Chapter 10
Development of ETL Processes Using the Domain–Specific Modeling Approach

Marko Petrović
University of Belgrade, Serbia

Nina Turajlić
University of Belgrade, Serbia

Milica Vučković
University of Belgrade, Serbia

Sladjan Babarogić
University of Belgrade, Serbia

Nenad Aničić
University of Belgrade, Serbia

ABSTRACT

ETL process development is the most complex and expensive phase of data warehouse development so research is focused on its conceptualization and automation. A new solution (model-driven ETL approach – M-ETL-A), based on domain-specific modeling, is proposed for the formal specification of ETL processes and their implementation. Several domain-specific languages (DSLs) are introduced, each defining concepts relevant for a specific aspect of an ETL process (primarily, languages for specifying the data flow and the control flow). A specific platform (ETL-PL) technologically supports the modeling (using the DSLs) and automated transformation of models into the executable code of a specific application framework. ETL-PL development environment comprises tools for ETL process modeling (tools for defining the abstract and concrete DSL syntax and for creating models in accordance with the DSLs). ETL-PL execution environment consists of services responsible for the automatic generation of executable code from models and execution of the generated code.

DOI: 10.4018/978-1-5225-5516-2.ch010
INTRODUCTION

The aim of data warehouse systems, as a specific type of information system, is to provide the necessary support for decision-making in an increasingly competitive and demanding business environment. To this end they enable the collection of vast amounts of business data and provide the necessary tools for extracting business intelligence from the collected data. As stated in (Jarke, Lenzerini, Vassiliou, & Vassiliadis, 2003) such systems are expected to have the right information in the right place at the right time with the right cost in order to support the right decision.

Extract-Transform-Load (ETL) processes are at the core of data warehouse systems as they are responsible for the actual acquisition and integration of business data from a number of different sources, its transformation into strategic business information and the subsequent storage of the transformed data in a format that facilitates business analysis.

It is evident that the development of appropriate ETL processes, in light of the dynamics of modern business systems, faces a number of challenges. Namely, the sheer volume of business data that is to be rapidly gathered, processed, transformed, stored and delivered, imposes strict constraints regarding the performance and scalability of data warehouse systems. Moreover, the accumulated data must somehow be integrated, as it is gathered from diverse, usually very heterogeneous, data sources. Because the data sources may use various data models and be based on different technologies, a wide array of transformations must be performed to first resolve the potential structural and semantic conflicts and then translate such transformed data into a form suitable for its further analysis. On the other hand, the processes must also be agile and flexible so that they are able to easily respond to frequent changes in business requirements and constantly absorb new data sources (and allow for changes in the state and structure of existing data sources). A change in business requirements calls for new business analysis to be conducted, which in turn means that new strategic information must be provided. In other words, according to (El Akkaoui, Mazón, Vaisman, & Zimányi, 2012) agile and flexible ETL tools are needed which can quickly produce and modify executable code based on constantly changing needs of the dynamic business environment. It can, thus, be concluded that the development of ETL processes is extremely complex and time-consuming and that, consequently, it requires significant financial resources. Given that it has been estimated that as much as 70% of the time and effort invested in the development of data warehouses is spent on the development of ETL processes (Kimball & Caserta, 2004; Kimball, Ross, Becker, Mundy, & Thornthwaite, 2010) it is clear that an appropriate methodological approach to ETL process development has become imperative.

To date, research and practice have led to significant progress being made regarding the formalization and automation of data warehouse development. The methodological approaches, developed during the past couple of decades, are aimed at resolving some of the problems inherent to this process such as: high development and maintenance expenses, low productivity, failure to adequately satisfy user requirements, etc. These issues emanate from the nature of the modern business systems that are to be automated by software solutions. Today’s business environment is characterized by rapid organizational and technological changes, and modern business systems are becoming increasingly complex. Furthermore, in light of the pervasiveness of the Internet and the transition to e-business there is a growing need for business integration. It is, thus, evident that the main issues that must be dealt with when developing modern software systems are: overcoming complexity, attaining adaptability to frequent changes and enforcing the interoperability of heterogeneous distributed software systems.