# Apoio no desenvolvimento de uma Data Warehouse da bovinocultura de corte brasileira

William Pereira dos Santos Wada[1]
Luís Gustavo Barioni[2]

## Introdução

Um diagnóstico e uma avaliação aprofundados sobre a condição e a perspectiva dos sistemas de bovinocultura de corte no Brasil requerem a integração de dados de sua estrutura, tecnologias adotadas, uso da terra, produtividade, clima, solos, recursos naturais e condições socioeconômicas regionais, entre outros. Esses dados estão atualmente dispersos em diferentes bancos de dados pertencentes a diversas instituições. Levantamentos da estrutura produtiva e da dispersão espacial do uso da terra, por exemplo, são obtidos pelos Censos Agropecuários do Instituto Brasileiro de Geografia e Estatística (IBGE). Dados de clima são monitorados pela rede AGRITEMPO, liderada pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa). Já dados relacionados aos custos de produção são obtidos pelo Centro de Estudos Avançados em Economia Aplicada (Cepea), da Escola Superior de Agricultura Luiz de Queiroz (Esalq) da Universidade de São Paulo (USP).

Assim, a integração de dados se faz necessária, para facilitar a análise, o manuseio das informações e o diagnóstico multidimensio-

[1] *Faculdade de Tecnologia Unicamp; willianpsw@cnptia.embrapa.br*

[2] *Embrapa Informática Agropecuária; barioni@cnptia.embrapa.br*

nal da situação atual da bovinocultura de corte por meio de análise histórica dos dados. Segundo Inmon (1987) o desenvolvimento de um *data warehouse* (armazém de dados) para armazenar esses dados, integrá-los e agrupá-los com uma arquitetura que suporte essas avaliações é uma solução recomendada para esse tipo de problema.

O *data warehouse*, aqui apresentado é produto do projeto AVISAR – Avaliação dos impactos ambientais, sociais e econômicos dos sistemas de bovinocultura de corte na Amazônia, no Cerrado e no Pantanal, em desenvolvimendo na Embrapa Informática Agropecuária em parceria com CEPEA/USP e IBGE.

## Material e métodos

Critérios para a escolha do Sistema de Gerenciamento de Banco de Dados (SGBD), estabelecidos conforme Machado (2006) e consultando especialistas na área, foi o inicio das etapas de construção do Data Warehouse (DWH). As ferramentas para as etapas de construção também foram escolhidas com as mesmas fontes de consulta do SGBD.

A importância em escolher BDs que tenham o *driver* ODBC/JDBC foi apontada pelo Dr. Prof Fabio Primak (informação pessoal). Esses *drivers* permitem a comunicação do BD com outros sistemas que precisam interagir com os dados. A capacidade de volume de dados suportado pelo BD e o alinhamento dos propósitos de criação foram observados pelo consultor Sênior em Business Intelligence Tiago Souza (informação pessoal). Como o DWH tende a apenas aumentar em volume, devido ao seu caráter histórico, é necessário um BD que suporte grandes volumes. O alinhamento ajuda a reduzir campos da tabela que serão desnecessários aos propósitos do projeto, reduzindo o volume do DWH. A compatibilidade entre BD e o SO (Sistema Operacional) é importante pois a Embrapa Informática Agropecuária dá preferência a SO de licença *Open Source*. Capacitação no BD pelos pesquisadores foi outro critério relevante. Finalmente, estabeleceu-se como requisito o suporte a arquivos de Sistemas de Informação Geográfica (SIG).

Para as ferramentas de migração dos dados via *Extract Transform Load* (ETL) e modelagem de dados, o enfoque do critério de seleção foram a licença *Open source*, a facilidade de comunicação com diversos sistemas de BD e a facilidade de entendimento.

Identificar a granularidade dos dados foi a próxima etapa. Determinar o nível de detalhe dos dados é de suma importância para: (a) evitar processamento demorado; (b) não perder informação importante com nível de detalhe de dados muito baixo e; (c) se preocupar com o volume quando aumentado o nível de detalhamento. Note-se que o volume de dados e a quantidade de informação dos dados são inversamente proporcionais, necessitando o balanceamento entre essas duas questões.“O maior problema para os desenvolvedores de *data warehouse* é determinar o nível apropriado de granularidade dos dados[...] quando o nível está configurado corretamente os aspectos da concepção e implementação flui suavemente[...]”(INMON, 1987).

A arquitetura e implementação surgem como necessidades a serem estabelecidas depois de resolvida a questão da granularidade dos dados e as ferramentas a serem utilizadas. A Infraestrutura a ser usada para armazenagem dos dados e a maneira como eles serão organizados e integrados são questões pertinentes à arquitetura e à implementação. Machado (2006) aponta três tipos de arquiteturas as mais conhecidas: (a) global; (b) *Data Marts* Independente; (c) e *Data Marts* Integrados. Há vantagens e desvantagens nos dois primeiros tipos, sendo o integrado a composição das vantagens dos demais. Para a implementação desta arquitetura existem duas abordagens mais comuns. Os *Star Schema*, termo para a designação de modelos de dados multidimensionais, onde o desenho da disposição das tabelas se assemelha a uma estrela, e o modelo *Snowflake* que é parecido com a *Star Schema*, porém, em suas dimensões encontram-se níveis hierárquicos.

O tratamento dos dados (extração, limpeza, integração, modelagem de dados multidimensional) é o passo seguinte no desenvolvimento que ainda não foi iniciado no projeto.

## Resultados e discussão

O SGBD escolhido, pela experiência que a Embrapa já tinha com o sistema e por atender os critérios estabelecidos, foi o PostGreSQL. Esse SGBD tem uma extensão POSTGIS para tratamento de dados espaciais, permitindo a visualização dos dados em mapas, outro item considerado como o critério.

Para a ferramenta ETL foi selecionado o *Kettle* da suíte Pentaho que atendia aos critérios já definidos, e descritos em material e métodos e demonstrou ser uma ferramenta eficaz para os propósitos do projeto.

A ferramenta para modelagem dos dados que apresentava as características de fácil usabilidade e ter licença livre é o SQL Power Architect.

A granularidade dos dados foi estipulada pelos níveis de sumarização que os dados se encontravam.

"A escolha da arquitetura, uma decisão gerencial do projeto e está baseada na Infra estrutura atualmente disponível" (MACHADO, 2006). Foi escolhida a arquitetura Global pois são poucas as bases a serem integradas e a alimentação desses dados, bem como a manipulação do DWH, será centralizada em apenas um local.

Para o modelo de dados foi definido o *Star Schema*. A melhor performance em relação a tempo de processamento faz do modelo *Star Schema* concenso entre os especialistas da área como a melhor solução(MACHADO, 2006). *Snow Flake* é interessante para o entendimento do modelo, por isso, muitos analistas desenham em *Snow Flake* mas implementam em *Star Schema*.

## Referências

INMON, W. H. **Building the Data Warehouse**, 4.ed. Indianopolis: Wiley, 1987.

MACHADO, F. N. R. **Tecnologia e projeto de Data Warehouse**. 4. ed. São Paulo: Érica, 2006.

SANTOS, F. M. **Modelo dimensional para data warehouse**. Disponível em: <http://fmsantos.infotuga.com/?p=219#more-219 > Acesso em: 03 jun. 2010.