Chapter 7.18
Data Warehouse Refreshment

Alkis Simitisis
National Technical University of Athens, Greece

Panos Vassiliadis
University of Ioannina, Greece

Spiros Skiadopoulos
University of Peloponnese, Greece

Timos Sellis
National Technical University of Athens, Greece

ABSTRACT

In the early stages of a data warehouse project, the designers/administrators have to come up with a decision concerning the design and deployment of the backstage architecture. The possible options are (a) the usage of a commercial ETL tool or (b) the development of an in-house ETL prototype. Both cases have advantages and disadvantages. However, in both cases the design and modeling of the ETL workflows have the same characteristics. The scope of this chapter is to indicate the main challenges, issues, and problems concerning the manufacturing of ETL workflows, in order to assist the designers/administrators to decide which solution suits their data warehouse project better and to help them construct an efficient, robust, and evolvable ETL workflow that implements the refreshment of their warehouse.

INTRODUCTION

In the past, research has treated data warehouses as collections of materialized views. Although this abstraction is elegant and possibly sufficient for the purpose of examining alternative strategies for view maintenance, it is sufficient enough to describe the structure and contents of a data warehouse in real-world settings. Vassiliadis, Quix, Vassiliou, and Jarke (2001) bring up the issue of *data warehouse operational processes* and deduce the definition of a table in the data warehouse as the outcome of the combination of the processes that populate it. This new kind of definition complements existing approaches, since it provides the operational semantics for the content of a data warehouse table, whereas the existing definitions give an abstraction of its intentional semantics. Indeed, in a typical mediation scheme one would pose a query to a "virtual"

data warehouse, dispatch it to the sources, answer parts of it there, and then collect the answers. On the contrary, in the case of data warehouse operational processes, the objective is to carry data from a set of source relations and eventually load them in a target (data warehouse) relation. To achieve this goal, we have to (a) specify data transformations as a workflow and (b) optimize and execute the workflow.

Data warehouse operational processes normally compose a labor intensive workflow and constitute an integral part of the backstage of data warehouse architectures. To deal with this workflow and in order to facilitate and manage the data warehouse operational processes, specialized workflows are used under the general title extraction transformation loading (ETL) workflows. ETL workflows are responsible for the extraction of data from several sources, their cleansing, their customization and transformation, and finally, their loading into a data warehouse.

ETL workflows represent an important part of data warehousing, as they represent the means by which data actually get loaded into the warehouse. To give a general idea of the functionality of these workflows we mention their most prominent tasks, which include:

- The *identification* of relevant information at the source side
- The *extraction* of this information
- The *transportation* of this information to the DSA
- The *transformation* (i.e., customization and integration) of the information coming from multiple sources into a common format
- The *cleaning* of the resulting dataset, on the basis of database and business rules
- The *propagation* and loading of the data to the data warehouse and the refreshment of 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/processes.

In Figure 1, we abstractly describe the general framework for ETL processes. On the left side, we can observe the original data stores (sources) that are involved in the overall process. Typically, data sources are relational databases and files. The data from these sources are extracted by specialized routines or tools, which provide either complete snapshots or differentials of the data sources. Then, these data are propagated to the data staging area (DSA) where they are transformed and cleaned before being loaded into the data warehouse. Intermediate results, again 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 Figure 1 and comprises the target data stores, that is, fact tables for the storage of information and dimension tables with the description and the multidimensional rollup hierarchies of the stored facts. The loading of the central warehouse is performed from the loading activities depicted in the right side before the data warehouse data store.

Despite the plethora of commercial solutions that offer ad-hoc capabilities for the creation of an ETL scenario, a designer/administrator needs a concrete method to develop an efficient, robust, and evolvable ETL workflow. Therefore, this chapter intends to point out the main challenges and issues concerning the generic construction of ETL workflows. As an outline, in the rest of the chapter, we proceed with a brief presentation about the state of the art in ETL technology. Afterwards, we discuss why the modeling of ETL workflows is important and we indicate the main problems that arise during all the phases of an ETL process. Moreover, we present a modeling approach for the construction of ETL workflows, which is based on the life cycle of the data warehouse, along with an exemplary research framework named Arktos II. Finally, we list several open research challenges that proclaim ETL as a commodity of future research.