Chapter 15

Ontology Development for ETL Process Design

Azman Ta’a
Universiti Utara Malaysia, Malaysia

Mohd Syazwan Abdullah
Universiti Utara Malaysia, Malaysia

ABSTRACT

The Extract, Transform, Load (ETL) process design is difficult to perform because of the ambiguity of user requirements and the complexity of data integration and transformation. Current studies have explored the ontology-based approach to overcome these limitations by reconciling the semantics of user requirements within the ETL process design for easy generation of the ETL process specification. The ontology for ETL process activities has been developed by using the Requirement Analysis Method for ETL Processes (RAMEPs) that is gathered from the perspectives of organization, decision-maker, and developer. Therefore, the ontology is used to generate the ETL process specification for a student affairs’ Data Warehouse (DW) system. The correctness of the ontology model was validated by using an appropriate reasoner. Moreover, the process of ontology development for the case study is presented and shows how the ontology-based approach was successful in implementing the design and generating the ETL process specification.

INTRODUCTION

ETL is a series’ extraction, transformation, and loading process of data sources to the required DW. Modeling and designing the ETL processes involves several tasks, such as defining DW schemas, and data integration and transformation activities. The problems related to the ETL process design are difficult to resolve due to the ambiguity of user requirements and the complexity of data integration and transformation activities. The high-level, multi-view user requirements (e.g., staff, manager, CEO) need to be analyzed properly for supporting the design of the ETL
Ontology Development for ETL Process Design

processes, that is, the correctness of semantics for user requirements in relation to the data sources. Moreover, this task is crucial for managing the complexity of ETL process activities such as extracting, filtering, merging, and conversion. This complexity is obvious within the heterogeneity of the DW environment.

Currently, an effort to tackle these problems has been explored by using an ontology-based approach. The existing techniques, such as Entity Relationship Modeling (ERM), Object Role Modeling (ORM), or Unified Modeling Language (UML), all have limitations on presenting the semantics of user requirements toward the heterogeneous data sources. Moreover, an understanding of the user requirement in the context of data sources is important to ensure that the DW system can be properly organized, determined, and supported by the ETL process activities (Kimball & Caserta, 2004). Therefore, a systematic method for developing the ontology for ETL processes is important to ensure that the user requirements and related data sources are clearly defined and mapped with the DW domain. The rest of the chapter is organized as follows.

The next section presents the background of ETL processes and the need for ontology in designing the ETL process. This is followed by the discussion on the Ontology-Oriented Approach and the ontology development process within the RAMEPs method is presented. Then, the case study using this approach using the Student Affairs data warehouse is discussed and followed by the ontology validation and evaluation processes. Finally, the discussion on using ontology in ETL design is and the concluding remarks are presented.

BACKGROUND

The representation of the ETL processes through an ontological approach is suitable to generate the ETL process specification programmatically. Apparently, the ontology formally provides meaning for data integration and transformation activities. However, current works have not properly explained the ontology development process for ETL process activities. Particularly, the ontology is concerned about the definition of ETL process operations and inter-mapping between user requirements toward the relevant data sources. Our approach will utilize the application ontology for DW domain, which has not been given much attention by researchers as well as practitioners in building the ontology. Moreover, the heterogeneous scenario of DW system will be structured according to the ontology in order to highlight the data source conflicts that need to be resolved.

In the ontology development process, the construction of ontology is semantically described by the user requirement glossaries. The semantics of the user requirements is described at a high-level meaning, so that the user requirements can be possibly mapped to the data sources for accomplishing the transformation and integration processes. Strong linkages between requirement glossaries and appropriate data sources through ontology structure will produce the ETL process specification programmatically. This can be done through invoking an appropriate algorithm and reasoning to the ontology. Furthermore, the use of ontology is based on Description Logic (DL), which constitutes the most commonly used knowledge representation formalism (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007).

This research used the Web Ontology Language (OWL) for knowledge representation that uses the DL formalism. The Resource Description Framework (RDF) is used together with OWL in presenting the ontology, which is capable to model the concepts of the domain, relationships between concepts to attributes, the attributes, and relationship that belong to each attribute. The concepts refer to the facts, whereas the dimensions, measures, business rules, and actions refer to the attributes. The concepts of the domain are represented by classes, while the relationships and attributes are represented by the properties.