Chapter 2.22
Extraction, Transformation, and Loading Processes

Jovanka Adzic
Telecom Italia, Italy

Valter Fiore
Telecom Italia, Italy

Luisella Sisto
Telecom Italia, Italy

ABSTRACT

ETL stands for extraction, transformation, and loading, in other words, for the data warehouse (DW) backstage. The main focus of our exposition here is the practical application of the ETL process in real world cases with extra problems and strong requirements, particularly performance issues related to population of large data warehouses. In a context of ETL/DW with strong requirements, we can individuate the most common constraints and criticalities that one can meet in developing an ETL system. We will describe some techniques related to the physical database design, pipelining, and parallelism which are crucial for the whole ETL process. We will propose our practical approach, "infrastructure based ETL"; it is not a tool but a set of functionalities or services that experience has proved to be useful and widespread enough in the ETL scenario, and one can build the application on top of it.

INTRODUCTION

ETL stands for extraction, transformation, and loading, in other words, for the data warehouse backstage. A variety of commercial ETL tools exist in the market (IBM, 2005; Informatica, 2005; Microsoft, 2005; Oracle, 2005), with a recent market review of Gartner Research (Gartner, 2005). A lot of research efforts exist (Golfarelli & Rizzi, 1998; Husemann, Lechtenborger, & Vossen, 2000; Tryfona, Busborg, & Christiansen, 1999; Vassiliadis, Simitsis, & Skiadopoulos, 2002, May; Vassiliadis, Simitsis, & Skiadopoulos, 2002, November) mostly targeting modeling (conceptual, logical) and methodology issues (like logical modeling of ETL workflows). Some works are focused on the end-to-end methodology for the warehouse and ETL projects (Kimball & Caserta, 2004; Kimball, Reeves, Ross, & Thornthwaite, 1998; Vassiliadis, Simitsis, Georgantas, & Terrovitis, 2003) targeting the complete life cycle of the DW project, describing how to plan, design, build, and run the DW and its ETL backstage. The main focus of our exposition here is the practical application of the ETL process in real world cases with extra problems and strong requirements, particularly performance issues related to population of large data warehouses (one case study is described in Adzic & Fiore, 2003).

In this chapter, we will first discuss the ETL scenario, requirements, criticalities, and so forth that constitute the general framework for ETL processes. Then we will describe some techniques related to the physical database design, pipelining, and parallelism which are relevant for performance issues. Finally, we will describe our practical approach, "infrastructure based ETL"; it is not a tool but a set of functionalities or services that experience has proved to be useful and widespread enough in the ETL scenario, and one can build the application on top of it.

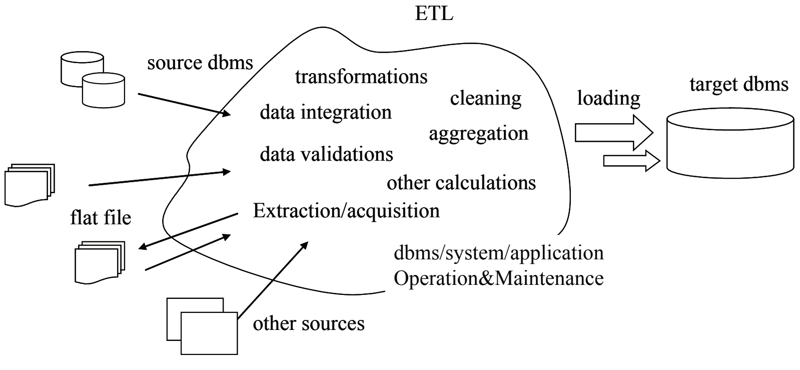
ETL SCENARIO

The primary scenario in which ETL takes place is a wide area between the sources of data and a target database management system (DBMS); in the middle, there are all the required functionalities to bring and maintain historical data in a form well suited for analysis (Figure 1).

All the work to collect, transform, and load data from different and multiple sources to a target DBMS structured for analysis is what we call ETL.

A DW project consists of three main technical tasks: ETL, database design, and analysis techniques and tools; each of them has particular issues and requirements. Above all, we must consider the problems in accessing data owned by other departments, groups, and so on; obtaining the necessary grants to access data is not always easy for both technical and nontechnical reasons. These political problems can impose constraints and work-around that make the ETL process more complex. Another topic is the absence of internal enterprise standardization. It can be very difficult to find the same rules (even in the same department) in naming files, in expressing a date, in choosing a structure for files, and so forth. These problems, involve very general questions like

Figure 1. ETL scenario


ETL
source dbms
transformations
cleaning
data integration
aggregation
data validations
other calculations
flat file
Extraction/acquisition
dbms/system/application
Operation&Maintenance
other sources
loading
target dbms