Revista de Sistemas de Informação n°. 02 Jul./Dez. 2008

DESMISTIFICANDO O CONCEITO DE ETL

Fábio Silva Gomes da Gama e Abreu- FSMA

Resumo

Este artigo aborda os conceitos de ETL (Extract, Transform and Load ou Extração, Transformação e Carga) com o objetivo de desmistificar o uso deste recurso como apenas um sub-processo da construção de um DW (Data Warehouse ou Armazém de Dados).

Palavras-chave: ETL, armazém de dados

Abstract

This article discusses the concepts of ETL (Extract, Transform and Load) with the objective to demystify the use of this feature only as a sub-process of building a DW (Data Warehouse).

Key Words – ETL, Data Warehouse

INTRODUÇÃO

ETL é um tema pouco disseminado nas universidades. Em algumas delas, o aluno se forma sem sequer ter ouvido falar neste conceito.

No entanto, a maioria das universidades ou faculdades – senão todas – que abordam o tema, o fazem como sendo um sub-processo do processo de construção de um DW ou Data Mart.

Não que isto esteja errado, porém não é somente para este fim que o ETL é utilizado. Com isso, o recém formando sai com uma visão limitada do conceito de ETL, o que é ruim para ele e pior ainda para o mercado de trabalho.

O CONCEITO

O processo de ETL (Extract, Transform and Load) destina-se à extração, transformação e carga dos dados de uma ou mais bases de dados de origem para uma ou mais bases de dados de destino (Data Warehouse).

A extração e carga são obrigatórias para o processo, sendo a transformação/limpeza opcional.

ETL LECIONADO NAS UNIVERSIDADES

O processo de ETL (Extract, Transform and Load) é o processo mais crítico e demorado na construção de um Data Warehouse, pois consiste na extração dos dados de bases heterogêneas, na transformação e limpeza destes dados, e na carga dos dados na base do DW.

As decisões gerenciais são tomadas com base nas informações geradas pelas ferramentas do tipo front-end. Estas informações são geradas através dos dados armazenados no Data Warehouse. Se estes dados não forem corretamente trabalhados no processo de extração, as informações geradas através deles farão com que decisões sejam tomadas erroneamente, podendo afetar diretamente os negócios da organização. Portanto, os dados devem representar a verdade, a mais pura verdade, nada mais que a verdade (KIMBALL, 1998 apud ABREU, 2007). A maior parte do esforço exigido no desenvolvimento de um DW é consumido neste momento e não é incomum que oitenta por cento de todo esforço seja empregado no processo de ETL, (INMON, 1997 apud ABREU, 2007).

Somente a extração dos dados leva mais ou menos 60 por cento das horas de desenvolvimento de um DW (KIMBALL, 1998 apud ABREU, 2007).

Esta etapa do processo deve se basear na busca das informações mais importantes em sistemas fontes ou externos e que estejam em conformidade com a modelagem do DW. Tal busca de dados pode ser obstruída por problemas como a distribuição das origens dos dados, que podem estar em bases distintas com plataformas diferentes gerando a demanda de utilização de formas de extração diferentes para cada local (ALMEIDA, 2006 apud ABREU, 2007).

No momento de criação do DW é comum uma carga de dados inicial que faça com que a extração busque todos os dados dos sistemas fontes, mas com o decorrer do tempo a extração deve estar preparada apenas para fazer cargas incrementais. A carga incremental que carrega apenas os registros que foram alterados ou inseridos desde a carga inicial é muito mais eficiente (KIMBALL, 1998 apud ABREU, 2007).

A transformação dos dados é a fase subseqüente à sua extração. Esta fase não só transforma os dados, mas também realiza a limpeza dos mesmos. A correção de erros de digitação, a descoberta de violações de integridade, a substituição de caracteres desconhecidos, a padronização de abreviações podem ser exemplos desta limpeza (GONÇALVES, 2003 apud ABREU, 2007). Segundo Kimball (1998), as características mais relevantes para garantir a qualidade dos dados são:

- unicidade, evitando assim duplicações de informação;
- precisão. Os dados não podem perder suas características originais assim que são carregados para o DW;
- completude, não gerando dados parciais de todo o conjunto relevante às análises; e
- consistência, ou seja, os fatos devem apresentar consistência com as dimensões que o compõem.

É necessário que os dados fiquem em uma forma homogênea para serem carregados no DW.

Durante o processo de homogeneização, são encontrados muitos conflitos de modelagem. Estes conflitos podem ser divididos em semânticos e estruturais.

Os conflitos semânticos são todos aqueles que envolvem o nome ou a palavra associada às estruturas de modelagem, por exemplo, mesmo nome para diferentes entidades ou diferentes nomes para a mesma entidade. Já os conflitos estruturais englobam os conflitos relativos às estruturas de modelagem escolhidas, tanto no nível de estrutura propriamente dita como no nível de domínios. Os principais tipos de conflitos estruturais são aqueles de domínio de atributo que se caracterizam pelo uso de diferentes tipos de dados para os mesmos campos (GONÇALVES, 2003 apud ABREU, 2007).

De acordo com Gonçalves (2003), os conflitos típicos de domínio de atributo são:

- diferenças de unidades: quando as unidades utilizadas diferem, embora forneçam a mesma informação (exemplo: distância em centímetros ou polegadas);
- diferenças de precisão: quando a precisão escolhida varia de um ambiente para outro (exemplo: o custo do produto é armazenado com duas posições ‘0,12’ ou com seis posições decimais ‘0,123456’);
- diferenças em códigos ou expressões: quando o código utilizado difere um do outro (exemplo: sexo representado por M ou F e por 0 ou 1);
- diferenças de granularidade: quando os critérios associados a uma informação, embora utilizando uma mesma unidade, são distintos (exemplo: quando horas trabalhadas correspondem às horas trabalhadas na semana ou às horas trabalhadas no mês);
- diferenças de abstração: quando a forma de estruturar uma mesma informação segue critérios diferentes (exemplo: endereço armazenado em um único atributo, ou subdividido em rua e complemento).

Depois de identificados os conflitos de modelagem, devem-se criar as regras de conversão para os padrões estabelecidos pelo Data Warehouse (GONÇALVES, 2003 apud ABREU, 2007). Essas regras podem ser criadas com o auxílio de ferramentas de integração utilizadas para o processo de extração e carga de dados. Após a criação das regras, a etapa de carga dos dados pode ser planejada.

Segundo Almeida (2006), basicamente são carregadas as dimensões estáticas, de modificação lenta ou remanescente e fatos integrantes ao modelo do DW. Este processo pode ter alto custo de processamento alem de implicar em tempo de carga que na maioria das vezes não pode ser extenso devido à utilização contínua do DW. Assim, algumas precauções podem ser tomadas antes de se iniciar a carga dos dados, como:

- desligamento de índices e referências de integridade (isso pode prejudicar na qualidade dos dados pois apesar de diminuir o processamento, os dados não são validados no momento da inserção);
- utilização de comandos do tipo TRUNCATE ao invés de DELETE pois nos SGBDs mais atuais este recurso não gera armazenamento de informações em áreas de recuperação de dados;
- ter a consciência de que no momento da carga alguns dados não serão carregados e deste modo os mecanismos do processo devem dar suporte a auditorias de carga para que a mesma possa ser reiniciada no momento em que foi parada e a possibilidade de manter logs com os dados rejeitados para a avaliação dos motivos pelo qual não foram carregados e assim ajustados para integrarem o conjunto a ser carregado.

Dimensões estáticas normalmente não oferecem problemas, pois estas mantêm dados que não sofrem alteração na sua origem e serão carregados uma única vez, assim como as remanescentes que normalmente são originadas de esforço manual na sua confecção, por exemplo, as planilhas eletrônicas. Já as dimensões de modificação lenta necessitam da verificação em suas fontes e nas auditorias das cargas para que se possa identificar qual o momento seguinte depois da última carga que deve iniciar o processo, gerando processamento na leitura de logs de sistemas operacionais e comparação de atributos, podendo então ser necessário sobrescrever todo o conteúdo de um registro, gerar um novo registro na dimensão ou criar um atributo a mais para armazenar o valor antigo (KIMBALL, 1998 apud ABREU, 2007).

Após as dimensões estarem corretamente carregadas, já é possível iniciar a carga dos fatos, que depois de modelados para conter apenas os dados de importância para a organização, direcionam quais regras serão utilizadas como, por exemplo, filtros do que será inserido ou somas a serem realizadas, provocando o aparecimento de regras que passaram despercebidas no início da modelagem.

No entanto, os fatos demandam cuidados na sua carga como o uso das chaves artificiais das dimensões para que se tenha uma integridade referencial, controle de valores nulos obtidos no momento da transação para que não gerem a falta de integridade referencial como datas que, estando nulas, invalidarão o histórico do fato. Técnicas para amenizar o processo devido ao grande volume de dados podem ser usadas, como a carga incremental dos fatos, que irá carregar apenas dados novos ou alterados, execução do processo em paralelo e em momentos de pouco ou nenhum uso do SGBD e a utilização de tabelas auxiliares que serão renomeadas como definitivas ao fim da carga (KIMBALL, 1998 apud ABREU, 2007).

MAS NÃO É SÓ ISSO

Como visto no item anterior, o conceito de ETL que é lecionado está interligado aos sistemas OLAP (On-Line Analytical Processing ou Processamento Analítico On-Line). Porém, antes mesmo do conceito de OLAP existir, o processo de ETL já era

utilizado pelos sistemas OLTP (On-Line Transaction Processing ou Processamento de Transações em Tempo-Real).

Um exemplo bem comum de utilização do ETL para sistemas transacionais é o uso de informações de bases corporativas para estes sistemas.

Um exemplo bem simples pode ser visto logo abaixo:

Uma empresa possui uma base de dados corporativa com o cadastro de todos os seus funcionários.

CADASTRO_DE_FUNCIONARIOS				
MATRICULA	NOME	ENDERECO	TELEFONE	SETOR
4568937	Cláudia Ferreira Gusmão	Rua A, 203	35840074	Financeiro
2893890	Juliana Torres de Almeida	Rua B, 21	78938830	Administrativo
3990487	Marcela de Sousa Aguiar	Rua F, 111	78307382	Financeiro

Quadro 1 – Parte da estrutura da base corporativa de funcionários

Um sistema financeiro da empresa está sendo desenvolvido e necessita carregar para sua base o último nome e a matrícula dos funcionários do setor de finanças. A figura abaixo exemplifica como o processo de ETL funcionaria para este caso.

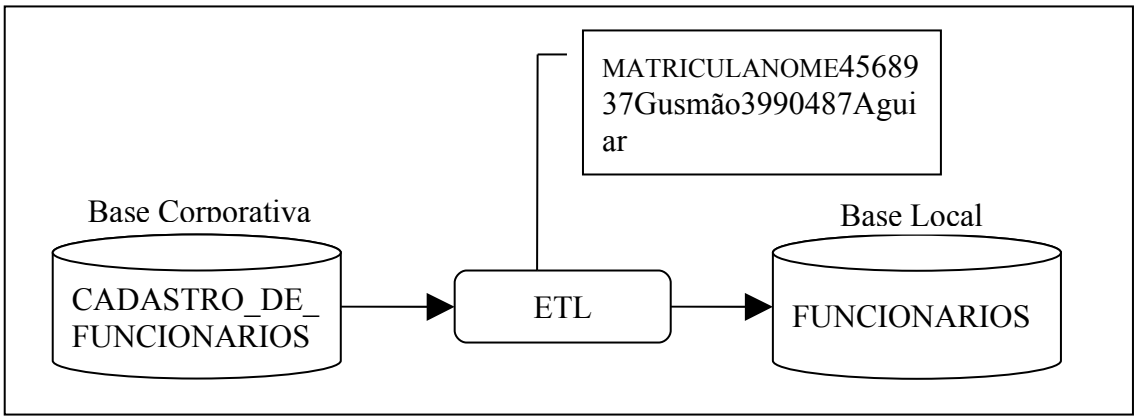

Figura 1 – Processo ETL

Situações como a do exemplo acima, e outras muito mais complexas acontecem o tempo todo no desenvolvimento de sistemas transacionais. Com isto, as ferramentas de integração de dados se tornam de extrema importância para o desenvolvimento dos processos de ETL da organização, pois otimizam o tempo nesta fase do projeto e possuem, dentre outras coisas, recursos de auditoria que ajudam na correção e investigação de possíveis problemas nas regras do ETL. No entanto, a aquisição de uma ferramenta de integração de dados só é cogitada por empresas de pequeno e médio porte quando se pensa no desenvolvimento de um Data Warehouse ou de um Data Mart.

Talvez isso aconteça pelo fato de os profissionais que dão suporte técnico a este tipo de decisão serem aqueles mesmos que saíram da universidade com o conceito limitado sobre ETL.

É claro que o custo de uma ferramenta de integração de dados é muito alto, e muitas destas empresas não tem capacidade financeira para adquiri-la. Porém existem ferramentas de integração de dados que são Open Source (Código Aberto) e que podem

atender com eficiência às demandas de ETL dessas empresas, como o software Talend Open Studio, por exemplo.

CONCLUSÃO

O tema ETL não deve ser pensado apenas como parte de um projeto de BI (Business Intelligence ou Inteligência de Negócio) ou como um sub-processo na construção de um DW. Ele é muito mais que isso. Se a organização possui uma área de TI que possui sistemas transacionais que necessitam migrar informações entre si, o uso do ETL com o auxílio de ferramentas de integração de dados já pode ser planejado e implementado.

Se as universidades ou faculdades trabalharem o conceito de ETL abrangendo todo seu campo de atuação, o mercado de trabalho terá profissionais com uma visão mais ampla do assunto, podendo otimizar com mais eficiência os processos de TI dentro das organizações.

Referências bibliográficas

ABREU, Fábio Silva Gomes da Gama e. **Estudo de usabilidade do software Talend Open Studio como ferramenta padrão para ETL dos sistemas-clientes da aplicação PostGeoOlap. 2007**. Monografia (Graduação em Sistemas de Informação) – Faculdade Salesiana Maria Auxiliadora, Macaé, 2007.

ALMEIDA, Alexandre Marques de. **Proposição de indicadores para avaliação técnica de projetos de Data Warehouse: um estudo de caso no Data Warehouse da plataforma Lattes.** 2006. Monografia (Pós-Graduação em Engenharia de Produção) – Universidade Federal de Santa Catarina, Florianópolis, 2006.

GONÇALVES, Marcio. **Extração de dados para Data Warehouse.** Rio de Janeiro: Axcel Books, 2003.

INMON, W. H.; HACKATHORN, Richard D. **Como Usar o Data Warehouse.** Tradução: Olávio Faria. Rio de Janeiro: Infobook, 1997.

KIMBALL, Ralph. **Data Warehouse Toolkit.** Tradução Mônica Rosemberg; Revisão Técnica Ronal Stevis Cassiolato. São Paulo: Makron Books, 1998.