Modelling ETL Processes of Data Warehouses with UML Activity Diagrams[*]

Lilia Muñoz[1], Jose-Norberto Mazón[2], Jesús Pardillo[2], and Juan Trujillo[2]

[1] Lucentia Research Group. Dep. of Information Systems, Control, Evaluation and
Computing Resources, University Technological of Panama, Panama
lilia.munoz@utp.ac.pa
[2] Lucentia Research Group, Dep. of Software and Computing Systems,
University of Alicante, Spain
{jnmazon,jesuspv,jtrujillo}@dlsi.ua.es

Abstract. *Extraction-transformation-loading* (ETL) processes play an important role in a *data warehouse* (DW) architecture because they are responsible of integrating data from heterogeneous data sources into the DW repository. Importantly, most of the budget of a DW project is spent on designing these processes since they are not taken into account in the early phases of the project but once the repository is deployed. In order to overcome this situation, we propose using the *unified modelling language* (UML) to conceptually model the sequence of activities involved in ETL processes from the beginning of the project by using *activity diagrams* (ADs). Our approach provides designers with easy-to-use modelling elements to capture the dynamic aspects of ETL processes.

Keywords: ETL, UML, activity diagrams, modelling, processes.

1 Introduction

In the nineties, Inmon [1] coined the term *data warehouse* (DW) as a "collection of integrated, subject-oriented databases designated to support the decision support function". Specifically, a DW is *integrated*, because data are collected from heterogeneous sources (legacy systems, relational databases, COBOL files, etc.) to adapt them for decision making. Importantly, the integration of these sources is achieved in the DW domain by defining a set of *extraction-transformation-loading* (ETL) processes. These processes are responsible for extracting data from heterogeneous sources, transforming their data into an adequate format (by conversion, cleaning, etc.) and loading the processed data into the DW.

Designing ETL processes is extremely complex, costly and time consuming [21]. However, it has been broadly argued in the literature that ETL processes are one of the most important parts of the development of a DW [1,10].

[*] Supported by Spanish projects ESPIA (TIN2007-67078) and QUASIMODO (PAC08-0157-0668). Lilia Muñoz is funded by SENACYT and IFARHU of the Republic of Panama. Jose-Norberto Mazón and Jesús Pardillo are funded under Spanish FPU grants AP2005-1360 and AP2006-00332, respectively.

R. Meersman, Z. Tari, and P. Herrero (Eds.): OTM 2008 Workshops, LNCS 5333, pp. 44–53, 2008.

Shilakes [7] reports that ETL and data cleaning tools are estimated to cost at least one third of effort and expenses in the budget of a DW, while [8] mentions that this number can rise up to 80% of the development time in a DW project. Currently, specialised tools provided by DBMS vendors such as [3,5,4], are widely used for specifying ETL processes in a DW environment. Unfortunately, these tools have some drawbacks since they lack in a conceptual modelling perspective: (i) lack of specificity and expressiveness, (ii) dependency on the target platform, (iii) highly complex configuration and setup. Furthermore, the high price of acquisition and maintenance of these tools makes that many organisations prefer to develop their own ETL processes by means of specific programs. Therefore, this scenario can make the design of ETL processes difficult for the integration of heterogeneous data sources in the DW.

To overcome these drawbacks, in recent years, several proposals have been defined for the conceptual modelling of ETL processes [10,11,13,12,14]. They advocate them from the perspective of the sources, and their transformation processes of those sources. However, they only propose static structures to model ETL processes, which do not allow to evaluate the behaviour of the designed ETL processes. Furthermore, they do not define formal mechanisms for representing special conditions, *e.g.*, sequence of the control flows or temporal restrictions. Finally, some of these proposals are not formally integrated in a concrete framework, thus providing only partial solutions and making difficult their application when a disparate set of sources is being integrated. Keeping in mind these considerations and the need of new modelling elements to represent special conditions of ETL processes, this paper proposes the developing of a *conceptual modelling framework*, that allows us to clarify the behaviour of an ETL process. To this aim, we take advantage of the high expressivity of the *unified modelling language* (UML) [9], specifically, the *activity diagrams* (ADs). The UML ADs are behavioural diagrams used to capture the dynamic aspects of a system. In this sense, we designed a set of *modelling elements* of AD to represent the activities involved in ETL processes, those will allow us to model the behaviour of a process and the seamless integration of the design of ETL processes together with the DW conceptual schema. This paper is organised in the following way. In Section 2, we present the related work. In Section 3, our proposal based on UML ADs for the conceptual modelling of ETL processes of DW is presented. In Section 4, we state a concrete example of application of our proposal. Finally, in Section 5, we encompass the main conclusions and future work.

2 Related Work

Conceptual Models of ETL Processes. Conceptual modelling of ETL processes have been developed from several perspectives: generic modelling [10], development methodologies [11] and ontology-based [14]. Although these approaches are interesting (becoming a milestone in ETL design), they lack in providing mechanisms for conceptually representing some important issues, such as temporal conditions or behaviour, and dynamic aspects of ETL processes. Furthermore,