Data Warehouse Back-End Tools

Alkis Simitsis
National Technical University of Athens, Greece

Dimitri Theodoratos
New Jersey Institute of Technology, USA

INTRODUCTION

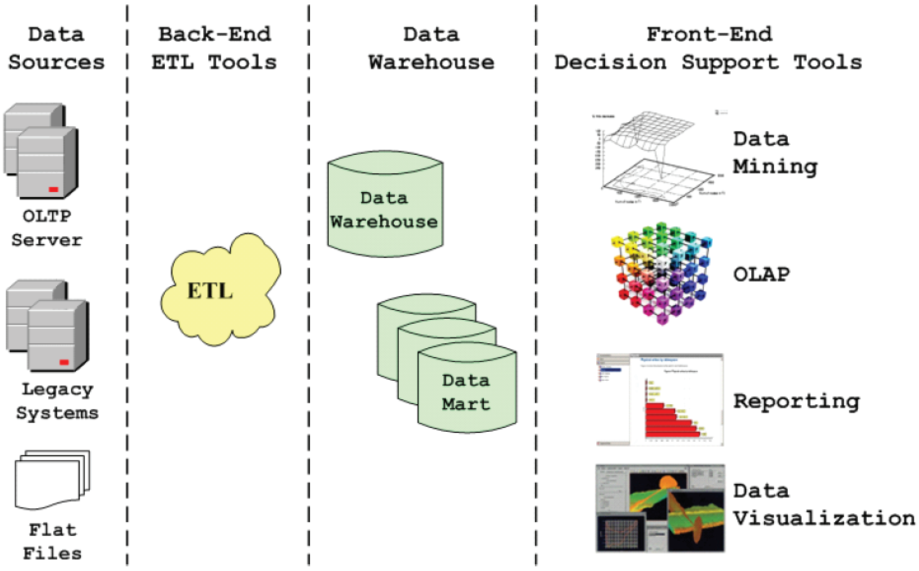
The back-end tools of a data warehouse are pieces of software responsible for the extraction of data from several sources, their cleansing, customization, and insertion into a data warehouse. In general, these tools are known as Extract – Transformation – Load (ETL) tools and the process that describes the population of a data warehouse from its sources is called ETL process. In all the phases of an ETL process (extraction and transportation, transformation and cleaning, and loading), individual issues arise, and, along with the problems and constraints that concern the overall ETL process, make its lifecycle a very complex task.

BACKGROUND

A Data Warehouse (DW) is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst, etc.) to make better and faster decisions. The architecture of the data warehouse environment exhibits various layers of data in which data from one layer are derived from data of the previous layer (Figure 1).

The front-end layer concerns end-users who access the data warehouse with decision support tools in order to get insight into their data by using either advanced data mining and/or OLAP (On-Line Analytical Processing) techniques or advanced reports and visualizations. The central data warehouse layer comprises the data warehouse fact and dimension tables along with the

Figure 1. Abstract architecture of a Data Warehouse



appropriate application-specific data marts. The back stage layer includes all the operations needed for the collection, integration, cleaning and transformation of data coming from the sources. Finally, the sources layer consists of all the sources of the data warehouse; these sources can be in any possible format, such as OLTP (On-Line Transaction Processing) servers, legacy systems, flat files, xml files, web pages, and so on.

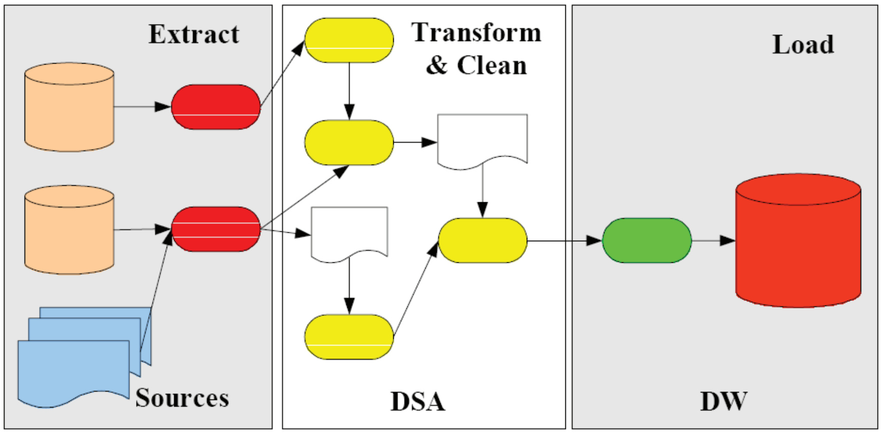
This article deals with the processes, namely ETL processes, which take place in the back stage of the data warehouse environment. The ETL processes are data intensive, complex, and costly (Vassiliadis, 2000). Several reports mention that most of these processes are constructed through an in-house development procedure that can consume up to 70% of the resources for a data warehouse project (Gartner, 2003). The functionality of these processes includes: (a) the identification of relevant information at the source side; (b) the extraction of this information; (c) the transportation of this information from the sources to an intermediate place called Data Staging Area (DSA); (d) the customization and integration of the information coming from multiple sources into a common format; (e) the cleaning of the resulting data set, on the basis of database and business rules; and (f) the propagation of the homogenized and cleansed data to the data warehouse and/or data marts. In the sequel, we will adopt the general acronym ETL for all kinds of in-house or commercial tools, and all the aforementioned categories of tasks.

Figure 2 abstractly describes the general framework for ETL processes. In the left side, the original data providers (Sources) exist. The data from these sources are extracted by extraction routines, which provide either complete snapshots or differentials of the data sources. Next, these data are propagated to the Data Staging Area (DSA) where they are transformed and cleaned before being loaded to the data warehouse. Intermediate results in the form of (mostly) files or relational tables are part of the data staging area. The data warehouse (DW) is depicted in the right part of Fig. 2 and comprises the target data stores, i.e., fact tables for the storage of information and dimension tables with the description and the multidimensional, roll-up hierarchies of the stored facts. The loading of the central warehouse is performed from the loading activities depicted right before the data warehouse data store.

State of the Art

In the past, there have been research efforts towards the design and optimization of ETL tasks. Among them, the following systems are dealing with ETL issues: (a) the AJAX system (Galhardas et al., 2000), (b) the Potter's Wheel system (Raman & Hellerstein, 2001), and (c) Arktos II (Arktos II, 2004). The first two prototypes are based on algebras, which are mostly tailored for the case of homogenizing web data; the latter concerns the modeling and the optimization of ETL processes in a customizable and extensible manner. Additionally, several research efforts have dealt with individual issues and problems of the ETL processes: (a) design and

Figure 2. The environment of Extract-Transform-Load processes


Extract
Transform
& Clean
Load
Sources
DSA
DW