BPMN-Based Conceptual Modeling of ETL Processes

Zineb El Akkaoui[1], José-Norberto Mazón[2],
Alejandro Vaisman[1], and Esteban Zimányi[1]

[1] Department of Computer and Decision Engineering (CoDE)
Université Libre de Bruxelles,
Brussels, Belgium
{zelakkao,avaisman,ezimanyi}@ulb.ac.be
[2] Department of Software and Computing Systems (WaKe)
Universidad de Alicante,
Alicante, Spain
jnmazon@dlsi.ua.es

Abstract. Business Intelligence (BI) solutions require the design and implementation of complex processes (denoted ETL) that extract, transform, and load data from the sources to a common repository. New applications, like for example, real-time data warehousing, require agile and flexible tools that allow BI users to take timely decisions based on extremely up-to-date data. This calls for new ETL tools able to adapt to constant changes and quickly produce and modify executable code. A way to achieve this is to make ETL processes become aware of the business processes in the organization, in order to easily identify which data are required, and when and how to load them in the data warehouse. Therefore, we propose to model ETL processes using the standard representation mechanism denoted BPMN (Business Process Modeling and Notation). In this paper we present a BPMN-based metamodel for conceptual modeling of ETL processes. This metamodel is based on a classification of ETL objects resulting from a study of the most used commercial and open source ETL tools.

1 Introduction

The term Business intelligence (BI) refers to a collection of techniques used for identifying, extracting, and analyzing business data, to support decision-making. BI applications include a broad spectrum of analysis capabilities, including On-Line Analytical Processing (OLAP) and data mining tools. In most cases, organizational data used by BI applications come from heterogeneous and distributed operational sources that are integrated into a data warehouse (DW). To achieve this integration, the data warehousing process includes the extraction of the data from the sources, the transformation of these data (e.g., to correct semantic and syntactic inconsistencies) and the loading of the warehouse with the cleansed, transformed data. This process is known as ETL (standing for Extraction, Transformation, Load). It has been widely argued that the ETL process development

A. Cuzzocrea and U. Dayal (Eds.): DaWaK 2012, LNCS 7448, pp. 1–14, 2012.