ADMINISTRAÇÃO DE BANCO DE DADOS

Márcio Motta

OBJETIVOS DE APRENDIZAGEM

Ao final deste texto, você deve apresentar os seguintes aprendizados:

- Conhecer o conceito de ETL
- Identificar as funções de Extração, Transformação e Carga
- Reconhecer os objetivos e vantagens da utilização de ETL em um D*ata Warehouse*

INTRODUÇÃO

Este texto tem como objetivo demonstrar o funcionamento e a utilização de um processo de ETL em sistemas *Data Warehouse*. ETL (*Extract, Transform and Load*) são procedimentos responsáveis pela extração de dados de várias fontes, a sua limpeza, otimização e inserção desses dados num *Data Warehouse.*

O processo ETL é uma das fases mais críticas na construção de um *Data Warehouse,* pois é nesta fase que grandes volumes de dados são processados. Será abordado de forma direta, o seu conceito, características, o modo como este processamento ocorre e ainda, as vantagens da sua utilização em um *Data Warehouse.*

APRESENTAÇÃO

ETL, do inglês, *Extract, Transform and Load* ou em português, Extração, Transformação e Carga consiste no trabalho com dados de diferentes fontes. A maioria dessas fontes tendem a ser bancos de dados relacionais, mas podem vir de outros tipos de fontes, como bases de dados Access, Planilhas de dados (Excel) arquivos de textos separados por vírgulas (CSV) e arquivos XML.

Um sistema ETL precisa ser capaz de se comunicar com essas bases de dados bastante heterogêneas, seja pelo seu formato, seja pela normatização dos dados. O ETL fará a transformação das informações facilitando a interpretação e análise desses dados, após ser armazenada esta informação fica disponível no *Data Warehouse* para consultas que visam auxiliar na tomada de decisão.

FASES DO ETL

O ETL é um processo que se divide em três fases:

Extração

É a coleta de dados dos sistemas de origem (também chamados *Data Sources* ou sistemas operacionais), extraindo-os e transferindo-os para o ambiente do *Data Warehouse*, onde o sistema de ETL pode operar independente dos sistemas operacionais.

Transformação

depois que os dados são extraídos é necessário tratá-los, efetuando a limpeza e a filtragem a fim de garantir sua integridade por meio de programas ou rotinas especiais que tentam identificar anomalias. Caso as encontre, é necessário resolvê-las antes de serem inseridas do *Data Warehouse.*

Correção de erros de digitação, a descoberta de violação de integridade, a substituição de caracteres desconhecidos, a padronização de abreviações podem ser exemplos de limpeza de dados. Pode-se encontrar conflitos de semântica nos dados, que são divididos de duas formas: conflitos semânticos e conflitos estruturais.

Conflitos Semânticos - 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.

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.

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).

Carga dos dados

Consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o modelo dimensional. Dependendo das necessidades da organização, este processo varia amplamente. Alguns *Data Warehouses* podem substituir as informações existentes semanalmente, com dados cumulativos e atualizados, ao passo que outros *Data Warehouses* podem adicionar dados a cada hora. A latência e o alcance de reposição ou acréscimo constituem opções de projeto estratégicas que dependem do tempo disponível e das necessidades de negócios.

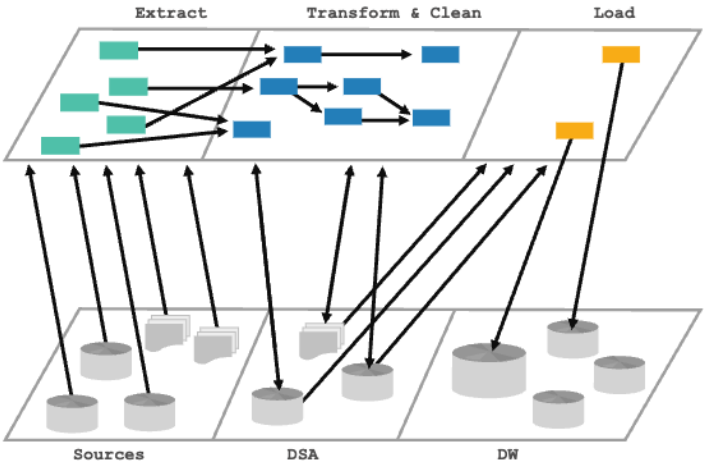
Figura nº 1 – Esquema da Infraestrutura de um sistema ETL

Fonte: Autor

Quando o *Data Warehouse* se encontra construído, uma das ferramentas mais utilizadas para o acesso e a análise dos dados é o *Online Analytical Processing* (OLAP). Através desta ferramenta é possível realizar o tratamento dos dados provenientes de diferentes fontes em tempo real, utilizando métodos mais rápidos e eficazes. Permite também usar uma grande variedade de ferramentas de visualizações dos dados e organizá-los através dos critérios de seleção pretendidos. A maior vantagem do OLAP é, no entanto, a capacidade de realizar análises multidimensionais dos dados, associadas a cálculos complexos, análises de tendências e modelação.

A Figura 2 descreve de forma geral o processo de ETL. A camada inferior representa o armazenamento dos dados que são utilizados em todo o processo. No lado esquerdo pode-se observar os dados originais provenientes, na maioria dos casos, de Bancos de Dados ou de arquivos com formatos heterogêneos (CSV ou XML). Os dados provenientes destas fontes são obtidos (como é ilustrado na área superior esquerda da Figura 2), por rotinas de extração que fornecem informação igual ou modificada, relativamente à fonte de dados original. Posteriormente, esses dados são propagados para a *Data Staging Area* (DSA) onde são transformados e limpos antes de serem carregados para o Data Warehouse. O *Data Warehouse* é representado na parte direita da figura e tem como objetivo o armazenamento dos dados. O carregamento dos dados no *Data Warehouse* é realizado através das atividades de carga representadas na parte superior direita da figura (*Load*).

Figura nº 2 – Ilustração do Processo de ETL



Fonte: Autor

CARACTERÍSTICAS

ETL, do inglês, *Extract, Transform and Load* ou em português, Extração, As ferramentas de ETL existentes encontram-se preparadas para o processo de extração, transformação e carga. Tem-se assistido a inúmeros avanços nessas ferramentas desde 1990, estando atualmente mais direcionadas para o usuário.

Uma boa ferramenta de ETL deve ser capaz de comunicar com as diversas BD e ler diferentes formatos. Atualmente há uma elevada oferta de opções, como registrado na Tabela .

Ferramentas ETL	Versão	Distribuidor
Adeptia Integration Server	4.9	Adeptia
Clover ETL	2.9.2	Javlin
Data Integrator	9.2	Pervasive
Data Integrator & Data Services	XI 3.0	SAP Business Objects
Data Manager/Decision Stream	8.2	IBM
Data Migrator	7.6	Information Builders
DataFlow Manager	6.5	Pitney Bowes Business
DB2 Warehouse	9.1	IBM
Elixir Repertoire	7.2.2	Elixir
ETL4ALL	4.2	IKAN
IBM Information Server (Datastage)	8.1	IBM
Open Text Integration Center	7.1	Open Text
Oracle Warehouse Builder (OWB)	11 R1	Oracle
Pentaho Data Integration	3.0	Pentaho
PowerCenter	9.0	Informatica US
SQL Server Integration Services	10	Microsoft
Talend Open Studio & Integration Suite	4.0	Talend
Transformation Manager	5.2.2	ETL Solutions Ltd.

VANTAGENS DA UTILIZAÇÃO DE ETL

- Maior garantia da qualidade dos dados: ferramentas de ETL podem disponibilizar meios para trabalhar a qualidade dos dados através de algoritmos complexos (lógica fuzzy, IA, etc).
- Separação entre funcionalidade e manipulação de dados: uma ferramenta de ETL já possui suas funcionalidades disponíveis (Lookup, Merge, Split, Expressões calculadas, etc). Só é necessário concentrar-se em como fluir os dados dentro da carga e não codificar cada tarefa da carga.
- Desenvolvimento das cargas: desenvolver uma rotina de carga em uma ferramenta de ETL é muito mais fácil e rápido que codificá-la. Dependendo da facilidade da ferramenta é possível inclusive que usuários não técnicos a utilizem para cargas mais simples.
- Manutenção das cargas: as tarefas de manutenção de uma rotina de carga são mais fáceis de realizar em relação à manutenção de código.
- Manutenção de Metadados: os metadados são gerados e mantidos automaticamente com a ferramenta evitando que problemas de conversão gerem dados não íntegros ao final do processo. A manutenção de metadados também evita ou alerta para alterações de esquema que invalidem a carga.
- Desempenho: as ferramentas de ETL utilizam métodos mais performáticos para trabalhar com grandes volumes e normalmente conseguem extrair, transformar e carregar dados com mais velocidade e menos utilização de recursos. Isso inclui operações não logadas, gravações em bloco, etc.
- Portabilidade: ferramentas de ETL podem ser transferidas de servidor mais facilmente e até eventualmente distribuir sua carga entre vários servidores.
- Múltiplas Conexões: a conexão de uma ferramenta de ETL com múltiplas fontes de dados é transparente. Caso apareça alguma fonte não trivial como o SAP, Mainframe e VSAM, é possível adquirir o conector sem a necessidade de codificar um.
- Reinicialização: ferramentas de ETL possuem a capacidade de reiniciar a carga de onde pararam sem a necessidade de codificar essa inteligência.
- Segurança: é possível tornar a segurança mais modular dividindo-se os papéis (criação de cargas, execução de cargas, agendamento, etc).

CONCLUSÃO

As ferramentas de ETL trabalham diretamente em conjunto com os conceitos de *Business Intelligence*, com o objetivo de transformar grandes quantidades de dados em informações de qualidade, para a tomada de decisão, de modo a possibilitar uma visão sistêmica do negócio e auxiliar na distribuição uniforme dos dados entre os analistas de negócios.

Sem as rotinas de ETL seria necessário disponibilizar uma grande quantidade de tempo na interpretação e filtragem dos dados contidos em diversas bases e um grande conhecimento técnico por parte dos analistas. Com todas as rotinas de transformação, a análise de dados se torna muito mais voltada para a parte gerencial e não de forma técnica, pois os dados são relatados já de forma clara, coerente e organizada. A definição das regras de transformação e detecção dos dados tomam a maior parte do projeto (na casa de 80%), tornando extremamente simples e direta para qualquer tipo de usuário a interpretação dos dados.

REFERÊNCIAS

KIMBALL, R. & ROSS, M. - **The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling** (Second Edition), Ed. John Wiley & Sons, ISBN 0-471-20024-7. 2002.

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

Conteúdo: