DDPP120 e ADPP120	N, Média, Mínimo, Máximo e Desvio Padrão
3	07	DDPP240 e ADPP240	N, Média, Mínimo, Máximo e Desvio Padrão
4	08	DDPP365 e ADPP365	N, Média, Mínimo, Máximo e Desvio Padrão
4	09	DDPP455 e ADPP455	N, Média, Mínimo, Máximo e Desvio Padrão
4	10	DDPP550 e ADPP550	N, Média, Mínimo, Máximo e Desvio Padrão
5	11	DDPE365 e ADPE365	N, Média, Mínimo, Máximo e Desvio Padrão
5	12	DDPE455 e ADPE455	N, Média, Mínimo, Máximo e Desvio Padrão
5	13	DDPE550 e ADPE550	N, Média, Mínimo, Máximo e Desvio Padrão

Tabela 6.2: Operações aplicadas sobre os atributos de DEPs.

folha, os atributos e as operações utilizadas. Enquanto nas folhas da Tabela 6.2 procurou-se aplicar várias operações diferentes sobre um ou dois atributos comuns, nas folhas da Tabela 6.3 procurou-se aplicar uma única operação sobre vários atributos que possuem em comum uma seqüencia temporal, como por exemplo, a DEP Direta de Perímetro Escrotal aos 365, 455 e 550 dias da folha 22. Além dos fatos apresentados, as consultas OLAP de DEPs também são compostas por atributos analíticos das dimensões Fazenda, Animal e Tempo. Os tipos de operações foram aplicadas levando-se em conta a natureza do atributo.

Por questão de desempenho e utilização de memória, as consultas OLAP do grupo de Medidas foram estruturadas de uma maneira diferente. Foram criados dois cadernos de trabalho, um para medidas de Peso e outro para medidas de Perímetro Escrotal. Foram aplicadas as operações agregadas de Média, Mínimo, Máximo, Desvio Padrão e Quantidade (N) sobre ambas as medidas. No caderno de medidas de Peso foram criadas quatro folhas, uma com os atributos analíticos das dimensões Medidas, Animal e Tempo, outra com os atributos analíticos das dimensões Medidas, Animal e Fazenda, uma terceira folha com os atributos das dimensões Medidas, Fazenda e Tempo e uma quarta com os atributos das dimensões Medidas, Animal e Fazenda. Nessa quarta folha está sendo enfatizada

Caderno	Folha	Atributos	Operações
6	14	DDPP120, DDPP240, DDPP365 DDPP455 e DDPP550	Média
6	15	DDPP120, DDPP240, DDPP365 DDPP455 e DDPP550	Mínimo
6	16	DDPP120, DDPP240, DDPP365 DDPP455 e DDPP550	Máximo
6	17	DDPP120, DDPP240, DDPP365 DDPP455 e DDPP550	Desvio Padrão
7	18	ADPP120, ADPP240, ADPP365 ADPP455 e ADPP550	Média
7	19	ADPP120, ADPP240, ADPP365 ADPP455 e ADPP550	Mínimo
7	20	ADPP120, ADPP240, ADPP365 ADPP455 e ADPP550	Máximo
7	21	ADPP120, ADPP240, ADPP365 ADPP455 e ADPP550	Desvio Padrão
8	22	DDPE365, DDPE455 e DDPE550	Média
8	23	DDPE365, DDPE455 e DDPE550	Mínimo
8	24	DDPE365, DDPE455 e DDPE550	Máximo
8	25	DDPE365, DDPE455 e DDPE550	Desvio Padrão
9	26	ADPE365, ADPE455 e ADPE550	Média
9	27	ADPE365, ADPE455 e ADPE550	Mínimo
9	28	ADPE365, ADPE455 e ADPE550	Máximo
9	29	ADPE365, ADPE455 e ADPE550	Desvio Padrão

Tabela 6.3: Operações aplicadas sobre os atributos de DEPs.

a análise do atributo Técnico Responsável da dimensão Fazenda. No caderno de medidas de Perímetro Escrotal também foram feitas as mesmas combinações de dimensões. Essa divisão foi feita para se obter uma boa combinação dos atributos de dimensão a serem analisados, levando-se em conta o desempenho e a utilização de memória, uma vez que a combinação das quatro dimensões, Medidas, Animal, Fazenda e Tempo em uma única folha torna a consulta lenta pelo fato da ferramenta precisar alocar uma grande quantidade de memória.

Caderno	Folha	Atributos	Operações
1	01	PN e PBD	Média, Mínimo, Máximo e Desvio Padrão
1	02	PVD	Média, Mínimo, Máximo e Desvio Padrão
1	03	P120 e P240	Média, Mínimo, Máximo e Desvio Padrão
2	04	P365 e P455	Média, Mínimo, Máximo e Desvio Padrão
2	05	P550 e P730	Média, Mínimo, Máximo e Desvio Padrão
3	06	C365 e C455	Média, Mínimo, Máximo e Desvio Padrão
3	07	C550 e C730	Média, Mínimo, Máximo e Desvio Padrão

Tabela 6.4: Operações aplicadas sobre os atributos Ponderais.

As Tabelas 6.4 e 6.5 apresentam para cada caderno de Ponderais e para cada

Caderno	Folha	Atributos	Operações
4	08	P120, P240, P365, P455, P550 e P730	Média
4	09	P120, P240, P365, P455, P550 e P730	Mínimo
4	10	P120, P240, P365, P455, P550 e P730	Máximo
4	11	P120, P240, P365, P455, P550 e P730	Desvio Padrão
5	12	C365, C455, C550 e C730	Média
5	13	C365, C455, C550 e C730	Mínimo
5	14	C365, C455, C550 e C730	Máximo
5	15	C365, C455, C550 e C730	Desvio Padrão

Tabela 6.5: Operações aplicadas sobre os atributos Ponderais.

folha, os atributos e as operações utilizadas. A divisão dos cadernos e folhas dessas tabelas seguiu a mesma idéia das Tabelas 6.2 e 6.3, conforme apresentado anteriormente.

Também por questão de desempenho e utilização de memória, as consultas OLAP do grupo de Reprodução foram estruturadas diferentemente. Foram criados seis cadernos de trabalho, um para cada medida de Reprodução: NC, PVP, IVP, IEP, PG e PS. Foram aplicadas as operações agregadas de Média, Mínimo, Máximo, Desvio Padrão e Quantidade (N) sobre ambas as medidas, além da operação de soma sobre o atributo NC.

Em todos os cadernos das medidas de Reprodução foram criadas quatro folhas, uma com os atributos analíticos das dimensões Reproduz, Animal e Fazenda, outra com os atributos analíticos das dimensões Reproduz, Animal e Tempo (onde é considerada a data de acasalamento), uma terceira folha com os atributos das dimensões Reproduz, Animal e Tempo (onde é considerada a data de nascimento) e uma quarta com os atributos das dimensões Reproduz, Animal e Fazenda (onde é considerado o Técnico Responsável). Essa divisão foi feita para se obter uma boa combinação dos atributos de dimensão a serem analisados, levando-se em conta o desempenho e a utilização de memória. Todas as consultas descritas anteriormente foram implementadas, mas como são muitas, serão apresentadas apenas algumas amostras.

Na Figura 6.2 é apresentada a consulta OLAP construída no Oracle Discoverer User Edition para a medida MGT. Todas as consultas sobre os atributos de DEPs possuem essa estrutura. Ela é composta pelos seguintes atributos analíticos: Ano e Mês da dimensão Tempo; Pessoa, Pesa ao Nascer, Estado, Município, Criador e Fazenda da dimensão Fazenda e Sexo, Raça, Categoria, Ocorrência ao Parto, Situação ao Nascimento e Situação da dimensão Animal.

Nessa consulta é possível realizar uma operação de *Slice* selecionando um

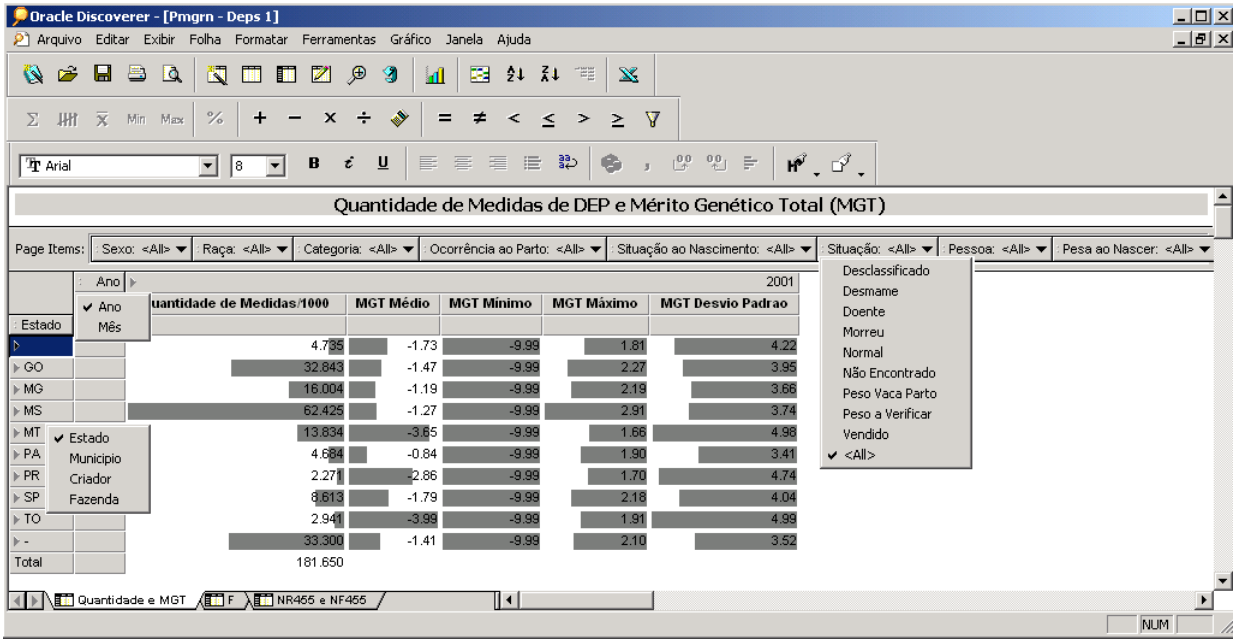


Figura 6.2: Consulta OLAP sobre a medida MGT do grupo de DEPs.

dos possíveis valores do atributo Situação (Desclassificado, Desmame, Doente, Morreu, ...) ou selecionando valores dos outros atributos que estão localizadas no *Page Items*. Pode-se também realizar operações de *Drill-down/up* sobre os atributos da hierarquia Estado, Município, Criador e Fazenda e também sobre os atributos da hierarquia Ano, Mês. Pode-se também executar operações de *Pivot* trocando de lugar os atributos Estado e Ano, ou ainda, trocando estes com os atributos do *Page Items*. Enfim, é possível fazer várias combinações de atributos e seus valores, a fim de se realizar diferentes tipos de análises.

Outro recurso que auxilia no processo de análise são os gráficos. Na Figura 6.3 é apresentado o gráfico correspondente à Figura 6.2. Pode-se observar que o Estado do MS é o que possui a maior quantidade de medidas de DEPs. Nas Figuras 6.4 e 6.5 são apresentados a consulta OLAP e o gráfico correspondente à consulta realizada sobre os atributos NR455 e NF455 do grupo de DEPs.

A Figura 6.6 apresenta uma outra consulta OLAP, sendo o foco dessa os valores médios para a DEP Direta de Peso Padronizado para Diferentes Dias. Nessa consulta foi realizada uma operação de *Pivot* entre os atributos Sexo e Estado e uma operação de *Slice* sobre o atributo Estado, onde é considerado apenas os Estado do PA. Analisando-se os valores médios do atributo Sexo para os diferentes Estados, verificou-se que o Estado do PA é o que apresenta a maior diferença entre os valores médios dos sexos fêmea e macho, ou seja, os valores médios da

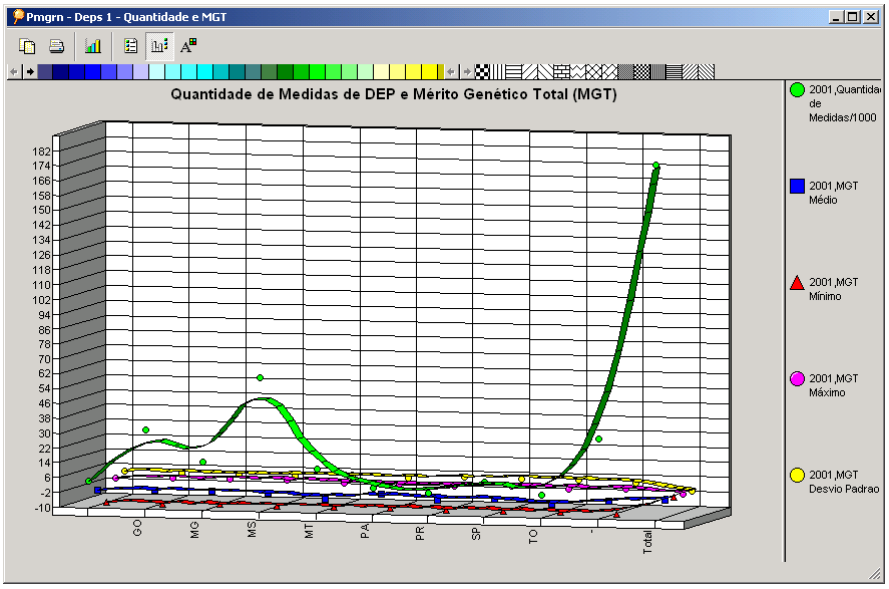


Figura 6.3: Gráfico da Consulta OLAP sobre a medida MGT do grupo de DEPs.

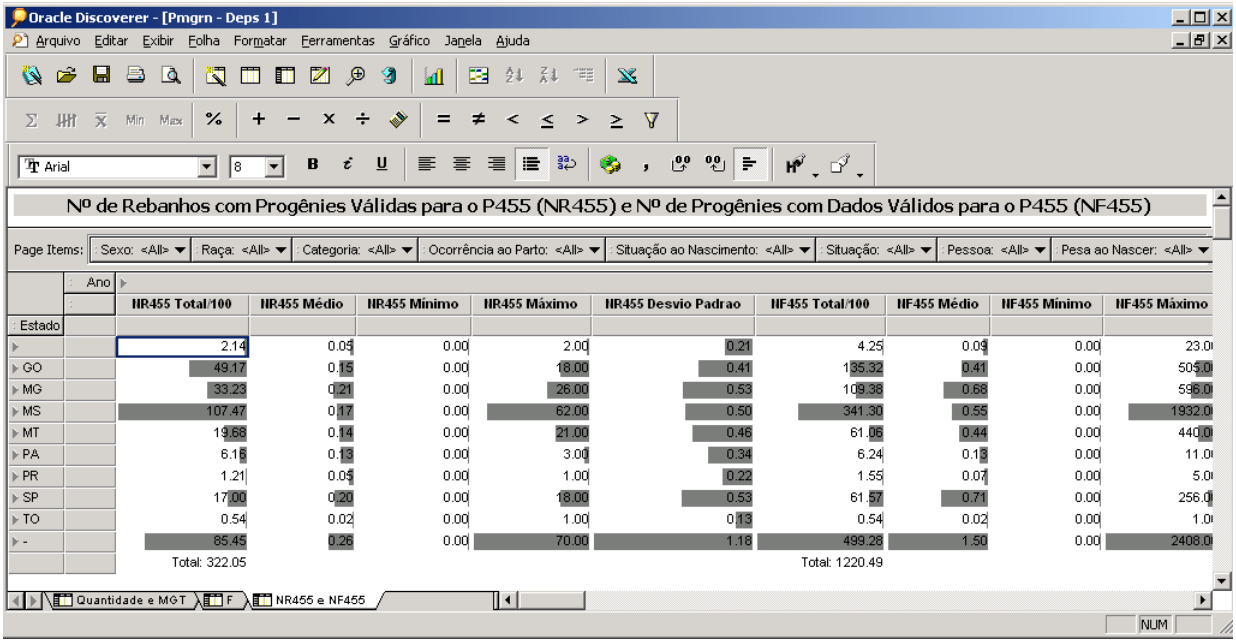


Figura 6.4: Consulta OLAP sobre a medida NR455 e NF455 do grupo de DEPs.

DEP Direta de Peso Padronizado dos machos é bem maior do que das fêmeas no Estado do PA em comparação com os outros estados.

As consultas OLAP dos atributos de DEPs levam aproximadamente 40 segundos para serem executadas, uma vez que essas consultas precisam resumir os dados da visão materializada Dep_Mv, que possui 429.717 tuplas.

São apresentados dois exemplos das seis consultas implementadas sobre as

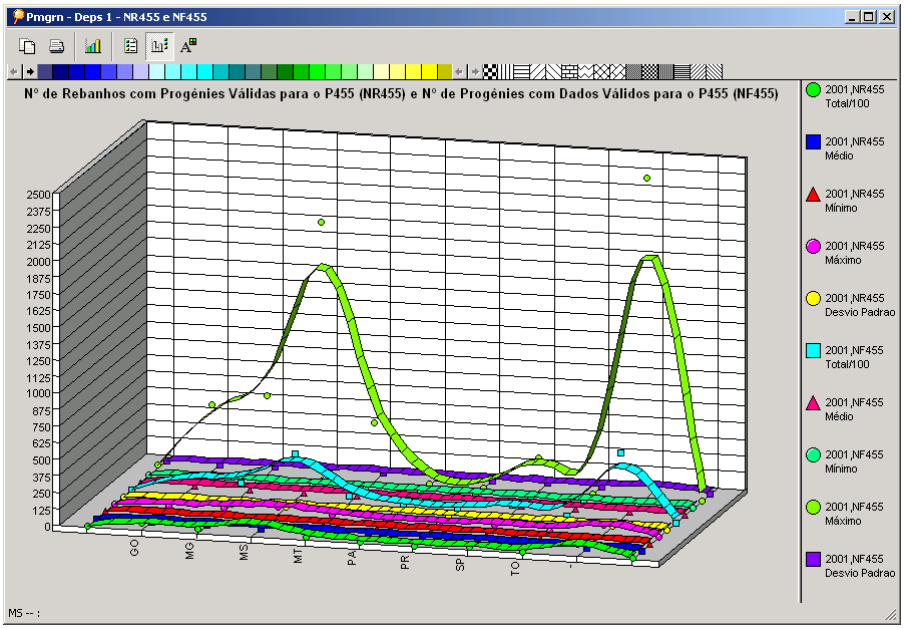


Figura 6.5: Gráfico da Consulta OLAP sobre a medida NR455 e NF455 do grupo de DEPs.

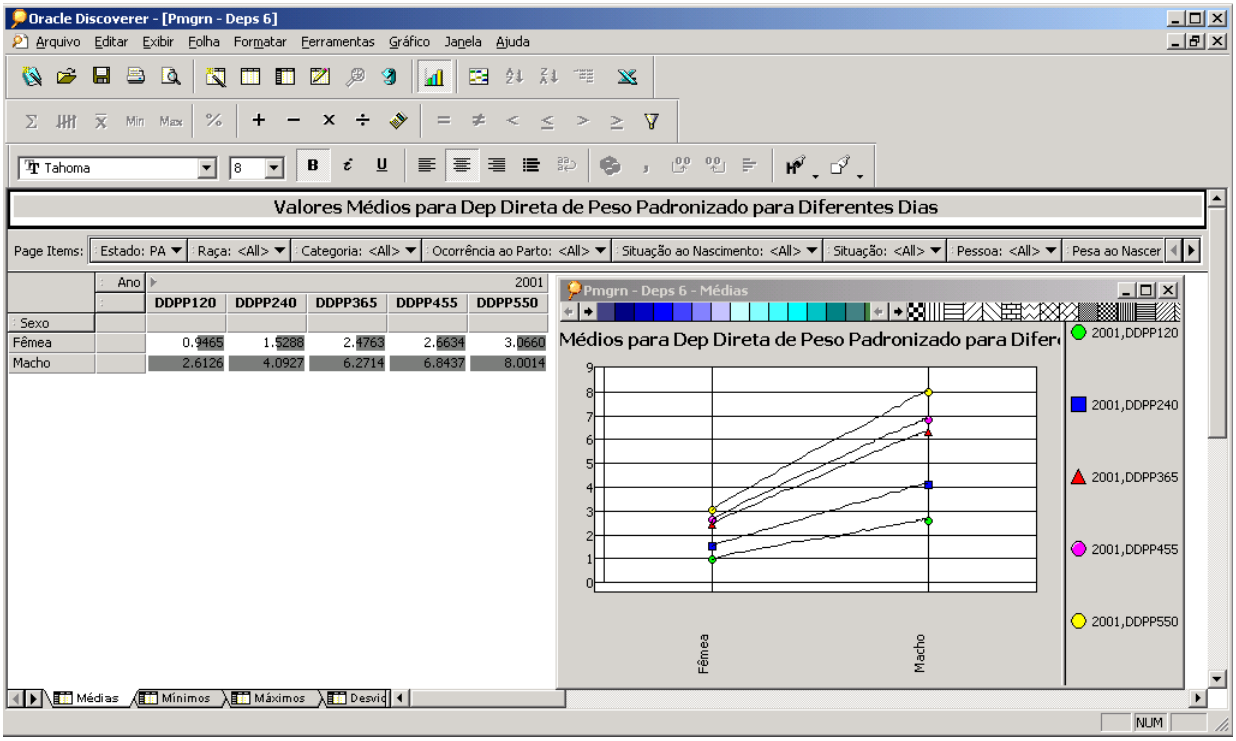


Figura 6.6: Consulta OLAP sobre as médias para DEP Direta de Peso Padronizado para Diferentes Dias.

medidas de Peso e Perímetro Escrotal. Na Figura 6.7 é apresentada uma das consultas OLAP sobre a medida Peso. Essa consulta tem como atributos analíti-

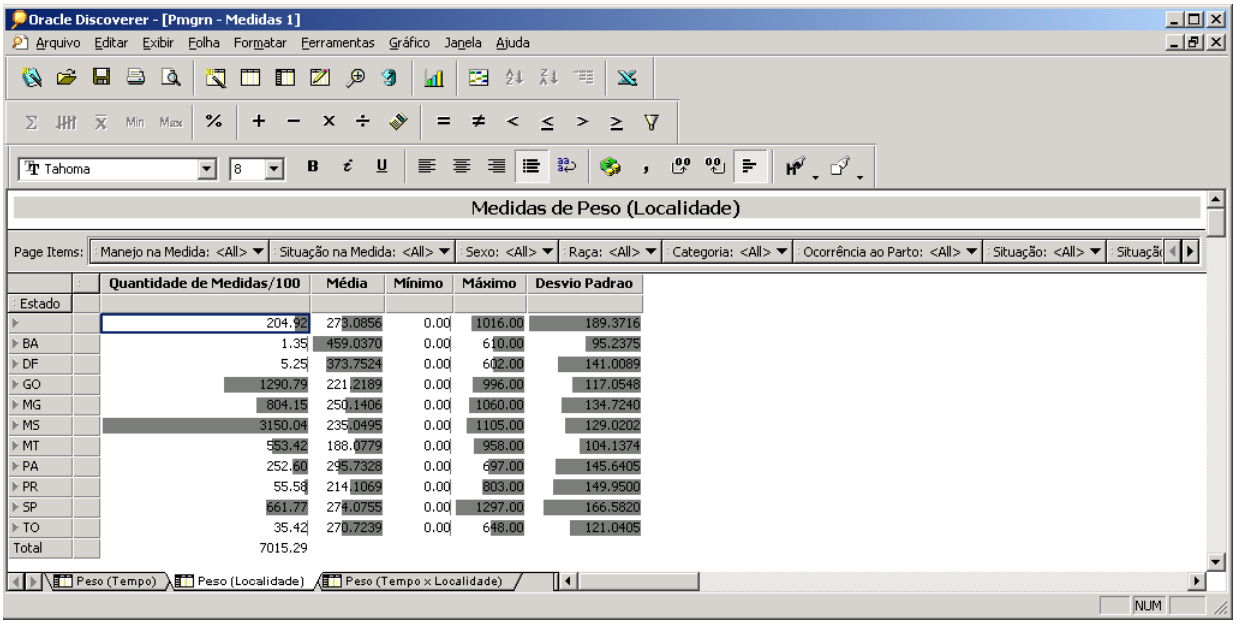


Figura 6.7: Consulta OLAP sobre a Medida Peso.

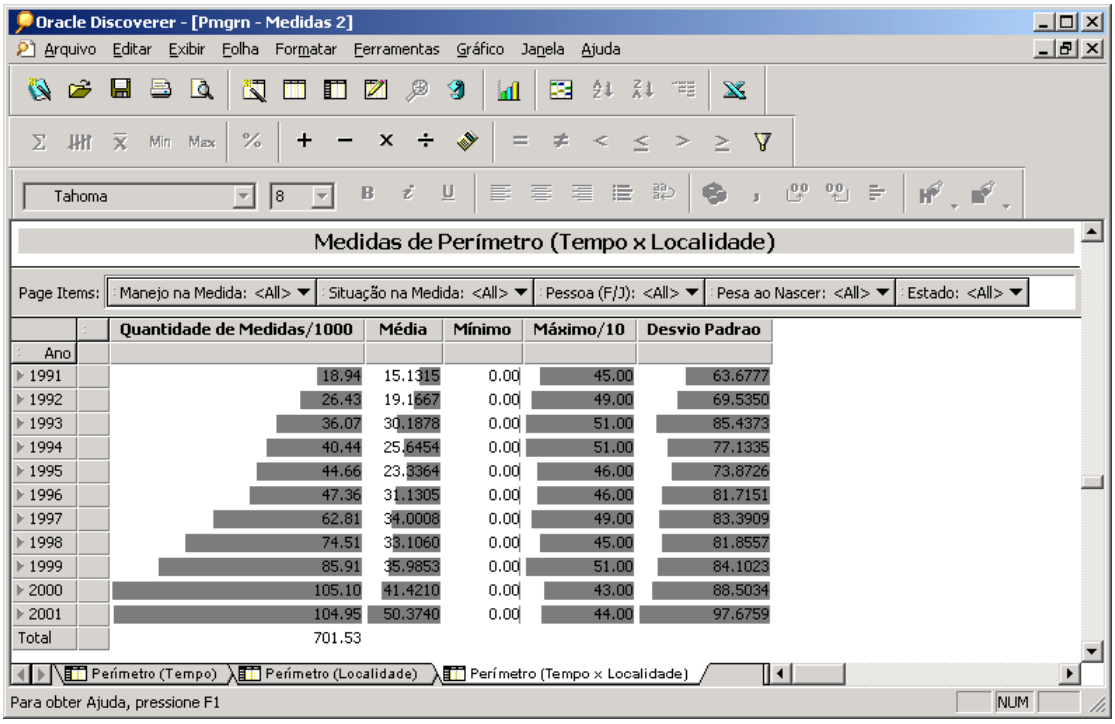


Figura 6.8: Consulta OLAP sobre a Medida Perímetro.

cos Manejo e Situação da dimensão Medida; os atributos Sexo, Raça, Categoria, Ocorrência ao Parto, Situação e Situação ao Nascimento da dimensão Animal e os atributos Estado, Município, Criador e Fazenda da dimensão Fazenda. Pode-se observar no resultado da consulta que o Estado de MS possui a maior quanti-

dade de medidas de Peso, o Estado da BA possui o maior Peso Médio e o Estado de SP possui o maior Peso.

Na Figura 6.8 é apresentada uma das consultas OLAP sobre a medida Perímetro Escrotal. Essa consulta têm como atributos analíticos Manejo e Situação da dimensão Medida; os atributos Pessoa, Pesa ao Nascer, Estado, Município, Criador e Fazenda da dimensão Fazenda e os atributos Ano e Mês da dimensão Tempo. Pode-se verificar por meio da consulta que a quantidade de medidas cresceu consideravelmente de 1960 a 2001, assim como, o valor médio das medidas de Perímetro.

As consultas OLAP dos atributos de Medidas levam aproximadamente 2:30 minutos para serem executadas, uma vez que elas precisam resumir os dados da visão materializada Medidas_Mv, que possui 1.484.938 tuplas.

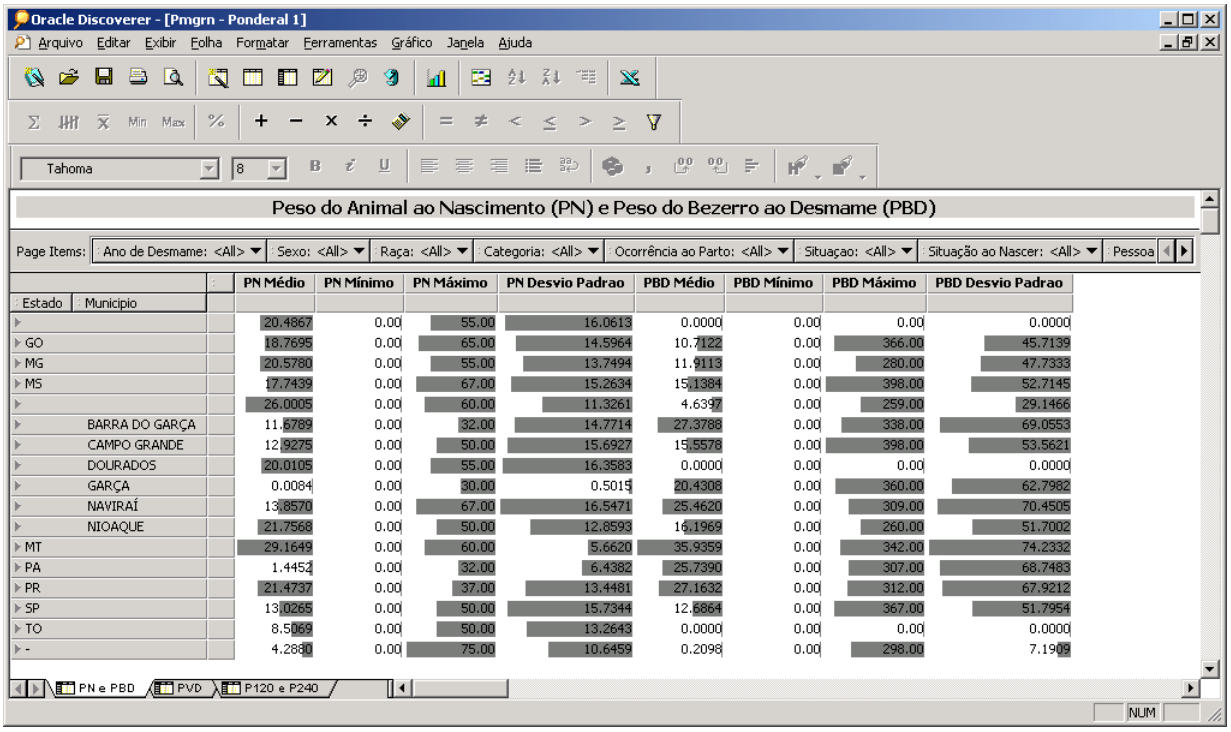
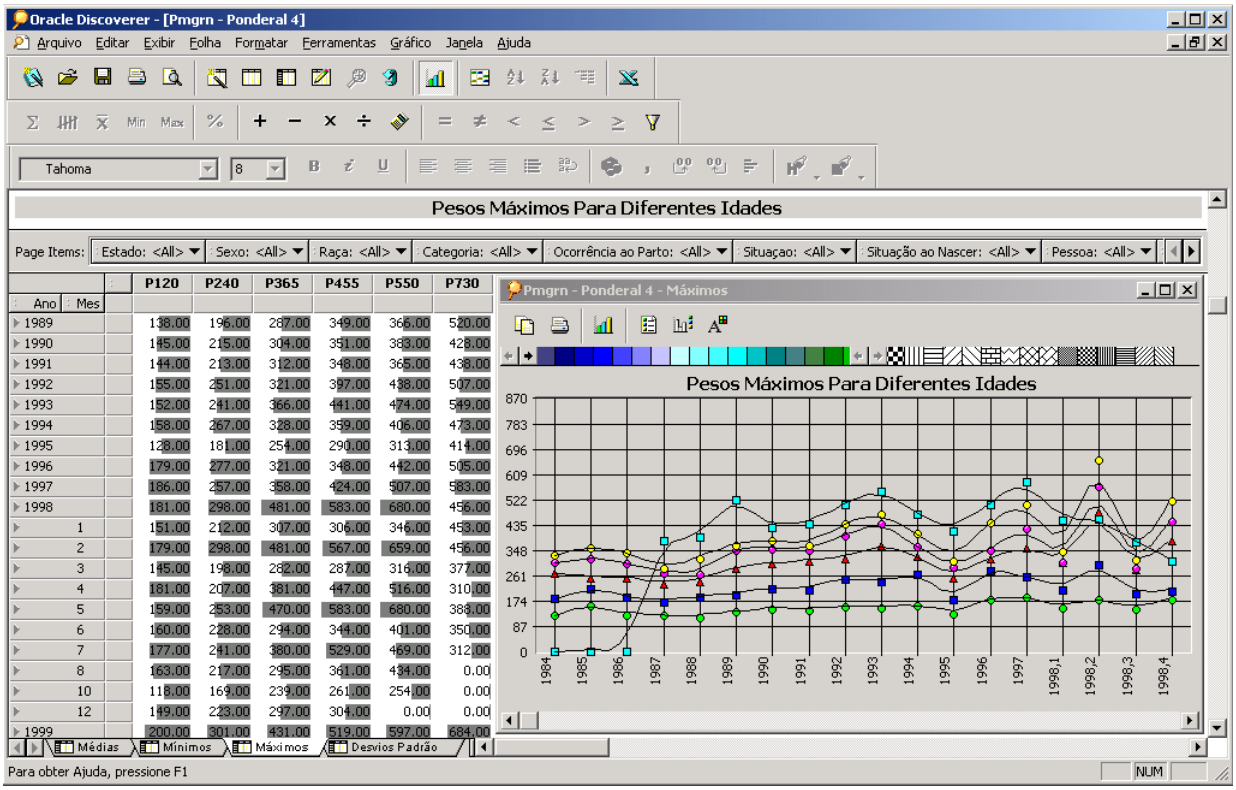


Figura 6.9: Consulta OLAP sobre as medidas Ponderais PN e PBD.

As Figuras 6.9 e 6.10 apresentam duas amostras de consultas OLAP implementadas para medidas Ponderais. Na Figura 6.9 é apresentada uma das consultas OLAP sobre as medidas Ponderais PN e PBD. Nessa consulta foi realizada uma operação de *drill-down* para o atributo Município, à partir do valor MS do atributo Estado. A consulta apresentada é composta pelos atributos analíticos das dimensões Fazenda, Tempo e Animal.



Ano / Mes	P120	P240	P365	P455	P550	P730
1989	138.00	196.00	287.00	349.00	366.00	520.00
1990	145.00	215.00	304.00	351.00	383.00	428.00
1991	144.00	213.00	312.00	348.00	365.00	438.00
1992	155.00	251.00	321.00	397.00	438.00	507.00
1993	152.00	241.00	366.00	441.00	474.00	549.00
1994	158.00	267.00	328.00	359.00	406.00	473.00
1995	128.00	181.00	254.00	290.00	313.00	414.00
1996	179.00	277.00	321.00	348.00	442.00	505.00
1997	186.00	257.00	358.00	424.00	507.00	583.00
1998	181.00	298.00	481.00	583.00	680.00	456.00
1	151.00	212.00	307.00	306.00	346.00	453.00
2	179.00	298.00	481.00	567.00	659.00	456.00
3	145.00	198.00	282.00	287.00	316.00	377.00
4	181.00	207.00	381.00	447.00	516.00	310.00
5	159.00	253.00	470.00	583.00	680.00	388.00
6	160.00	228.00	294.00	344.00	401.00	350.00
7	177.00	241.00	380.00	529.00	469.00	312.00
8	163.00	217.00	295.00	361.00	434.00	0.00
10	118.00	169.00	239.00	261.00	254.00	0.00
12	149.00	223.00	297.00	304.00	0.00	0.00
1999	200.00	301.00	431.00	519.00	597.00	684.00

Figura 6.10: Consulta OLAP sobre os pesos máximos ponderais para diferentes idades.

Já a consulta e o gráfico da Figura 6.10 objetivam apresentar valores agregados de pesos ponderais máximos para diferentes idades. Nessa consulta foi realizada uma operação de *Pivot* entre os atributos Estado e Ano. Também foi realizada uma operação de *drill-down* sobre o valor 1998 do atributo ano. Ela também é composta pelos atributos analíticos das dimensões Fazenda, Tempo e Animal.

As consultas OLAP dos atributos Ponderais levam aproximadamente 40 segundos para serem executadas, uma vez que elas precisam resumir os dados da visão materializada Ponderal_Mv, que possui 417.617 tuplas.

Para os atributos de Reprodução foram criadas 24 consultas OLAP. Na Figura 6.11 é apresentada uma dessas consultas. Como um registro de Reprodução possui um pai, uma mãe e um filho, a dimensão Animal é referenciada três vezes. Portanto, essa dimensão contribui com os atributos analíticos da Mãe, do Pai e do Filho. Outros atributos analíticos são Tipo de Acasalamento e Manejo ao Parto da dimensão Reprodução e a hierarquia Estado, Município, Criador e Fazenda da dimensão Fazenda.

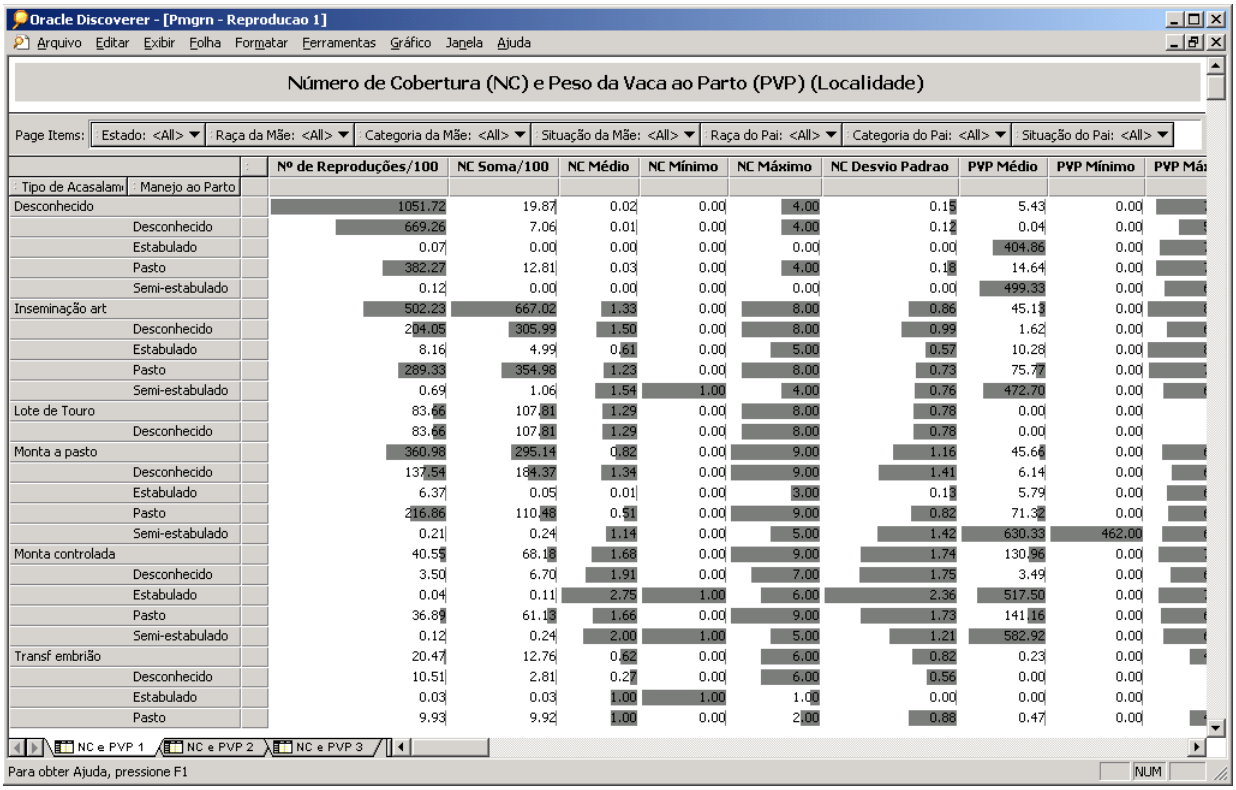


Figura 6.11: Consulta OLAP sobre as medidas de Reprodução.

As consultas realizadas sobre os atributos de Reprodução levam aproximadamente 40 segundos para serem executadas, uma vez que estas consultas resumem os dados da visão materializada Reproducao_Mv, que possui 479.315 tuplas.

Esta seção objetivou apresentar uma amostra das consultas OLAP que foram construídas *a priori* no ambiente desenvolvido neste trabalho. A próxima seção se focará em apresentar também algumas amostras dos tipos de análises que podem ser realizadas com a ferramenta de *Data Mining* Visual Spotfire descrita no Capítulo 4.

6.2 Data Mining

Para realizar análises visuais com o Spotfire foram feitas divisões semelhantes às da seção anterior. Foram considerados quatro grupos de dados: DEPs, Medidas, Ponderal e Reprodução, cada um correspondendo a um esquema estrela do *Data Mart* do PMGRN.

O Spotfire acessa os dados do DW armazenados no SGBD Oracle via ODBC ou OLEDB. O acesso via ODBC apresenta uma taxa de recuperação dos dados baixa. Logo, toda vez que o especialista for fazer suas análises, ele terá que es-

perar bastante tempo até que os dados sejam recuperados e os gráficos gerados. O acesso via OLEDB é bem mais eficiente, com esse tipo de acesso milhares de registros podem ser recuperados em poucos segundos. Sendo portanto essa a solução adotada neste trabalho.

Foram criados arquivos do tipo SFS (Spotfire Analysis Files), que armazenam as configurações dos gráficos feitas pelo analista, bem como, a consulta SQL que recupera os dados do SGBD. Quando esses arquivos são abertos é requerido o usuário Oracle, a senha e o nome do servidor, para que seja feita a conexão com o SGBD.

Um outro problema é a quantidade de registros que o Spotfire consegue manipular. Tentou-se de início analisar todos os dados da visão Reproducao_Mv, que possui 479.315 registros. Porém, o Spotfire apresentou um péssimo desempenho com essa quantidade de dados. Foram realizados vários testes com quantidades diferentes de registros e constatou-se que seria melhor particionar os dados em quantidades que variavam em torno de 150 mil registros. Pois com esta quantidade o Spotfire apresenta um melhor desempenho.

Para os dados de DEPs, Ponderais e Reprodução foi feita uma divisão por Estado. Foram criados três arquivos para cada grupo. Nas Tabelas 6.6, 6.7 e 6.8 são apresentados os arquivos criados e as respectivas quantidades de registros. Já para as Medidas foi feita a divisão dos dados por Ano. Foram criados nove arquivos, conforme apresentado na Tabela 6.9.

Estados	**Registros**
GO e MS	158.811
MG, MT, SP, BA, DF, MA, PA, PR, RO, TO, VE	97.146
Desconhecidos	97.146

Tabela 6.6: Divisão dos dados de DEPs por Estado.

Estados	**Registros**
MS e MT	144.381
SP e GO	133.176
BA, DF, MA, MG, PA, PR, RN, RO, TO, VE e Desconhecidos	140.060

Tabela 6.7: Divisão dos dados de Ponderais por Estado.

Na Figura 6.12 é apresentado o gráfico do arquivo SFS de DEPs cujo Estado é o de GO. No *menu popup* ao lado direito do gráfico são mostrados os atributos analíticos que podem ser escolhidos para compor as dimensões do gráfico. Para a construção do gráfico *Scatter Plot 3D* apresentado, foram escolhidos os atributos

Estados	Registros
GO e MS	176.200
MG, MT, SP, BA, DF, MA, PA, PR, RN, RO, TO, VE	191.325
Desconhecidos	111.790

Tabela 6.8: Divisão dos dados de Reprodução por Estado.

Ano	Registros	Ano	Registros	Ano	Registros
1960-1992	158.279	1993-1994	139.206	1995-1996	184.447
1997	114.173	1998	131.735	1999	147.207
2000	187.618	2001	221.243	2002	201.030

Tabela 6.9: Divisão dos dados de Medidas por Ano.

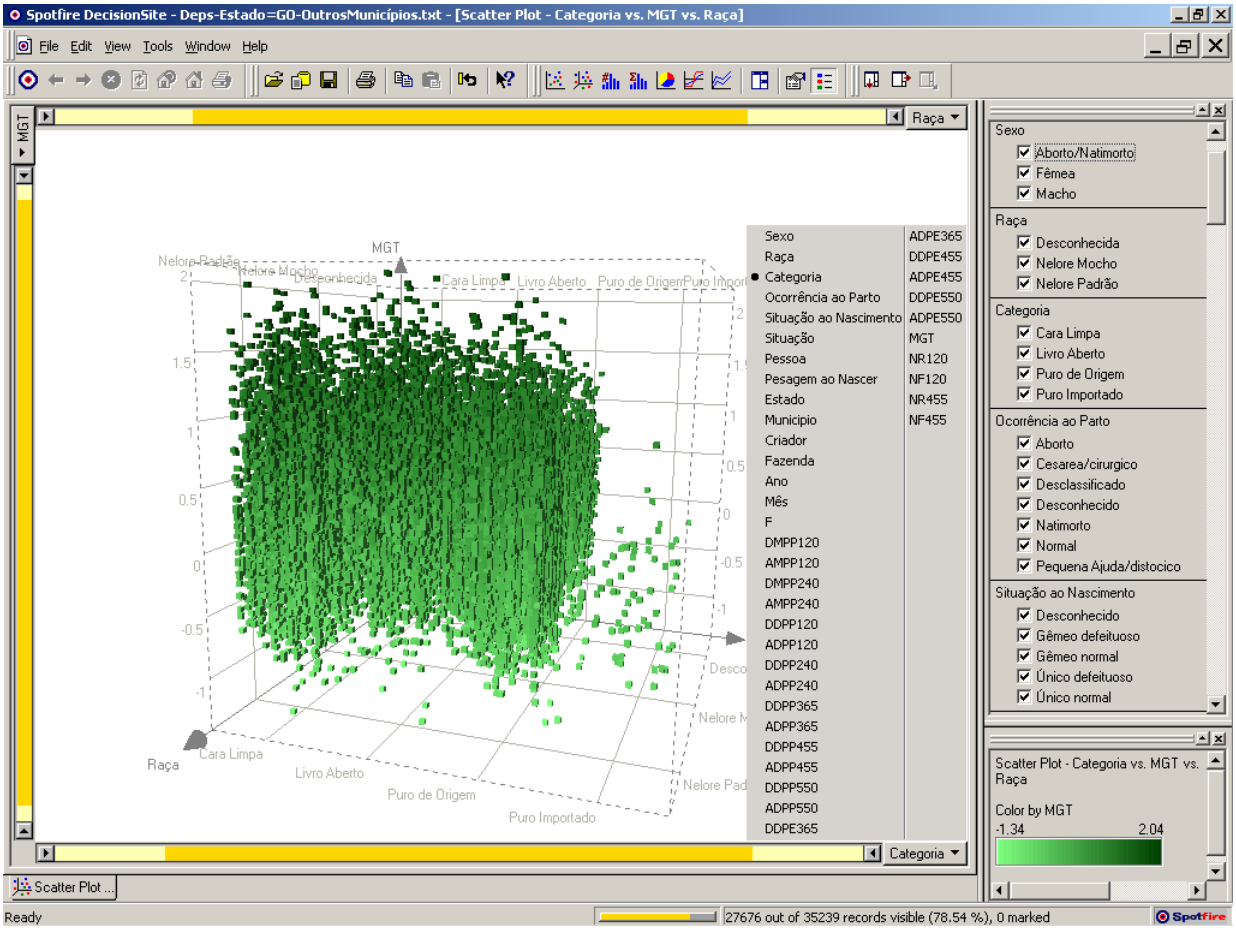


Figura 6.12: Gráfico *Scatter Plot 3D* dos atributos Categoria, Raça e MGT.

Raça, Categoria e MGT para as respectivas dimensões. Para o atributo MGT foi escolhido um verde claro para indicar valores baixos do MGT e verde escuro para indicar valores altos, conforme apresentado na legenda no canto inferior direito da figura. Pode-se observar no gráfico que a maioria dos valores de MGT estão

entre -1.35 e 2.04 e que a Categoria Puro Importado possui poucos valores de MGT.

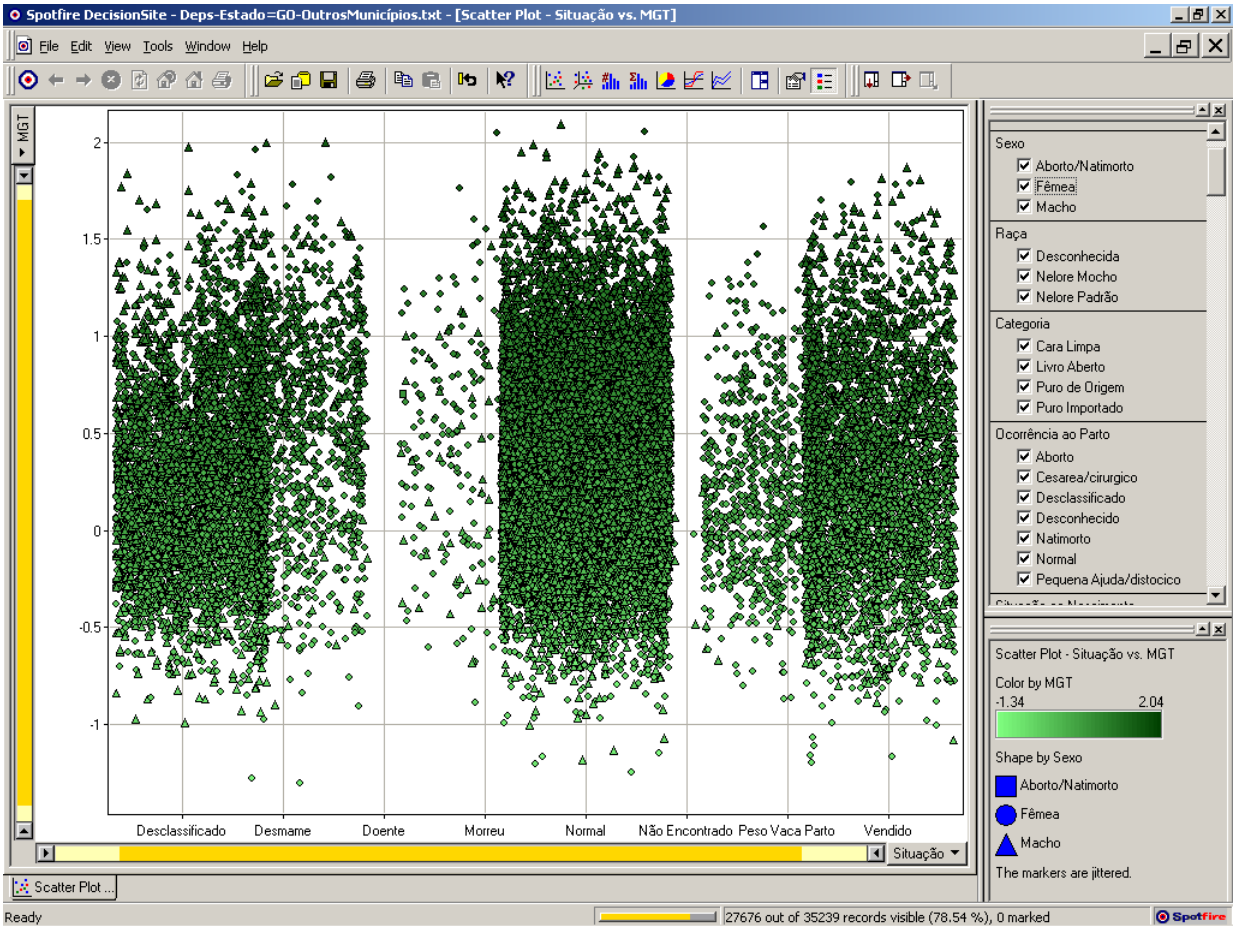


Figura 6.13: Gráfico *Scatter Plot 2D* dos atributos Município e MGT.

Na Figura 6.13 é apresentada uma versão em duas dimensões do gráfico *Scatter Plot*. O eixo horizontal representa o atributo Situação e o eixo vertical o atributo MGT. Pode-se observar no gráfico que a situação Normal concentra mais valores de MGT.

Os gráficos da Figura 6.14 também foram gerados à partir do arquivo de DEPs do Estado de GO e suas dimensões são representadas pelos atributos Sexo, Município e NF455. Em termos de valores os dois gráficos são idênticos, só possuindo uma renderização diferente. Pode-se observar nos gráficos que existem apenas três valores do NF455 acima de 100 e apenas um valor acima de 400. Os restantes dos valores estão abaixo de 100, ou seja, existem três valores anormais para o atributo NF455 nos dados de DEPs. Além disso pode-se observar que os registros estão mais concentrados na intersecção dos Municípios Goiânia e São

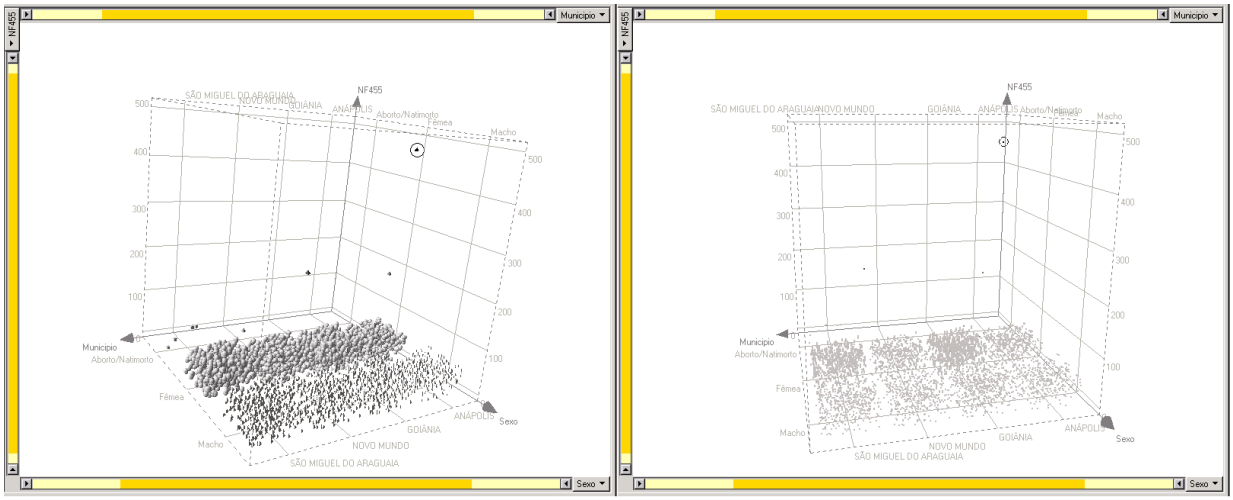


Figura 6.14: Gráfico *Scatter Plot 3D* dos atributos Sexo, Município e NF455.

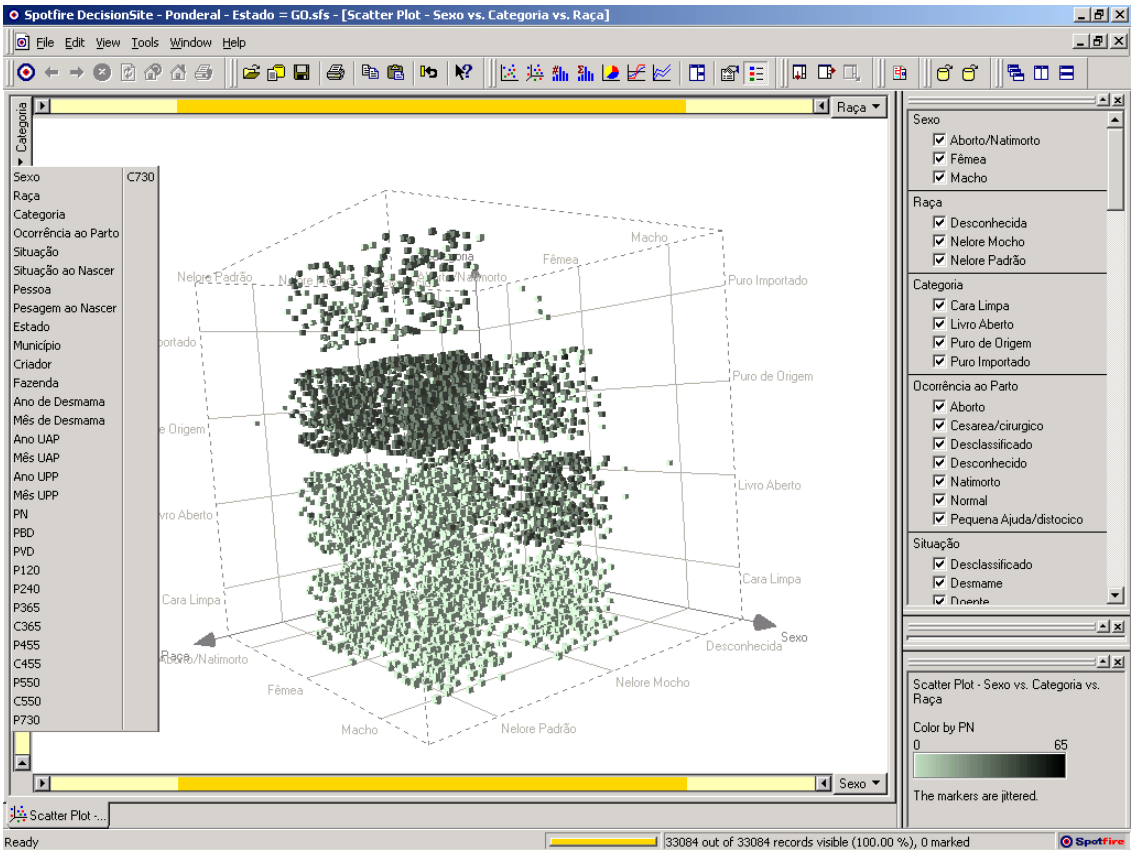


Figura 6.15: Gráfico *Scatter Plot 3D* dos atributos Sexo, Raça, Categoria e PN.

Miguel do Araguaia com o Sexo Fêmea.

Na Figura 6.15 é apresentado um gráfico do arquivo SFS de Ponderais cujo Estado é o de GO. No *menu popup* ao lado esquerdo do gráfico são mostrados

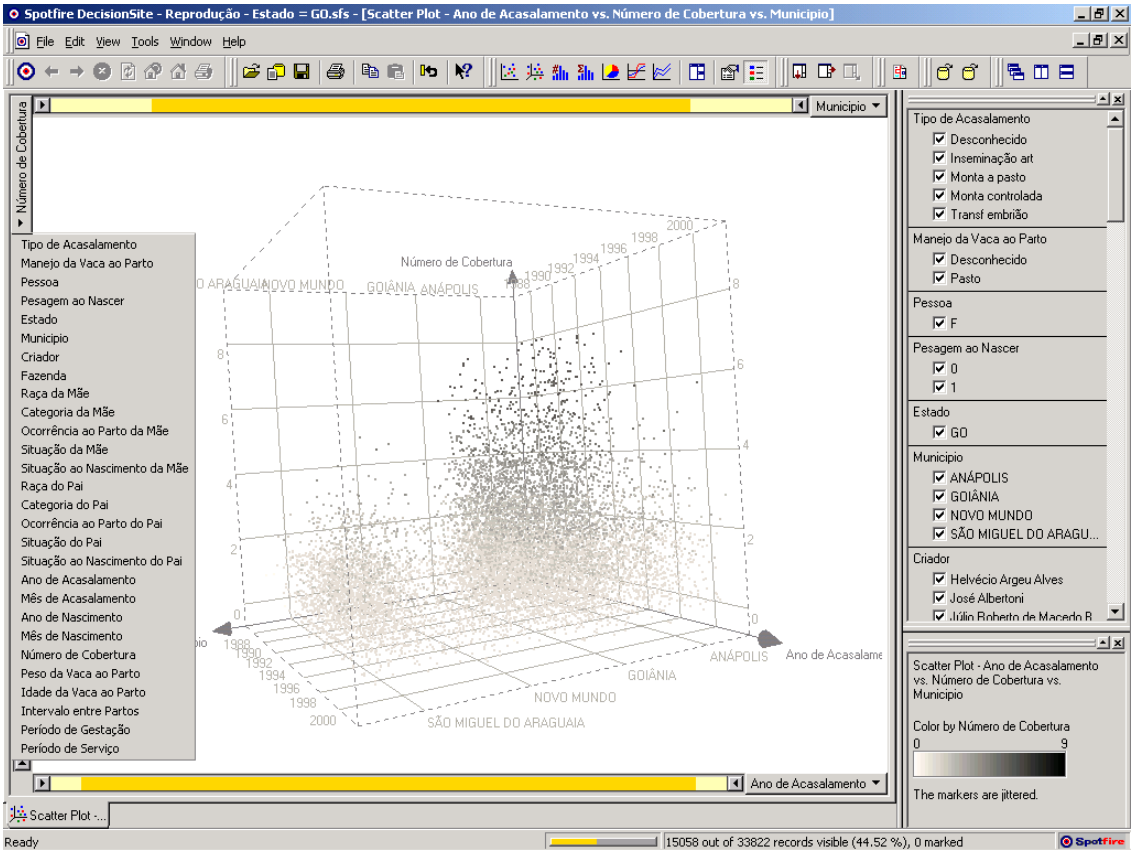


Figura 6.16: Gráfico *Scatter Plot 3D* dos atributos Município, Ano de Acasalamento e Número de Cobertura.

os atributos analíticos que podem ser escolhidos para compor as dimensões do gráfico. Esse gráfico é composto por quatro dimensões. Os três eixos do gráfico são representados pelos atributos Sexo, Raça e Categoria e a quarta dimensão é representada pela cor, sendo a mesma baseada nos valores do atributo PN. Cores fracas representam valores baixos e cores fortes representam valores altos, conforme a legenda apresentada no canto inferior direito da figura. Pode-se observar no gráfico que os maiores valores de PN estão concentrados na intersecção da Categoria Puro de Origem com a Raça Nelore Padrão.

Na Figura 6.16 é apresentado um gráfico do arquivo SFS de Reprodução cujos dados também correspondem ao Estado de GO. No *menu popup* ao lado esquerdo do gráfico são mostrados os atributos analíticos que podem ser escolhidos para compor as dimensões do gráfico. Os três eixos do gráfico são representados pelos atributos Município, Ano de Acasalamento e Número de Cobertura. Pode-se observar no gráfico que no período de 1987 a 2001 o Município de Goiânia possui a maior quantia de registros de reprodução. Esse município possui também os

mais elevados Números de Cobertura.

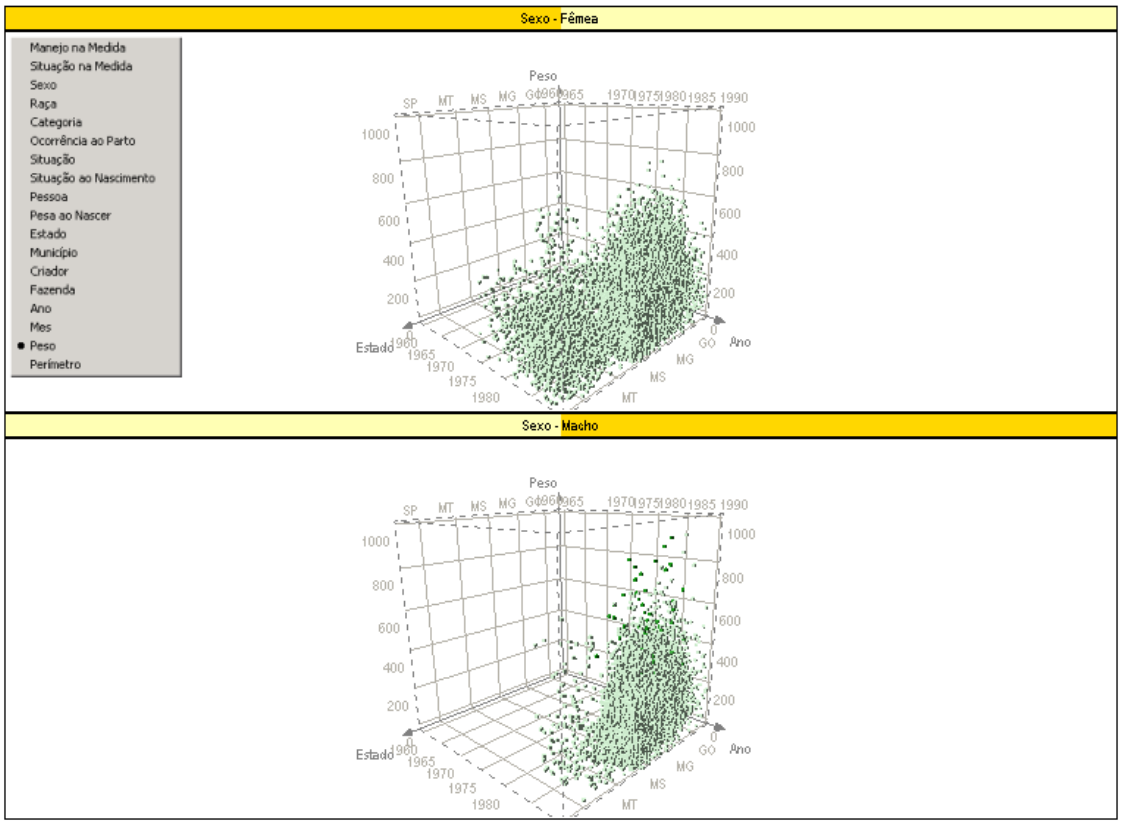


Figura 6.17: Gráfico *Scatter Plot 3D* dos atributos Peso, Ano, Estado, Sexo.

Nas Figuras 6.17 e 6.18 são apresentados os gráficos do arquivo SFS de Medidas cujos dados correspondem aos anos maiores do que 1960 e menores do que 1989. No *menu popup* localizado no canto superior esquerdo do gráfico da Figura 6.17 são mostrados os atributos analíticos que podem ser escolhidos para compor as dimensões do gráfico. Os três eixos dos gráficos da Figura 6.17 são representados pelos atributos Ano, Estado e Peso. Nesse exemplo, a quarta dimensão é representada pelo atributo Sexo, sendo que para isso foi criado um gráfico para cada valor do atributo Sexo. Pode-se observar na figura que a maior parte dos registros são do Sexo Fêmea.

Os gráficos da Figura 6.18 possuem cinco dimensões. Além das três dimensões dos eixos, são criados seis gráficos para representar todas as combinações possíveis dos valores dos atributos Sexo e Raça, que representam mais duas dimensões. Pode-se observar nos gráficos que a maioria dos registros de Peso pertencem à Raça Nelore Padrão.

As atividades apresentadas nesta seção correspondem à fase de pré-proces-

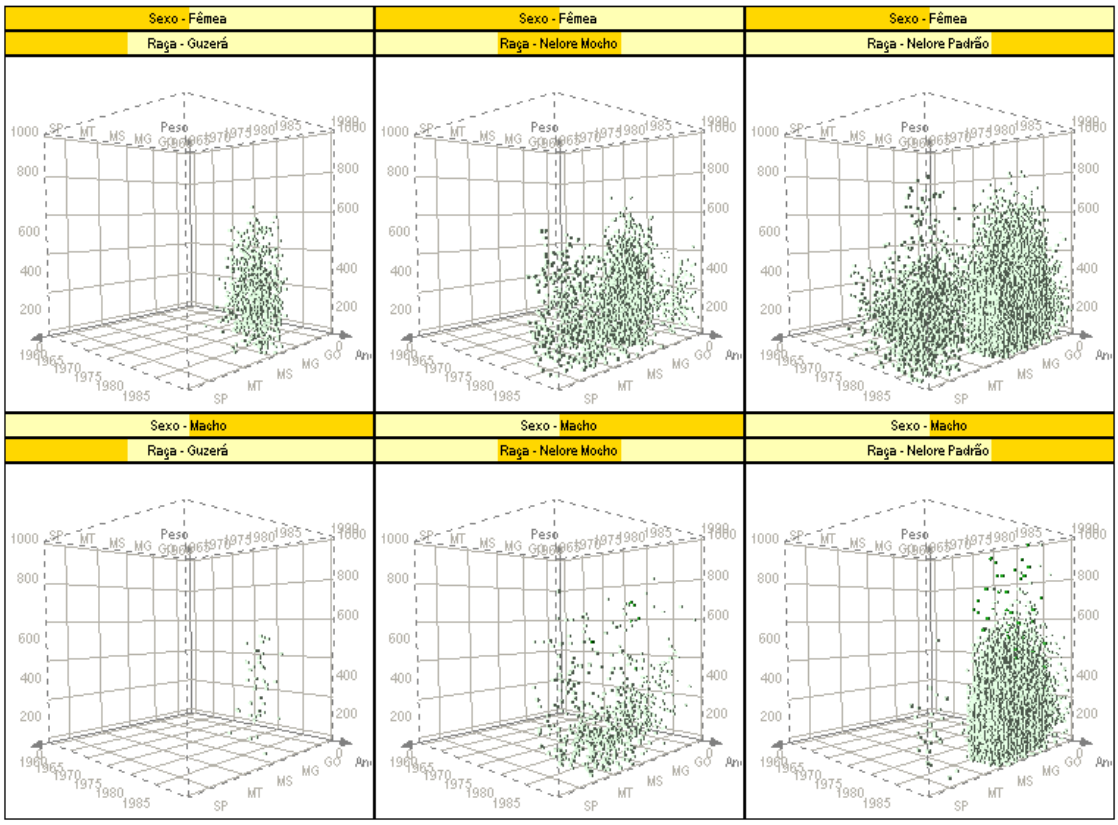


Figura 6.18: Gráfico *Scatter Plot 3D* dos atributos Peso, Ano, Estado, Sexo e Raça.

samento do processo de *Data Mining*. A atividade de compreensão do domínio foi realizada no Capítulo 2 e ficarão por conta dos especialistas do PMGRN as atividades de Extração de Padrões visualmente das outras diversas combinações de atributos que podem ser feitas, Pós-processamento e Utilização desses padrões.

De forma geral, foram verificados alguns padrões interessantes nas poucas amostras que foram apresentadas. Sendo eles, os três valores anormais do atributo NF455 da Figura 6.14. Uma grande concentração dos maiores valores do atributo PN nos registros que possuem o valor Puro de Origem para o atributo Categoria e o valor Nelore Padrão para o atributo Raça, conforme apresentado na Figura 6.15. Esta pequena amostra ressalta que é possível extrair visualmente muitos padrões interessantes dos dados do PMGRN.

Nesta seção foi feita uma breve demonstração das análises visuais que podem ser feitas com o Spotfire, com o intuito de encontrar tendências e padrões nos dados. Os pesquisadores do PMGRN poderão realizar muitos tipos de análises, considerando diferentes atributos através de várias dimensões.

6.3 Considerações Finais

Com as tecnologias apresentadas neste capítulo podem ser feitas as mais variadas análises com os dados do PMGRN. Com o Oracle Discoverer foi possível obter uma visão multidimensional dos dados agregados, ao contrário da visão tabular dos relatórios gerados nos ambientes operacionais. Esse fato foi comprovado através da apresentação de uma amostra das consultas OLAP que foram implementadas e também de uma breve análise feita sobre as mesmas. Com o Spotfire é possível visualizar graficamente os dados e extrair padrões deles, considerando-se diversas dimensões. Isto também foi comprovado através dos gráficos apresentados e analisados. Em suma, com ambas as ferramentas é possível fazer uma diversidade de combinações de atributos, tornando possível realizar uma grande variedade de análises. Além da análise, estas ferramentas proporcionam aos especialistas do PMGRN uma maneira de encontrar erros em seus dados, os quais poderão ser corrigidos, ou seja, elas também auxiliam na limpeza dos dados do Programa.

Este capítulo apresentou amostras dos tipos de análises que podem ser feitas com os dados do ambiente de suporte à decisão do PMGRN. No próximo capítulo serão apresentadas as conclusões deste trabalho.

CAPÍTULO

7

Conclusão

Este trabalho objetivou a construção de um ambiente de apoio à tomada de decisão para o Programa de Melhoramento Genético da Raça Nelore, aumentando assim, a capacidade deste em extrair informação e conhecimento de seus dados. Atualmente, o Programa é apoiado por dois sistemas, o SisNe (Sistema Nelore) e o ANCPWeb. O SisNe é o sistema responsável pelo gerenciamento dos dados operacionais do Programa, enquanto o ANCPWeb é um sistema de consulta voltado para a Internet, no qual os criadores podem fazer pesquisas relacionadas às características genéticas dos animais cadastrados no Programa. Através desse sistema é possível saber, por exemplo, quais são os melhores animais para uma característica específica.

A maior contribuição, deste trabalho, refere-se ao projeto e desenvolvimento de um ambiente que incremente o poder de análise dos dados do PMGRN, uma vez que os sistemas existentes (SisNe e ANCPWeb) não dão suporte aos tipos de análises que podem ser realizadas com as tecnologias aqui apresentadas.

Este trabalho iniciou-se com a compreensão do domínio em questão e o entendimento de como o banco de dados operacional do PMGRN estava estruturado. Após essa etapa passou-se para a construção do *Data Mart* do PMGRN, bem como a adição desse assunto no *Data Warehouse* do Departamento de Genética, seguindo a metodologia apresentada na Subseção 3.1.4.

Na primeira etapa da metodologia adotada foram apresentadas as justificativas para o desenvolvimento deste DW. Na etapa de planejamento foram apresentadas a visão corporativa do DW, a arquitetura e topologia do sistema e a área de negócio a ser trabalhada. Na etapa de análise foram levantados os re-

quisitos do ambiente, os quais resultaram nos modelos de dados do DW e do *Data Mart*. Foram criados um DER com dezessete entidades para o DW e em um esquema constelação com quatro estrelas para o *Data Mart*, utilizando o Oracle Warehouse Builder. Na etapa de projeto foram acrescentados mais detalhes de implementação a esses modelos, foram criados também os mapeamentos para o povoamento do DW à partir do banco de dados operacional e para o povoamento do *Data Mart* à partir do DW, por meio do Oracle Warehouse Builder. Verificou-se a necessidade de particionar os dados e qual estratégia seria utilizada para capturar as mudanças ocorridas no ambiente operacional. Na etapa de implementação foram criados os objetos físicos no SGBD Oracle e foi executada a atividade de povoamento do DW e do *Data Mart*. Foram criadas as consultas OLAP com o Oracle Discoverer e os gráficos para a mineração visual dos dados com o Spotfire.

Durante a construção das consultas OLAP verificou-se a necessidade de deixar previamente pronta a junção das tabelas fatos com suas dimensões, através da criação de visões materializadas, as quais, contribuíram para melhorar o desempenho das referidas consultas, pois a execução das consultas sobre o esquema estrela eram computacionalmente custosas devido as junções. Verificou-se também a necessidade de organizar corretamente os fatos e os atributos a serem analisados, para que o Oracle Discoverer executa-se com um consumo adequado de memória. Quanto aos gráficos construídos no Spotfire, o uso das visões materializadas também foi fundamental para que a ferramenta pudesse recuperar os dados com rapidez.

A forma na qual os dados do *Data Mart* foram estruturados faz com que as consultas OLAP sejam executadas mais eficientemente, além de não sobrecarregar o ambiente operacional, uma vez que o *Data Mart* está fisicamente separado deste.

O *Data Mart* foi construído com os atributos mais significativos do ambiente operacional. Esses atributos foram chamados de fatos e atributos analíticos. Os fatos são os atributos que podem ser quantificados, sendo os mesmos qualificados pelos atributos analíticos. Assim, o *Data Mart* é composto pelos atributos considerados como mais significativos para análise, tanto por consultas OLAP, como por ferramentas de *Data Mining*.

Como o ambiente está em fase de implantação, espera-se que ele possa contribuir para as pesquisas relacionadas ao melhoramento genético da raça Nelore e melhorar o processo de tomada de decisão dos criadores e pesquisadores do PMGRN. Isso pode ser atestado pelo interesse desses em conhecer e analisar o

que foi proposto.

Pelas poucas análises apresentadas, pôde-se verificar que as técnicas utilizadas para o desenvolvimento deste ambiente analítico trarão grandes benefícios aos criadores e pesquisadores do PMGRN. Uma vez que eles poderão verificar tendências e encontrar padrões que contribuirão para melhorar a seletividade dos animais da raça Nelore.

Outra importante contribuição, alcançada com a construção do DW, refere-se ao fato da extração de padrões utilizando tanto ferramentas visuais quanto automáticas poder ser realizada com mais facilidade.

Durante o desenvolvimento deste projeto surgiram muitos problemas e dificuldades. A primeira das dificuldades foi instalar e configurar as ferramentas necessárias para o desenvolvimento, gerenciamento e utilização do ambiente. Outra dificuldade surgiu na fase de desenvolvimento do DW. A etapa mais crítica dessa fase foi a extração dos dados da base de dados operacional do PMGRN, devido aos erros que ocorreram e que foram apresentados na Seção 5.5. Na fase de construção das consultas OLAP ocorreram problemas de desempenho e utilização de memória. Por meio deste trabalho pôde-se adquirir muita experiência no desenvolvimento de sistemas de apoio à tomada de decisão, bem como, no aperfeiçoamento do conhecimento de algumas ferramentas e aprendizado de outras.

Com relação a trabalhos futuros que possam complementar o que foi desenvolvido, destacam-se:

- A implementação do *Data Mart* do PMGRN utilizando um SGBD Multimensional;
- A implementação de consultas MOLAP;
- A comparação de desempenho entre a solução multidimensional e a solução relacional aqui utilizada;
- A mineração dos dados do PMGRN utilizando ferramentas que gerem modelos simbólicos, conexionistas ou estatísticos;
- A adição de novos assuntos ao *Data Warehouse* do Departamento de Genética;
- A validação, pelos usuários do PMGRN, do ambiente desenvolvido com relação ao desempenho e qualidade das consultas implementadas.

Referências Bibliográficas

Adriaans, P. & D. Zantinge (1996). *Data Mining*. Addison-Wesley Longman. 40

Ahlberg, C. (1996). *Spotfire: an information exploration environment. ACM SIGMOD Record 25*(4), 25–29. 64

Alison, S., G. Robinson, & P. Terhune (2001, Novembro). *Oracle9i Warehouse Builder User's Guide*. Technical Report A95931-01, Oracle Corporation, Califórnia. 59

Azmy, A. (1998, Maio). *SuperQuery: Data Mining for Everyone. Azmy Thinkware.* Disponível em: `http://www.azmy.com/wp1.htm`, [03/2001]. 38

Baranauskas, J. A. & M. C. Monard (2000). *Reviewing Some Machine Learning Concepts and Methods*. Technical Report 102, ICMC-USP, São Carlos. Disponível em: `ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_102.ps.zip`, [01/2001]. 44

Barquini, R. (1996). *Planning and Designing the Warehouse*. New Jersey: Prentice-Hall. 18, 24, 25, 27, 29

Batista, G. E. A. P. A. (1997, Setembro). *Um Ambiente de Avaliação de Algoritmos de Aprendizado de Máquina Utilizando Exemplos*. Dissertação de Mestrado, ICMC-USP. 43, 46

Batista, G. E. A. P. A. (2000, Março). *Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado*. Minidissertação para Qualificação de Doutorado, ICMC-USP. 41, 48

Bauer, A. & W. Lehner (1997). *The Cube-Query-Language (CQL) for Multidimensional Statistical and Scientific Database Systems*. In *International Conference*

on Database Systems for Advanced Applications, Melbourne - Australia, pp. 263–272. 29

Bohn, K. (1997, Junho). *Converting Data for Warehouse*. *DBMS 10*(07), 61–68. 36

Braga, A. P., A. P. L. F. Carvalho, & T. B. Ludermir (2000). *Redes Neurais Artificiais: Teoria e Aplicações*. Rio de Janeiro: LTC Editora. 44

Brand, E. & R. Gerritsen (1998). *Classification and Regression. DBMS, Data Mining Solutions Supplement*. Disponível em: `http://www.dbmsmag.com/9807m04.html`, [03/2001]. 42

Brownbridge, P. R. (2000, Dezembro). *Oracle Discoverer Plus User's Guide Release 4.1 For Windows*. Technical Report A86732-01, Oracle Corporation, Califórnia. 62, 63

Bustamente, G. & K. Sorenson (1994). *Decision support at Lands End - An evolution*. *IBM Systems Journal 33*(2), 228–238. 76

Cabena, P., P. Hadjinian, R. Stadler, J. Verhees, & A. Zanasi (1998). *Discovering Data Mining, From Concept to Implementation*. Upper Saddle River, New Jersey: Prentice Hall PTR. 40

Campos, M. L. & A. V. Rocha (1997). *Data Warehouse*. *Jornada de Atualização em Informática - JAI 16*, 221–261. 30

Carbonel, J. G. & P. Langley (1987). *Machine Learning - Encyclopedia of Artificial Intelligence*. Ed. John Wiley & Sons. 46

Chaudhuri, S. & U. Dayal (1997, Março). *An Overview of Data Warehousing and OLAP Technology*. *SIGMOD Record 26*(1), 65–74. 31, 36

Codd, E. F. (1993). *Providing OLAP (On-Line Analytical Processing) to User-Analyst: an IT Mandate*. E. F. Codd and Assoc. 31, 33

Colliat, G. (1996, Setembro). *OLAP, Relational and Multidimensional Database Systems*. *SIGMOD Record 25*(03), 64–69. 29

Corey, M., M. Abbey, I. Abramson, & B. Taub (2001). *Oracle8i Data Warehouse*. Rio de Janeiro-RJ: Editora Campus. 21, 22, 26, 29, 31, 32, 33, 36, 56, 57, 69

Decker, K. & S. Focardi (1995, Fevereiro). *Technology Overview: A Report on Data Mining*. Technical report, Swiss Center for Scientific Computing. Disponível em: http://www.cscs.ch/Official/TechReports/1995/CSCS-TR-95-02.ps.gz, [02/2001]. 1

Elder, J. & D. Pregibon (1996). *A Statistical Perspective on Knowledge Discovery in Databases*. pp. 83–116. 44

Fayyad, U. (1996). *Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, 11*, 20–25. 44

Fayyad, U., D. Haussler, & P. Stolorz (1996). *KDD for Science Data Analysis: Issues and Examples*. In E. Simoudis, J. W. Han, & U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 50–56. AAAI Press. 39

Fayyad, U., G. Piatetsky-Shapiro, & P. Smith (1996, Novembro). *The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of ACM 39*(11), 27–34. 40, 44

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. In *Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press. 38, 42

Fayyad, U. & E. Simoudis (1997). *Data Mining and KDD: An Overview*. New Port Beach. *Tutorial in the Third International Conference on Knowledge Discovery in Data Mining*. 39

Fayyad, U. M., G. Platestsky-Shapiro, P. Smyth, & R. Uthurusamy (1996). *From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining*, pp. 1–34. Menlo Park, CA: AAAI/MIT Press. 2, 38, 47

Félix, L. C. M. (1998, Agosto). *Data Mining no Processo de Extração de Conhecimento de Bases de Dados*. Dissertação de Mestrado, ICMC-USP. 41

FNP (1995). *ANUALPEC 95, Anuário Estatístico de Pecuária de Corte*. Technical report, Consultoria & Comércio FNP, São Paulo. 2

Gardner, S. R. (1998). *Building the Data Warehouse. Communications of the ACM 41*(9), 52–60. 19

Gatziu, S. & A. Vavouras (1999). *Data Warehousing: Concepts and Mecanisms*. Heidelberg - Germany: DMDW. 27, 28

Glymour, C. (1997). *Statistic Themes and Lessons for Data Mining*. In *Data Mining and Knowledge Discovery, Vol 1*, pp. 11–28. Kluwer Academic Publishers. 44, 46

Golfarelli, M., D. Maio, & S. Rizzi (1998). *Conceptual Design of Data Warehouses from E/R Schemes*. In *Hawaii International Conference on System Sciences*, Kona - Hawaii. 27, 28

Gupta, V. R. (1997). *An Introduction to Data Warehousing*. Technical report, System Services Corporation, Chicago, Illinois. 23

Han, J. (1995). *Mining Knowledge at Multiple Concept Levels*. In B. Columbia (Ed.), *Proceedings of School of Computing Science Simon Fraser University*, Canada. 38

Han, J. (1999). *Data Mining*. In J. Urban & P. Dasgupta (Eds.), *Encyclopedia of Distributed Computing*. Kluwer Academic Publisher. 42

Haykin, S. (1994). *Neural Networks - A Comprehensive Foundation*. Macmillan College Publishing Company. 44

Holsheimer, M., M. Kerten, H. Mannila, & H. Toivonen (1995). *A Perspective on Database and Data Mining*. Technical report. 2

Horst, P. S. (1999, Agosto). *Avaliação do Conhecimento Adquirido por Algoritmos de Aprendizado de Máquina Utilizando Exemplos*. Dissertação de Mestrado, ICMC-USP. 44

Inmon, W. H. (1996, Novembro). *The Data Warehouse and Data Mining. Communications of ACM 39*(11), 49–50. 42

Inmon, W. H. (1997). *Como construir o Data Warehouse*. Rio de Janeiro: Editora Campus. 2, 17, 18, 24

Inmon, W. H. & R. D. Hackarthorn (1997). *Como usar o Data Warehouse*. Rio de Janeiro: Infobook / IBPI Press. 33, 37

Inmon, W. H., J. D. Welch, & K. L. Glassey (1999). *Gerenciando o Data Warehouse*. São Paulo: Makron Books. 23, 31

Kerber, R., B. Livezey, & E. Simoudis (1995). *A Hybrid System for Data Mining*. John Wiley & Sons. 43

Kimball, R. (1997). *Data Warehouse Toolkit*. São Paulo: Makron Books. 2, 18, 19, 24, 27, 28, 30, 42

Kliber, D., B. Livezey, & E. Simound (1988). *Machine Learning as a Experimental Science*. *Machine Learning 3*(1), 5–8. 43

Labio, W., D. Quass, & B. Adelberg (1997). *Physical Database Design for Data Warehousing*. In *International Conference on Data Engineering*, Binghamton - UK. 25

Lane, P. (2001, Junho). *Oracle9i Data Warehousing Guide*. Technical Report A90237-01, Oracle Corporation, Califórnia. 54, 57

Li, B. (1996, Janeiro). *Data Mining* NOW, *a survey and thesis proposal*. Technical report, New York University. 2

Lôbo, R. B., L. A. F. Bezerra, H. N. Oliveira, C. U. Magnabosco, M. A. R. Freitas, & J. A. G. Bergmann (2002). *Avaliação Genética de Animais Jovens, Touros e Matrizes*. Technical report, GEMAC - Departamento de Genética - FMRP - USP, Ribeirão Preto. 6, 7, 8, 9

Luscher, L. M. (2001, Junho). *Oracle9i Database Concepts*. Technical Report A88856-02, Oracle Corporation, Califórnia. 52, 53

Mannila, H. (1997a). *Methods and Problems in Data Mining*. In *Proceedings International Conference on Database Theory (ICDT-97)*, Delphi, Greece. Springer-Verlag. 46

Mannila, H. (1997b, Junho). *Data Mining: Machine Learning, Statistic and Databases*. *Eight International Conference on Scientific and Statistical Database Management*, 1–8. Disponível em: `http://www.cs.helsinki.fi/~mannila/`, [03/2001]. 38, 40

Marques, V. F., R. H. Suzuqui, G. S. Araújo, & J. C. C. Neto (2000, Fevereiro). *Projeto e Construção de um Sistema de Banco de Dados que Permita Acesso Pela*

Internet Usando o SGBD Oracle. Projeto de Graduação apresentado no curso de Bacharelado em Ciência da Computação - DCT - CCET - UFMS. 28

Miley, M. (1997, Setembro). *Bring the World to Your Warehouse*. *Oracle Magazine 11*(05), 51–74. 37

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. 44, 46

Monard, M. C., G. E. A. P. A. Batista, S. Kawamoto, & J. B. Pugliesi (1997). *Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos*. Technical report, ICMC-USP, São Carlos. Disponível em: `http://labic.icmc.sc.usp.br/didatico/PostScript/ML.ps.zip`, [01/2001]. 43

Moriarty, T. & R. P. Greenwood (1996, Outubro). *Data's Quest from Source to Query*. *Database Programming & Design 9*(10). 37

Netz, A., S. Chaudhuri, J. Bernhardt, & U. M. Fayyad (2000). *Integration of Data Mining with Database Technology*. In A. E. Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, & K.-Y. Whang (Eds.), *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pp. 719–722. Morgan Kaufmann. 40

Oracle (2000, Dezembro). *Oracle Discoverer Administration Edition Administration Guide Release 4.1 For Windows*. Technical Report A86730-01, Oracle Corporation, Califórnia. 60, 61

Orli, R. J. (2001, Março). *Data Extraction, Transformation and Migration Tools*. Disponível em: `http://www.kismeta.com/ex2.html`, [03/2001]. 22

Orr, K. (2000). *Data Warehousing Technology*. *The Ken Orr Institute*. Disponível em: `http://www.kenorrinst.com/pg 33 d.w. whitepaper.htm`, [04/2001]. 23

Padilha, T. P. P. (1999, Março). *Investigação de Algoritmos de Aprendizado de Máquina Pertencentes ao Paradigma Estatístico para Aquisição de Conhecimento*. Dissertação de Mestrado, ICMC-USP. 44, 46

Padmanabhan, B. & A. Tuzhilin (1999, Junho). *Unexpectedness as a measure of interestingness in knowledge discovery*. *Decision Support Systems 27*, 303–318. Disponível em:

http://e5500.fapesp.br/cgi-bin/sciserv.pl?collection=journals-&journal=01679236&issue=v27i0003&article=303_uaamoiikd, [04/2001]. 44

PMGRN (2002, Agosto). *Programa de Melhoramento Genético da Raça Nelore - Grupo de Genética, Melhoramento Animal e Computação*. Disponível em: http://www_gen.fmrp.usp.br/gemac/pmgrn, [08/2002]. 9, 10, 11, 12, 13, 14

Poe, V., P. Klauber, & S. Brobst (1998). *Building a Data Warehouse for Decision Support*. New Jersey: Prentice-Hall. 2, 17, 18, 24

Pugliesi, J. B. (2001, Março). *O Pós-Processamento em Extração de Conhecimento de Bases de Dados*. Minidissertação para Qualificação de Doutorado, ICMC-USP. 44

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Los Altos, California, USA: Morgan Kaufmann Publishers, Inc. 47

Ramoni, M. & P. Sebastiani (1997). *Bayesian Inference With Missing Data Using Bound and Collapse*. Technical Report 58, Knowledge Media Institute - The Open University. 47

Rezende, S., R. Oliveira, L. Félix, & C. Rocha (1998). *Visualization for Knowledge Discovery in Databases*. In *Ebecken, N.F.F.*, England, pp. 81–95. Data Mining, WIT Press. 40, 45

Rezende, S. O. & E. S. Moreira (2000, Dezembro). *Tecnologia da Informação*. In *Fábrica do Futuro - Entenda hoje como sua indústria vai ser amanhã*, São Paulo, pp. 99–104. Editora Banas. 16

Rezende, S. O., C. A. J. Rocha, & R. B. Lôbo (2000). *Bayesian Networks for Knowledge Discovery in a Database from the Program for Genetic Improvement of the Nelore Breed*. In N. Ebecken (Ed.), *Second International Conference on Data Mining, Volume II*, England, pp. 15–24. WIT Press - Computational Mechanics Publications. 9

Rich, E. & K. Knight (1993). *Inteligência Artificial* (2^a ed.). Makron Books. 46

Rocha, C. A. J. (1999, Março). *Redes Bayesianas para Extração de Conhecimento de Bases de Dados, Considerando a Incorporação de Conhecimento de*

Fundo e o Tratamento de Dados Incompletos. Dissertação de Mestrado, ICMC-USP. 38

Russell, S. & P. Norvig (1995). *Artificial Intelligence - A Modern Approach*. Prentice Hall. 46

Salzberg, S. L. (1997). *On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach*. In *Data Mining and Knowledge Discovery, Vol 1(3)*. 43

Samos, J., F. Saltor, J. Sistac, & A. Bard'es (1998). *Database Architeture for Data Warehousing: An Evolutionary Approach*. In G. Quirchmayr, E. Schweighofer, & T. J. M. Bench-Capon (Eds.), *Database and Expert Systems Applications*, *Lecture Notes in Computer Science*, Vienna, Austria, pp. 746–766. Springer-Verlag. 19

Sherman, R. (1997, Agosto). *Metadata: The Missing Link*. *DBMS 10*(09), 73–82. 23

Shoshani, A. (1997). *OLAP and Statistical Database: Similarities and Differences*. *ACM TODS*, 185–196. 30

Spotfire (2002, Janeiro). *Spotfire DecisionSite 7.0 - User's Guide and Reference Manual*. Technical report, Spotfire, Sweden. 66

Todman, C. (2001). *Designing a Data Warehouse - Supporting Customer Relationship Management*. New Jersey: Prentice-Hall. 28

Weiss, S. M. & N. Indurkhya (1998). *Predictive Data Mining: A Pratical Guide*. Morgan Kaufmann Publishers Inc. 42

Weiss, S. M. & C. A. Kulikowski (1991). *Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. In *Computer Systems that Learn*. Morgan Kaufmann Publishers inc. 46

Weldon, J. L. (1997, Janeiro). *Warehouse Cornerstones*. *Byte 22*(1), 82–88. 23

Williams, J. (1997, Junho). *Tools for Traveling Data*. *DBMS 10*(07), 69–76. 36

Wu, M. & A. Buchmann (1997). *An Overview of Data Warehousing and OLAP Technology*. BTW'97, Ulm. 30