O Processo ETL em Sistemas *Data Warehouse*

João Ferreira, Miguel Miranda, António Abelha e José Machado

Universidade do Minho, Departamento de Informática,
Braga, Portugal
tiago_jtx@hotmail.com
{miranda,abelha,jmac}@di.uminho.pt
http://www.di.uminho.pt

Resumo. Extração, Transformação e Carga (*Extract Transform Load* - ETL) são procedimentos de uma técnica de *Data Warehouse* (DW), que é responsável pela extracção de dados de várias fontes, a sua limpeza, optimização e inserção desses dados num DW. Este artigo tem como objectivo demonstrar o funcionamento genérico do processo ETL em sistemas DW. O processo ETL é uma das fases mais críticas na construção de um sistema DW, pois é nesta fase que grandes volumes de dados são processados. Será abordado de forma sucinta, o modo como este processamento ocorre, e ainda, as ferramentas de ETL disponíveis no mercado. Por fim, serão abordados quais os critérios a ter em consideração na escolha de uma destas ferramentas.

Palavras-chave: *Extract Transform Load* (ETL), *Data Warehouse* (DW), Ferramentas ETL .

1 Introdução

A ideia principal de um sistema de *Data Warehouse* (DW) (ilustrado na figura 1), consiste em agregar informação proveniente de uma ou mais Bases de Dados (BD), ou de outras fontes, para posteriormente a tratar, formatar e consolidar numa única estrutura de dados. Um sistema DW está associado a BD com um grande volume de dados devido quer ao volume proveniente das fontes heterogéneas quer da baixa normalização habitualmente utilizada. A estrutura de dados do DW é desenvolvida de forma a facilitar a análise desses dados. Após ser armazenada, estas informação, fica disponivel no DW ou em *DataM*arts (DM) para consultas que visam ajudar na tomada de decisão. Devido ao custo elevado, o DW muitas vezes é dividido em partes menores, nomeadamente os DM. Um DM consolida apenas as informações de uma determinada área e após a sua criação podem se unir vários DM para formarem um único DW [1].

INForum 2010 - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, pp. 757–765

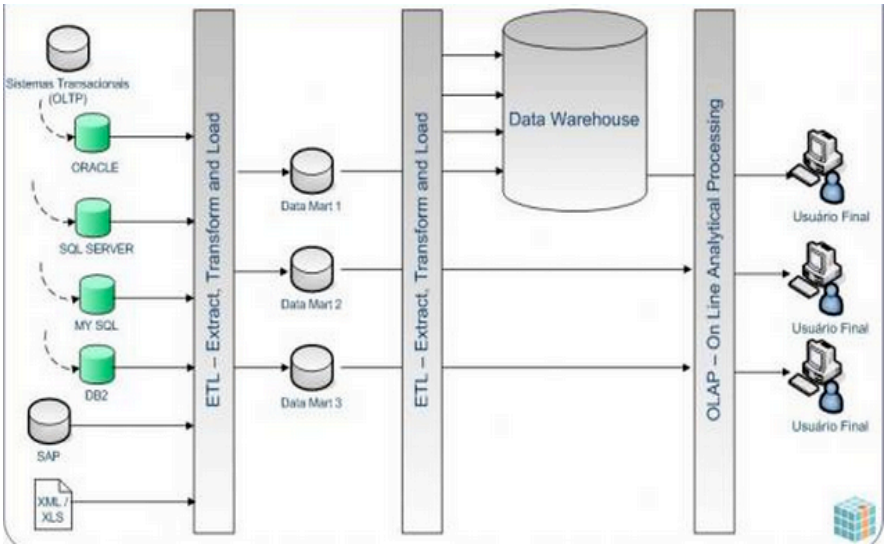

Figura 1. Esquema da Infra-estrutura de um sistema DW [1]

Para a construção de um DW são necessários diferentes passos principalmente ao nível da extracção e processamento de dados. O processo ETL destina-se à extracção e transformação dos dados e termina com a inclusão destes no DW. Esta fase caracteriza-se por englobar procedimentos de limpeza, integração e transformação de dados. Segundo a literatura este é o processo mais crítico e demorado na construção de um DW [1].

Quando o DW 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 proveniente 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 selecção 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 [3,2].

2 O Processo ETL

O ETL é um processo para extrair dados de um sistema de Bases de Dados (BD), sendo esses dados processados, modificados, e posteriormente inseridos numa outra BD. Estudos relatam que o ETL e as ferramentas de limpeza de dados consomem um terço do orçamento num projecto de DW, podendo, no que respeita ao tempo de desenvolvimento de um projecto de DW, chegar a consumir 80% desse valor. Outros estudos mencionam, ainda, que o processo de ETL tem custos na ordem dos 55% do tempo total de execução do projecto de DW [4,5,6].

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 BD ou, então, de ficheiros com formatos heterogéneos, por exemplo de texto. Os dados provenientes destas fontes são obtidos (como é ilustrado na área superior esquerda da figura 2), por rotinas de extracção 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 DW. O DW é representado na parte direita da figura e tem como objectivo o armazenamento dos dados. O carregamento dos dados no DW, é realizado através das actividades de carga representadas na parte superior direita da figura.

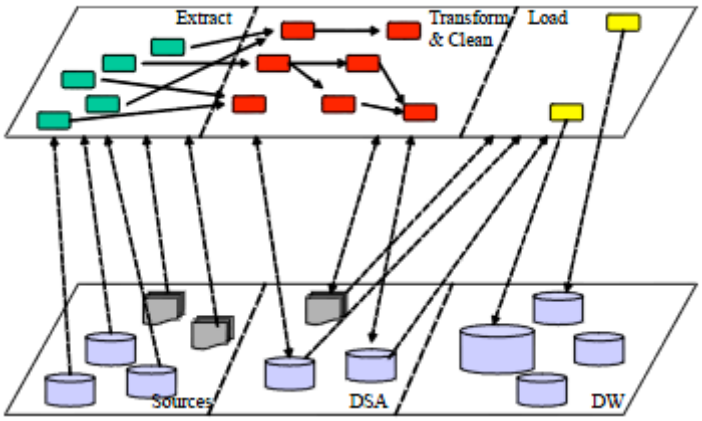

Figura 2. Ilustração do processo de ETL [13].

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

1. Extração;
2. Transformação;
3. Carga.

Segundo alguns autores a concepção de um processo ETL incide sobre o mapeamento dos atributos dos dados de uma ou várias fontes para os atributos das tabelas do DW [7,8].

2.1 Utilização do processo ETL em BD e Ferramentas disponíveis

No DW, os dados normalmente utilizados estão localizados em BD multidimensionais. É importante que se tenha consciência que as alterações nos dados

não afectam as fontes originais, mas sim, os dados no momento de extracção para o repositório da DW. Mais ainda, que os ajustes são modelados de acordo com as necessidades do modelo de DW, atendendo assim às restrições que são necessárias para esse modelo [12].

Depois do processo de transformação ocorre o processo de carga. Neste processam-se os mapeamentos sintácticos e semânticos entre os esquemas, respeitando as restrições de integridade e criando assim uma visão concretizada e unificada das fontes. Este processo é dos mais árduos e complexos de obter devido a sua complexidade que dependerá da heterogeneidade das BD [10] [11].

No mercado existem muitas ferramentas capazes de executar processos de ETL, a tabela 1 apresenta uma visão geral da evolução destas ferramentas [3].

Tabela 1. As várias gerações de ETL ao longo dos anos

Ano	**Título**	**Significado**
Início de 1990	Codificação manual de ETL	Códigos personalizados escitos à mão
1993-1997	A primeira geração de ferramentas de ETL	Código baseado em ferramentas de ETL
1999-2001	Segunda geração de ferramentas de ETL	Código baseado em ferramentas de ETL
2003-2010	Ferramentas de ETL actualmente	A maioria das ferramentas eficientes

As ferramentas de ETL disponíveis actualmente encontram-se bem preparadas para o processo de extracção, transformação e carga. Tem-se assistido a inúmeros avanços nestas ferramentas desde 1990, estando actualmente mais direccionadas para o utilizador [3].

Uma boa ferramenta de ETL deve ser capaz de comunicar com as diversas BD e ler diferentes formatos. Actualmente a oferta é elevada, como registado na tabela 2.

Tabela 2. Diferentes ferramentas de ETL

Lista de ferramentas ETL	Versão	ETL vendedores
Oracle Warehouse Builder (OWB)	11gR1	Oracle
Data Integrator & Data Services	XI 3.0	SAP Business Objects
IBM Information Server (Datastage)	8.1	IBM
PowerCenter	9.0	Informatica
Elixir Repertoire	7.2.2	Elixir
Data Migrator	7.6	Information Builders
SQL Server Integration Services	10	Microsoft
Talend Open Studio & Integration Suite	4.0	Talend
DataFlow Manager	6.5	Pitney Bowes Business Insight
Data Integrator	9.2	Pervasive
Open Text Integration Center	7.1	Open Text
Transformation Manager	5.2.2	ETL Solutions Ltd.
Data Manager/Decision Stream	8.2	IBM (Cognos)
Clover ETL	2.9.2	Javlin
ETL4ALL	4.2	IKAN
DB2 Warehouse	9.1	IBM
Pentaho Data Integration	3.0	Pentaho
Adeptia Integration Server	4.9	Adeptia

A selecção de uma ferramenta de ETL adequada é uma decisão muito importante a ser tomada. A ferramenta de ETL opera no núcleo do DW, com a extracção de dados de múltiplas fontes e a sua transformação. Estas características tornam-na numa ferramenta acessível para os analistas de sistemas de informação.

Ao contrário de outros componentes de uma arquitectura de *Data Warehousing*, é muito difícil mudar de uma ferramenta ETL para outra, devido à falta de normas, definições de dados e regras de transformação.

Ao seleccionar uma ferramenta de ETL devem ser tomados em consideração os seguintes pontos [9]:

- Suporte à plataforma: Deve ser independente de plataforma, podendo assim correr em qualquer uma.
- Tipo de fonte independente: Deve ser capaz de ler directamente da fonte de dados, independentemente do seu tipo, saber se é uma fonte de RDBMS (*Relational Database Management System*), ficheiro simples ou um ficheiro XML.
- Apoio funcional: Deve apoiar na extracção de dados de múltiplas fontes, na limpeza de dados, e na transformação, agregação, reorganização e operações de carga.
- Facilidade de uso: Deve ser facilmente usada pelo utilizador.
- Paralelismo: Deve apoiar as operações de vários segmentos e execução de código paralelo, internamente, de modo que um determinado processo pode tirar proveito do paralelismo inerente da plataforma que está sendo executada. Também deve suportar a carga e equilíbrio entre os servidores e capacidade de lidar com grandes volumes de dados. Quando confrontados com cargas muito

elevadas de trabalho, a ferramenta deve ser capaz de distribuir tarefas entre múltiplos servidores.

- Apoio ao nível do *debugging*: Deve apoiar o tempo de execução e a limpeza da lógica de transformação. O utilizador deve ser capaz de ver os dados antes e depois da transformação.
- Programação: Deve apoiar o agendamento de tarefas ETL aproveitando, assim, melhor o tempo não necessitando de intervenção humana para completar uma tarefa particular. Deve também ter suporte para programação em linha de comandos usando programação externa.
- Implementação: Deve suportar a capacidade de agrupar os objectos ETL e implementa-los em ambiente de teste ou de produção, sem a intervenção de um administrador de ETL.
- Reutilização: Deve apoiar a reutilização da lógica de transformação para que o utilizador não precise reescrever, várias vezes, a mesma lógica de transformação outra vez.

3 Caso de estudo

Na sequência da necessidade de validar os dados dos recursos humanos de um centro hospitalar português foi extraída a informação dos seus repositórios para um ambiente de *data warehouse*. A ferramenta escolhida para o tratamento de dados e construção do repositório foi a *release* 2 da *Oracle Database* 11g, que possui embebida em si a plataforma de desenvolvimento de *data warehouse* denominada *Oracle Warehouse Builder*. A fonte principal era uma instância *Oracle 8i*, na qual estavam integrados em diferentes perfis dados de recursos humanos e outros sistemas como o de controlo de ponto.

A informação encontrava-se dispersa em mais de uma centena de tabelas com registos processados e a processar. A dispersão de informação obrigou a alterar a fundo o esquema normal de destino procurando uma normalização de nível mais baixo para a construção dos diferentes *data marts*. Desta forma foram necessários desenvolver métodos para o ETL do repositório dos recursos humanos que garantissem a qualidade da informação e permitissem a construção de um novo repositório que fosse mais adequado para *alimentar* a DW.

Nesta fase tentou-se garantir que toda a informação estava correcta e consistente, teve-se algum receio que dados incorrectos pudessem conduzir a erros críticos de tomada de decisão. Dada esta importância de detecção de erros serão de seguida explicitados alguns objectivos de teste que se estabelecem para o sistema ETL:

3.1 Preenchimento de dados

Neste teste procura-se assegurar que todos os dados esperados eram carregados.

- Comparam-se o número de registos entre os dados das fontes e o número de registos carregados para o DW.

- Comparam-se valores únicos de determinados atributos entre as fontes e os dados carregados para o DW.
- Procura-se fazer um bom esquema de dados para perceber as limitações dos valores atribuídos.
- Procura-se validar os conteúdos de cada atributo, ou seja, não permitir que por razões de codificação o limite de caracteres entre cada esquema relacional (fonte e destino) não resulta na falha do fluxo de dados.
- Transformação de Dados - Neste teste tenta-se assegurar que os dados são transformados correctamente de acordo com as regras de negócio especificadas.
- Procuram-se criar testes, os mais diversos possíveis para antever algumas situações consequentes.
- Tenta-se validar o processamento correcto de campos no ETL tais como chaves estrangeiras.
- Procura-se verificar sempre se os tipos de dados presentes no DW são os que se tinham planeado.
- E ainda procura-se testar a integridade referencial entre as tabelas.

3.2 Qualidade de dados

Neste teste procura-se assegurar que o sistema ETL rejeita ou substituí valores por defeito, corrige ou ignora dados e reporta dados inválidos.

- Procura-se realizar as conversões dos dados sempre correctamente.
- Nos casos de atributos NULL procura-se sempre inserir valores equivalentes a "desconhecido".
- Sempre que algum atributo não está correcto procura-se validar e corrigir o problema.
- Sempre que aparecem valores duplicados analisam-se os códigos e corrige-se o problema

3.3 Performance e Escalabilidade

Nesta fase procura-se, assegurar que o carregamento dos dados e a performance das interrogações são eficientes e que a arquitectura é escalonável.

- Os carregamentos de teste são efectuados com volumes de dados pequenos para garantir o bom funcionamento.
- Comparam-se estes valores de performance de carregamento do ETL para antecipar questões de escalabilidade. Assim pontos de fraqueza que sejam detectados podem ser melhorados.
- Efectuam-se operações simples com junções para validar a performance das interrogações em volumes de dados muito grandes.

3.4 Integridade de dados

Neste teste procura-se verificar que o processo de ETL funciona correctamente em relação a outros processos de *upstream* e *downstream*.

3 Conclusão

O processo ETL é o mais complexo e moroso na construção de um sistema DW, devido a aspectos já anteriormente vistos neste artigo. Nos dias de hoje são disponibilizadas diversas ferramentas de ETL no mercado, cada uma com as suas particularidades. Entre estas ferramentas destacam-se a *Oracle Warehouse Builder* (OWB), *SQL Server Integration Services*, entre outras referidas no presente artigo. As suas capacidades de tratamento e manipulação de informação, aliadas a facilidade e simplicidade de utilização, tornam-nas uma referência entre as ferramentas ETL abordadas. Na aquisição de uma ferramenta deste tipo é muito importante saber adequar essa escolha ao problema em questão, sendo que a produtividade na obtenção das informações geradas pelo DW irá reflectir o grau de acerto dessa escolha.

Referências

1.http://imasters.uol.com.br/artigo/11721/bi/arquitetura_de_data_warehouse_parte_02/imprimir acedido em 8 Junho 2010
2. Rudman, W.; Brown, C.; Hewitt, C. The use of data mining tools in identifying medication error near misses and adverse drug events. Top Health Information Management; 23(2). p. 94–103; 2002.
3. Evaluating ETL and Data Integration Platforms http://www.evolve.mb.ca/dw/etlreport.pdf acedido 8 Junho 2010
4. Cza. Shilakes, J. Tylman. Enterprise Information Portals. Enterprise Software Team, em http://www.sagemaker.com/company/downloads/eip/indepth.pdf acedido em 8 Junho 2010
5. M. Demarest, The politics of data warehousing. http://www.uncg.edu/ism/ism611/politics.pdf acedido em 8 Junho 2010
6. B. Inmon. The Data Warehouse Budget. DM Review Magazine, January 1997, em www.dmreview.com/master.cfm?NavID=55&EdID=1315
7. R. Kimbal, L. Reeves, M. Ross, W. Thornthwaite. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying DataWarehouses. John Wiley & Sons, February 1998.
8. P. Vassiliadis. Gulliver in the land of data warehousing: practical experiences and observations of a researcher. In Proc. DMDW (Stockholm, Sweden, 2000), pp. 12.1 -12.16.
9. Rob Karel and Michael Goulde Market Overview: Open Source ETL Tools http://www.bismart.be/docs/forrester_research_market_overview_open_source_ETL.pdf acedido em 8 Junho 2010
10. Jorg, T., Dessloch, S.: Towards generating ETL processes for incremental loading. IDEAS, 101-110, 2008
11. Jorg, T., Dessloch, S.: Formalizing ETL Jobs for Incremental Loading of DataWare-houses. BTW, 327-346, 2009
12. Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004

13. Panos Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., Skiadopoulos, S.: A generic and customizable frameworkfor the design of ETL scenarios. Information Systens 30, 492-525, 2005