Ministério da Agricultura, Pecuária e Abastecimento

Comunicado Técnico 3

Novembro, 2001
Campinas, SP

ISSN 1677-8464

Carregando Dados na Arquitetura Data Warehousing - Armazém de Dados de Frutas

Luiz Manoel Silva Cunha[1]
Carlos Aberto Alves Meira[2]
Taylor Encinas[3]

Uma das principais necessidades identificadas no projeto Sistema de Integração e Qualificação de Informação para a Cadeia de Frutas (Seixas Neto et al., 2000) é a possibilidade de integração dos dados da fruticultura brasileira, hoje sistematizados de forma dispersa em bancos de dados de diferentes instituições, como o Instituto Brasileiro de Geografia e Estatística - IBGE, com seus levantamentos sobre a produção agrícola, e a Secretaria de Comércio Exterior - SECEX, com os dados de exportação e importação.

Atualmente, esses dados são recebidos do IBGE e da SECEX ou consultados nessas instituições pelos interessados e os cruzamentos de informação têm que ser feitos manualmente. Ou seja, análises e balanços a respeito da fruticultura em nível estadual, regional e nacional não estão sistematizados e nem automatizados.

Algumas dificuldades e problemas desse processo não sistematizado são:

- demora em extrair os dados específicos da cadeia de frutas;
- alguns cruzamentos de informação das diferentes fontes são feitos manualmente;
- outros dados têm que ser transportados para um sistema de planilhas, para permitir os cruzamentos; e
- erros e inconsistências podem ser incorporados nesse processo.

Para superar as dificuldades apresentadas, entre outras, elaborou-se o projeto Sistema de Integração e Qualificação da Informação para Cadeia de Frutas (Seixas Neto, 2000). Seu objetivo é buscar a integração de bancos de dados distribuídos para atender as necessidades do Programa de Desenvolvimento da Fruticultura (PROFRUTA), coordenado pelo Ministério da Agricultura, Pecuária e Abastecimento, agilizando o acompanhamento e a análise do desempenho dos produtos agrícolas da cadeia de frutas e oferecendo suporte a decisões estratégicas.

O produto esperado é um Armazém de Dados "*Data Warehouse*" de Frutas disponível na Web inserido na *home page* do programa PROFRUTA, onde o usuá-

[1] *M.Sc. em Ciências e Matemática Computacional, Técnico de Nível Superior III, Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo, 13083-970 – Campinas, SP (email: luizm@cnptia.embrapa.br)*

[2] *M.Sc. em Ciências e Matemática Computacional, Pesquisador da Embrapa Informática Agropecuária. (email: carlos@cnptia.embrapa.br)*

[3] *Estagiário da Embrapa Informática Agropecuária. (email: taylor_encinas@yahoo.com.br)*

Embrapa

rio poderá requerer a geração de relatórios específicos, para atender demandas do programa e realizar consultas complexas para suporte à decisão.

Data Warehouse/Data Mart

Conceitos

Data Warehouse é um banco de dados voltado para o suporte à decisão de usuários finais, derivado de diversos outros bancos de dados operacionais. Data Warehouse é, portanto, não volátil, integrado, variante no tempo e orientado a assunto (Inmon, 1998). O ambiente no qual se insere o DW pode ser visto também como um conjunto de diversas tecnologias, como ferramentas de extração e conversão, bancos de dados voltados para consultas complexas, ferramentas inteligentes de prospecção e análise de dados e ferramentas de administração e de gerenciamento.

Uma boa solução de DW tem como finalidade atender as necessidades de análise de informações dos usuários, como monitorar e comparar as operações atuais com as passadas, e prever situações futuras. Ao transformar, consolidar e racionalizar as informações dispersas por diversos bancos de dados e plataformas, permite que sejam feitas análises estratégicas bastante eficazes com base em informações antes inacessíveis ou subaproveitadas. Usar os arquivos das aplicações tradicionais, com seus dados operacionais e muitas vezes redundantes, para análises de tendências é simplesmente impossível.

Data Mart é um banco de dados voltado somente para uma determinada área, faz uso das mesmas tecnologias de DW, abriga um volume menor dados e, o processo de construção e "população" é bem menos complexo do que em um DW (Toni, 1999).

Arquiteturas

A arquitetura de um DW pode variar de acordo com as necessidades da organização. A Fig. 1, mostra a arquitetura mais comum. Nesta, o DW é construído com base em um ou mais sistemas, onde os usuários os acessam diretamente o DW.

Na Fig. 2, é apresentado uma arquitetura mais complexa de um ambiente DW, com uma base de dados centralizada, um sistema utilizado para limpar e integrar os dados e múltiplos repositórios de dados utilizados para atender determinadas áreas de negócio da organização de Data Mart .

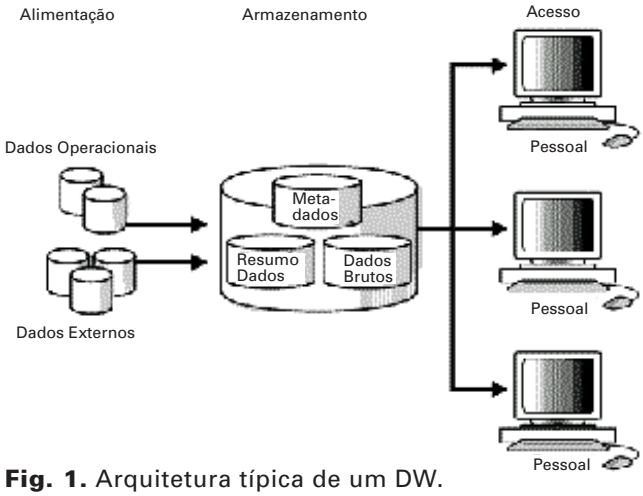

Fig. 1. Arquitetura típica de um DW.
Fonte: Oracle Corporation (2000).

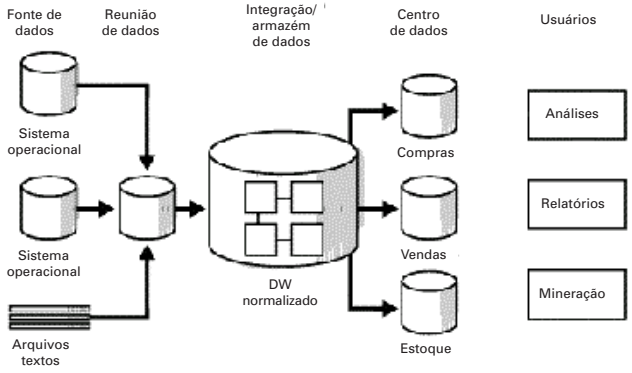

Fig. 2. Arquitetura mais complexa de um DW.
Fonte: Oracle Corporation (2000).

O Processo ETT (Extração, Transformação e Carga de Dados)

O processo de extração de dados, oriundos de várias fontes, e a sua colocação em um DW é chamado ETT (Oracle Corporation, 2000) que compreende as fases de Extração, Transformação e Transferência de dados. Esse processo deve ser entendido e visto como um grande e único processo e não como três processos bem definidos. Construir e manter um processo ETT é uma das partes mais difíceis no projeto de um DW, além de consumir bastante recurso e exigir um bom planejamento.

A primeira fase do processo ETT, diz respeito à definição das fontes de dados e extração deles. Os dados são copiados de uma base de dados para um arquivo ou para um ponto da rede de computadores. Os dados podem estar armazenados em diversos meios magnéticos e, em diferentes formatos, tais como: arquivos textos, arquivos tipo DBF (Data Base File), planilhas eletrônicas etc. Fatores relevantes que devem ser observados antes do início da fase extração de dados são apontados por Cielo (2001).

A fase seguinte, que trata da transformação de dados costuma ser também uma tarefa complexa e, em termos de tempo de processamento, é a que mais consome tempo. Normalmente, observa-se que os dados extraídos oriundos de várias fontes, contêm muito lixo e algumas inconsistências. Para que estes possam ser integrados ao DW, é necessário limpá-los para depois transformá-los. A limpeza dos dados é necessária para resolver problemas, tais como, remover 999999999-99 do campo CPF, por exemplo. A transformação irá tratar dos casos onde uma mesma informação possui diferentes formatos. Exemplo: em um sistema de cadastro de funcionários utiliza-se "H" para Homem enquanto que, em outro, utiliza-se "M", também para indicar Homem. Ao encerrar-se esta fase, os dados estarão padronizados, segundo as necessidades dos usuários e aptos a serem transferidos para o DW.

A etapa de transferência dos dados é, efetivamente, o ato de mover os dados de um sistema para o outro. Em um ambiente DW é possível transferir dados de: um sistema para uma base de dados ou para um DW; de uma base de dados para um DW; e de um DW para um *Data Mart*. Esta fase também é complexa e merece atenção para os seguintes aspectos (Cielo, 2001):

- Integridade dos dados;
- Se a tabela deve receber uma carga de dados incremental ou substitutiva;
- Embora existam ferramentas prontas para ETT, muita das vezes é necessário criar rotinas para carga de dados visando atender situações não previstas.

Exemplos que ilustram cada uma das fases do processo de ETT e técnicas possíveis de serem aplicadas são apresentadas na documentação da Oracle (Oracle Corporation, 2000), capítulos 11, 12 e 13.

Material e Métodos

Os materiais utilizados para construir o ADF e armazenar os dados foram: o Banco de Dados Oracle8i, software SQL*LOADER para carregar os dados armazenados em arquivos em diversos formatos para tabelas no banco de dados, ferramenta Case ERWin para modelagem dos dados, SQL*PLUS para processar os *scripts* de transformação e transferência dos dados armazenados nas tabelas, a linguagem SQL para construir os *scripts* e o sistema operacional Windows 2000.

A primeira etapa do processo de construção do Armazém de Dados de Frutas foi destinada à identificação das fontes de dados, obtenção dos arquivos de dados e de seus respectivos layouts. Em seguida, passou-se à construção de um modelo de dados, que representasse a integração das fontes de dados: Instituto Brasileiro de Geografia e Estatística - IBGE, com seus levantamentos sobre a produção agrícola e Secretaria de Comércio Exterior - SECEX. Num momento posterior, análises das fontes de dados foram realizadas e definiu-se quais transformações necessárias para tornar os dados aptos para o armazém. Em paralelo, investigou-se como utilizar a ferramenta SQL*LOADER. Terminada a fase investigação, foram construídos os arquivos de controle para dar carga no banco de dados e os *scripts,* seqüências de comandos na linguagem SQL, para transformar os dados armazenados no banco de dados e para carregar o ADF.

Arquivos de controle são arquivos texto escritos em uma linguagem que o SQL*LOADER entenda. Esses arquivos informam ao SQL*LOADER onde encontrar o arquivo de dados, como examiná-lo e interpretá-lo, bem como onde inserir os dados.

Resultados

Os resultados alcançados no projeto até o momento são:

a) Modelo de dados integrados (Meira et al., 2001b);
b) Modelo conceitual para o processo ETT aplicado à construção do Armazém de Dados de Frutas;
c) Um conjunto de regras para transformação e transferência dos dados do banco de dados para o armazém de dados;
d) Os arquivos de controles para carga no banco de dados e os *scripts* de transformação dos dados e carregamento do ADF;
e) O armazém de dados carregado e operacional;
f) Vários relatórios voltados para tomada de decisão (Meira et al., 2001a).

A Fig. 3 apresenta um arquivo de controle utilizado para carregar os dados para o ADF. A Fig. 4, exibe relatório em forma de gráfico, gerado com base nos dados do ADF.

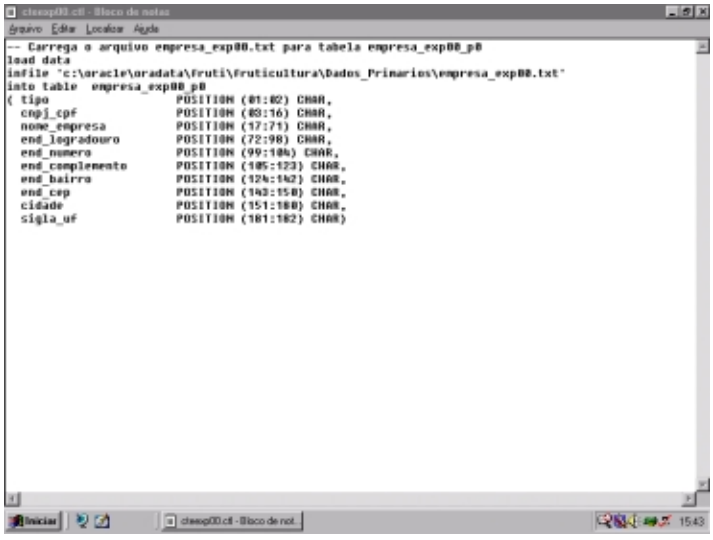

Fig. 3. Exemplo de arquivo de controle.

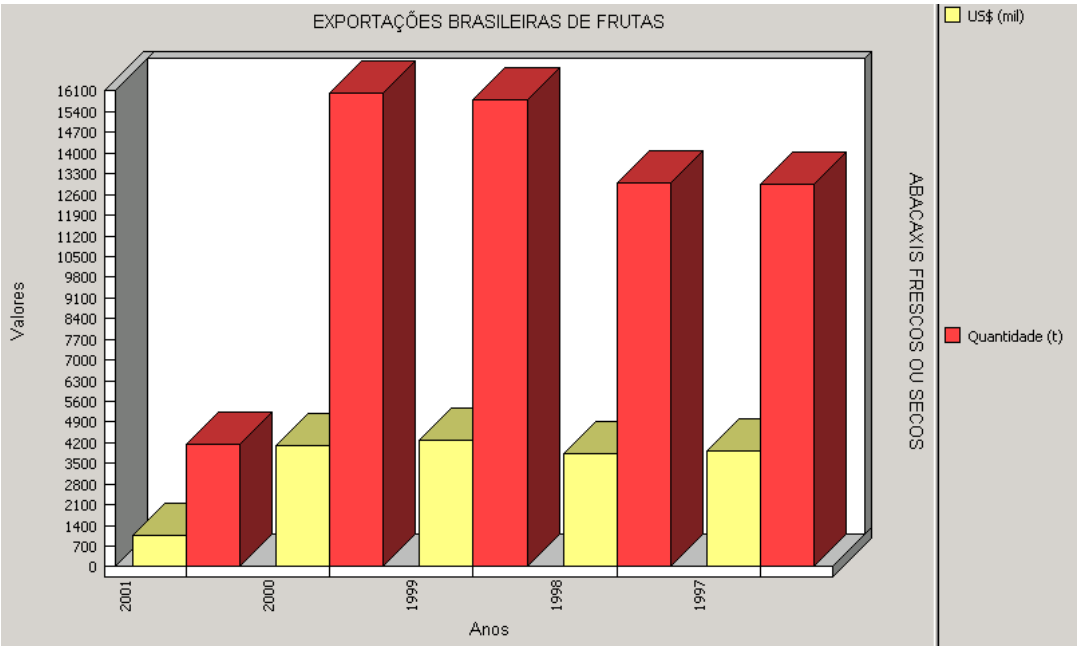

Fig. 4. Gráfico sobre exportação de frutas.
Fonte: Meira et al. (2001a).

Referências Bibliográficas

CIELO, I. **ETL – Extração, transformação e carga de dados**. Disponível em: <http://www.datawarehouse.inf.br/etl.asp>. Acesso em: 16 out. 2001.

INMON, W. D. **Como construir o data warehouse**. Rio de Janeiro: Campus, 1998.

MEIRA, C. A. A.; SEIXAS NETO, A.; CUNHA, L. M. S.; NAKA, J. **Análise de comércio exterior de frutas a partir do armazém de dados da fruticultura brasileira**. Campinas: Embrapa Informatica Agropecuária, 2001a. (Embrapa Informática Agropecuária. Comunicado Técnico, 17).

MEIRA, C. A. A.; SEIXAS NETO, A.; CUNHA, L. M. S.; NAKA, J. Armazém de dados da fruticultura brasileira. **Revista Tecnologia da Informação**, Brasília, 2001b. No prelo.

ORACLE CORPORATION. Oracle8i/personal edition on-line documentation CD-ROM: release 2(8.1.6) for Microsoft Windows 98. In: ORACLE CORPORATION. **Oracle database 8i realise 2(8.1.6) for Microsoft Windows**. Redwood Shores, 2000.

SEIXAS NETO, A; CUNHA, L.M.S.; MEIRA, C. A. A. **Sistema de integração e qualificação de informação para cadeia de frutas**. Campinas: Embrapa Informática Agropecuária, 2000. 19 p. (Embrapa. Programa 14 – Intercâmbio e Produção de Informação em Apoio às Ações de Pesquisa e Desenvolvimento. Projeto 14.2001.368). Projeto em andamento.

TONI, J. A. de. **Definição de um DATA MART em cooperativas agropecuárias**. 2000. 172 f. Dissertação(Mestrado) – Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

Comunicado Técnico, 3

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO
GOVERNO FEDERAL
Trabalhando em todo o Brasil

Embrapa Informática Agropecuária
Área de Comunicação e Negócios
Av. Dr. André Tosello s/nº
Cidade Universitária - "Zeferino Vaz"
Barão Geraldo - Caixa Postal 6041
13083-970 - Campinas, SP
Telefone/Fax: (19) 3789-5743
E-mail: sac@cnptia.embrapa.br

1ª edição

© Embrapa 2001

Comitê de Publicações

Presidente: *Francisco Xavier Hemerly*
Membros efetivos: *Amarindo Fausto Soares, Ivanilde Dispato, Marcia Izabel Fugisawa Souza, José Ruy Porto de Carvalho, Suzilei Almeida Carneiro*
Suplentes: *Fábio Cesar da Silva, João Francisco Gonçalves Antunes, Luciana Alvim Santos Romani, Maria Angélica de Andrade Leite, Moacir Pedroso Júnior*

Expediente

Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel* Fugisawa Souza
Capa: *Intermídia Publicações Científicas*
Editoração Eletrônica: *Intermídia Publicações Científicas*