Atualização Incremental de *Data Warehouse*

Pedro Paulo de Souza Bento da Silva[1] , Pedro Losco Takecian[1], João Eduardo Ferreira[1]

[1]Instituto de Matemática e Estatística da USP.

Resumo

O grupo de Banco de Dados do Instituto de Matemática e Estatística da USP participa de diversos projetos de pesquisa relacionados a análise de dados clínicos e moleculares. Era sabido que utilizar um Data Warehouse para integrar esses dados seria a solução mais adequada. No entanto, devido ao grande volume de dados inconsistentes aliado a grande frequência com que os requisitos do sistema são modificados, optou-se por construir um DW não incremental para cada projeto. Em um DW não incremental sempre que houver necessidade de acrescentar um novo conjunto de dados, é estritamente importante que todo o conteúdo das tabelas seja apagado para que após este procedimento os dados antigos juntamente com os novos sejam reinseridos. O objetivo deste trabalho é desenvolver um DW incremental que integre dados heterogêneos provenientes de hemocentros de âmbito nacional. Esses dados serão usados como fonte para futuras pesquisas (Figura 1). Além disso, para abastecer o DW, foi proposto também o projeto de modelagem e implementação de rotinas de Extração, Transformação e Carga (*Extract, Transform and Load* - ETL) que fossem de fácil manutenção, reutilizáveis e que, além de aplicar corretamente as regras de tratamento sobre os dados, fizessem isso de maneira eficiente. Tendo em vista o tempo limitado, optou-se, inicialmente, por desenvolver e aplicar as rotinas para o Hemocentro de São Paulo/Fundação Pró-Sangue.

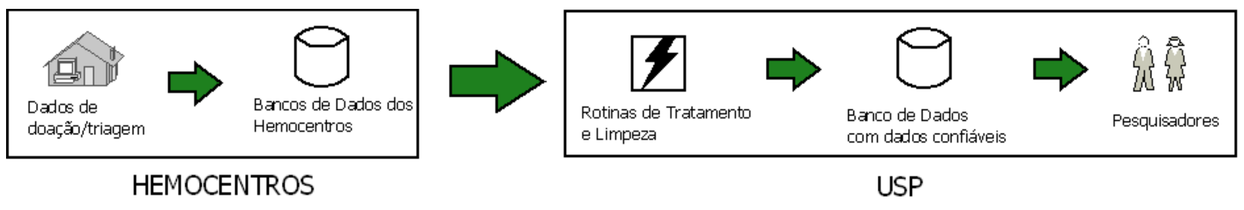

Figura 1. Ciclo de dados aplicado para o Hemocentro de São Paulo. Este esquema demonstra o fluxo de dados desde sua obtenção, passando por rotinas de tratamento e limpeza até o envio dos dados para análises posteriores, que serão feitas por pesquisadores da USP e NIH-USA (*National Institutes of Health*).

Problema e Proposta

Diante do cenário atual (Figura 2), percebeu-se que há grande dificuldade em manter a coleção de rotinas ETL. Isso se deve em geral aos seguintes fatores: grande quantidade de dados; descentralização dos dados dos hemocentros; instabilidade dos requisitos; grande quantidade de dados errados ou inconsistentes. Para solucionar esses problemas faz-se necessário:

1. Modelar um Data Warehouse normalizado e incremental que integre os dados dos hemocentros (Figura 3);

2. Ter grande familiaridade com os dados dos hemocentros, o que possibilita:

   2.1. Projetar um bom modelo conceitual;
   2.2. Modelar e implementar de maneira correta rotinas de transformação, limpeza e consistência de dados;
   2.3. Geração de dados confiáveis para inserção no Data Warehouse (Figura 3).

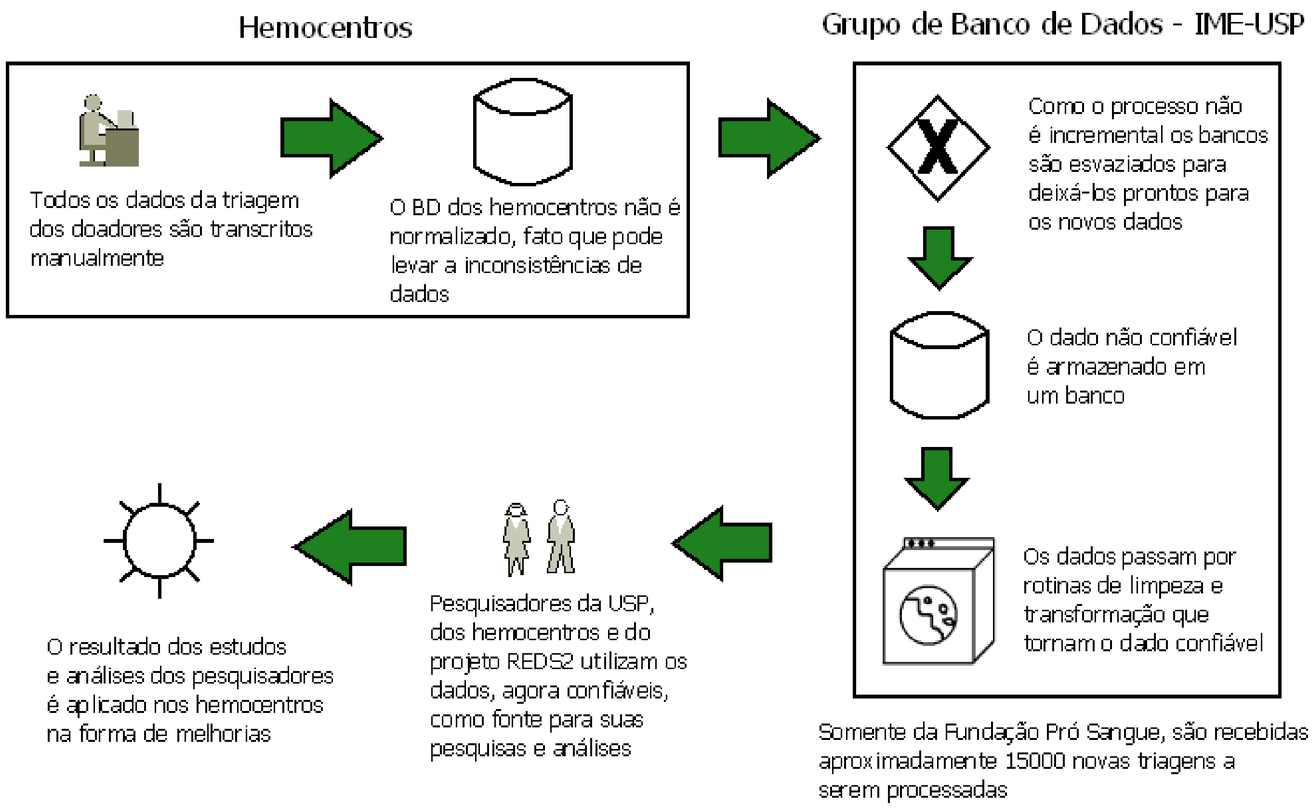

Figura 2. Processo de limpeza e transformação de dados utilizado atualmente.

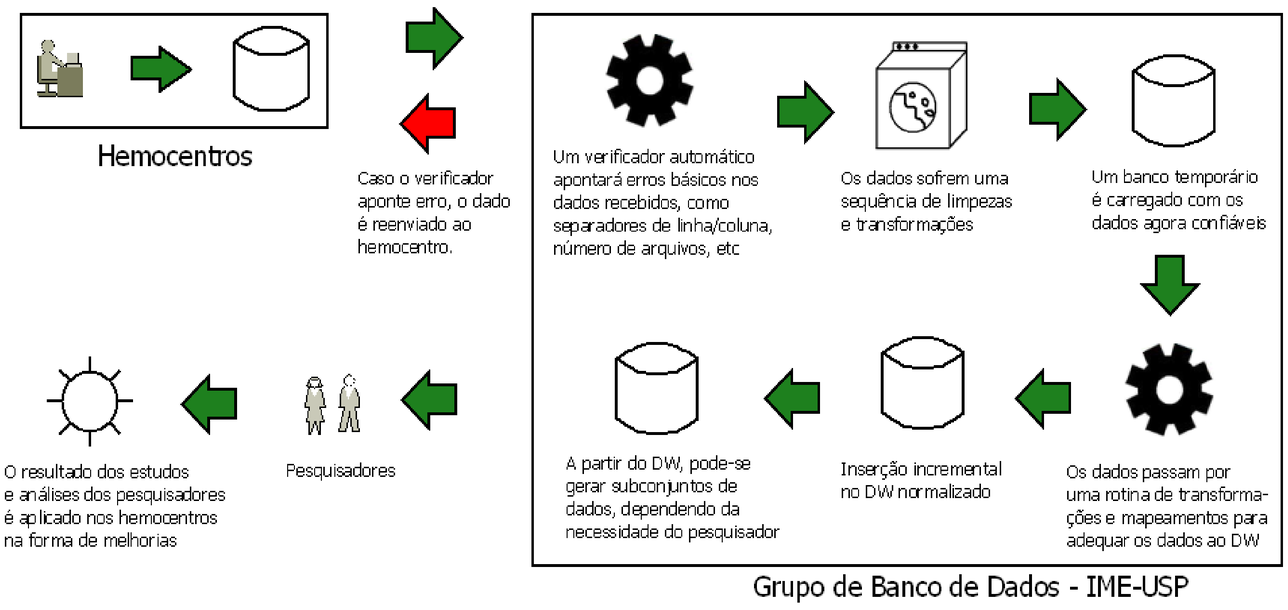

Figura 3. Processo de limpeza e transformação de dados proposto neste trabalho.

Resultados

1. Tecnologia

Neste trabalho, foi usado o pacote Microsoft SQL Server 2008, pois um dos requisitos do sistema era a geração de relatórios dinâmicos que pudessem ser visualizados no Excel. Comparado à versão utilizada anteriormente (Microsoft SQL Server 2000), foi bastante perceptível a melhora em se tratando de reutilização de código, facilidade de manutenção e integração com outras linguagens. Essas facilidades adicionais possibilitaram uma melhoria significativa na qualidade das rotinas criadas, tornando-as mais maleáveis e de fácil modificação.

2. Projeto

Após analisar os requisitos dos pesquisadores e as origens de cada um dos atributos provenientes dos hemocentros, foi possível desenvolver o modelo conceitual do DW normalizado, o modelo de procedência, limpezas e consistências de cada um dos seus atributos.

3. Implementação

- Sistema do DW normalizado;
- Rotinas que populam o DW de forma incremental;
- Rotinas que carregam, limpam, transformam, mapeiam e compõem os dados da Fundação Pró-Sangue (Figura 4);
- Desenvolvimento de rotinas ETL com a ferramenta *Pentaho Data Integration* de código aberto.

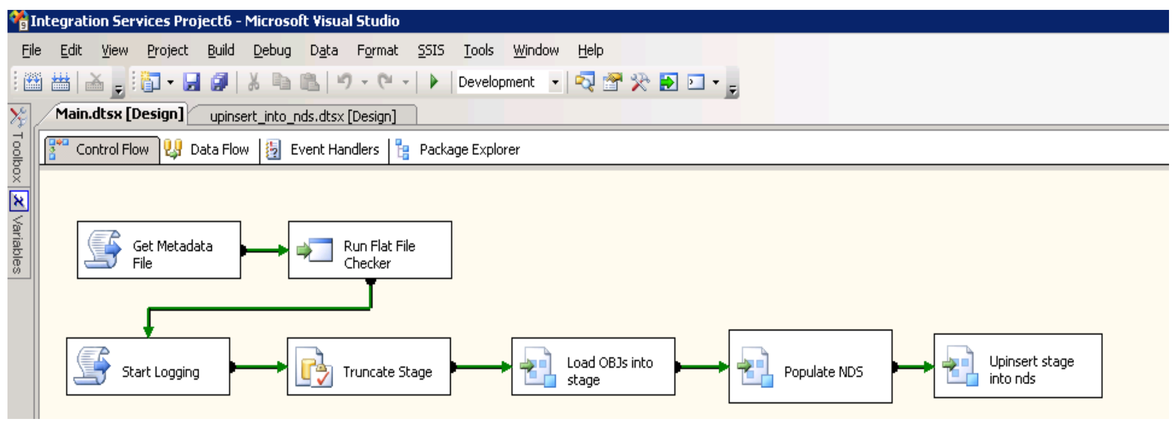

Figura 4. Tela de visualização da rotina principal de limpeza e tranformação dos dados do Hemocentro de São Paulo/Fundação Pró-Sangue.

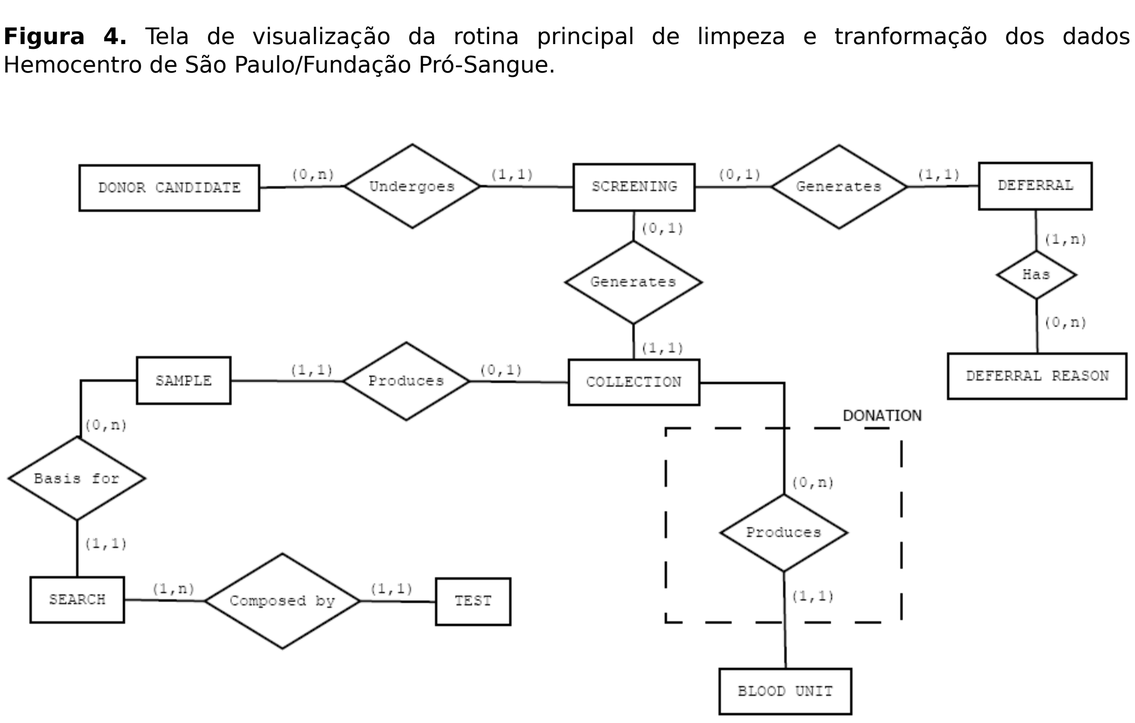

Figura 5. Modelo entidade-relacionamento simplificado (sem atributos) do *Data Warehouse* normalizado.

APOIO

FUNDAÇÃO PRÓ-SANGUE HEMOCENTRO DE SÃO PAULO

NATIONAL INSTITUTES OF HEALTH

Instituto de Matemática e Estatística USP