A New Approach for Conceptual Extraction-Transformation-Loading Process Modeling

Neepa Biswas, Department of Information Technology, Jadavpur University, Kolkata, India
Samiran Chattapadhyay, Department of Information Technology, Jadavpur University, Kolkata, India
Gautam Mahapatra, DRDO, Ministry of Defence, Govt of India, Research Centre Imarat, Kurmalguda, India
Santanu Chatterjee, DRDO, Ministry of Defence, Govt of India, Research Centre Imarat, Kurmalguda, India
Kartick Chandra Mondal, Department of Information Technology, Jadavpur University, Kolkata, India

ABSTRACT

Erroneous or incomplete data generated from various sources can have direct impact in business analysis. Extracted data from sources need to load into data warehouse after required transformation to reduce error and minimize data loss. This process is also known as Extraction-Transformation-Loading (ETL). High-level view of the system activities can be visualized by conceptual modeling of ETL process. It provides the advantage of pre-identification of system error, cost minimization, scope and risk assessment etc. A new modeling approach is proposed for conceptualization ETL process by using a standard Systems Modeling Language (SysML). For handling increasing complexity of any system model, it is preferable to go through verification and validation process in early stage of system development. In this article, the authors’ previous work is extended by presenting a MBSE based approach to automate the SysML model’s validation by using No Magic simulator. Here, the main objective is to overcome the gap between modeling and simulation and to examine the performance of the proposed SysML model. The usefulness of the authors’ approach is exhibited by using a use case scenario.

KEYWORDS

Conceptual Model, Data Warehouse, ETL, MBSE, Simulation, SysML

INTRODUCTION

Data warehouse (Franconi & Kamblet, 2004) is a repository of historical data which is consolidated in multidimensional format. In warehouse, data is stored in a standard structure which is obtained by integrating data from different operational sources of an organization. Business analyst (Ayhan et al., 2013; Snezana & Violeta, 2010) can access that data, perform analysis, apply business intelligence tool and make a prediction as well as take strategic decision. For maintaining a data warehouse, the main focus is to manage the large amount of data generated from different type of systems (SAP, ERP, Oracle, Mainframe etc.) and store those data in a uniform structure. For managing the uniformity of data, ETL has a very important role. ETL is a widely used process in business organizations. It
identifies and extracts data from various sources, filters and customizes those data according to the required format, at last integrate and update it into data warehouse (Vassiliadis, 2009). Configuring an ETL process is one of the key factors having direct impact over cost, time and effort for establishment of a successful data warehouse. Data modeling (Cagiltay, Topalli, Aykac, & Tokdemir, 2013) gives an abstract view about how the data will be arranged in an organization and how they will be managed. By applying data modeling techniques, the relationship between different data items can be visualized. The modeling concept has a great benefit over organizational data to manage it in a structured way. At starting phase, it is highly recommended to make an efficient modeling and design of the total workflow. Due to the expensive nature of warehouse implementation, good modeling, as well as documentation should be maintained. Based on the report of (Eckerson & White, 2003), designing a well-established ETL workflow consumes almost one-third of cost and effort in a DW implementation. A well-designed ETL process is one of the important aspects to accomplish an effective DW. Each vendor provided tool has their own specific methodology for designing the ETL process (Barateiro & Galhardas, 2005; Kherdekar & Metkewar, 2016). It requires understanding about functionality, language, standards etc. about that particular tool. Moreover, the integrated design is not suitable to execute in other platforms.

During the ETL processing, conceptual modeling reflects the high-level view of entities and relationship among them. It only provides an abstract view of the workflow instead of the implementation details. Different research work has been done for conceptual modeling of ETL. UML, BPMN and Semantic Web are commonly used so far for conceptual modeling techniques. A new way for modeling is proposed an ETL process using a system modeling language (SysML) in our previous work (Biswas et al., 2017).

Although there are many contributions towards ETL abstract modeling is done, SysML is a new direction for conceptualizing and validating of ETL workflow. There is a lot of research scope using SysML to practically implement ETL model, validation, simulation, executable code production in a specific way for the sake of both technical and non-technical users.

Gradually for last few years, the complexity of any system is noticeably growing up. Integration of heterogeneous system component like electrical, software, mechanical etc. are the reason behind it. A formal approach has been discussed by (Week, n.d.) for quantitative estimation of complexity of any system. Similar approach will be used later to analyze complexity of the system model which is taken as an example. Besides this, the system developers always bound to maintain their goal of building the correct product within low cost and fixed delivery date. Besides, clear perception of the overall project scope is required to verify the compliance with requirements. Till now the design of a correct system is a big obstacle for system engineers. Sometimes erroneous system design which remains unrecognized at an early phase of system design can result as uneconomical. Therefore, it is very much practical to validate any complex system design as early as possible. Model-Based System Engineering (MBSE) (Estefan, 2007) has come to tackle all these problems. MBSE restore the previous document-oriented approach with model-based approach. MBSE is a formalized way of modeling in each phase of system development life cycle. MBSE methodology guided modeling phases are requirement analysis, designing, analyzing, verification and verification. Model is capable of expressing all functional and non-functional requirements, structural and behavioral component of a system. Various high-level system model for complex embedded system designed by MBSE supported SysML (SysML, n.d.) language is gaining popularity in these days.

As per INCOSE, system modeling and simulation is a common practice today for system requirement and functionality evaluation. Some portion of system or whole system can be validated as per the requisite. Some analytic routine can be executed for formal analysis purpose or any simulation-oriented analysis can also be used.

This paper aims to propose a new technique for designing conceptual model of ETL by using SysML standard supporting Model-based System Engineering (MBSE) approach and simulate that designed model. SysML is a general-purpose system modeling language which facilitates the system
Towards a Unified Middleware for Ubiquitous and Pervasive Computing
Felix J. Villanueva, Francisco Moya, Fernando Rincon Santofimia, David Villa, Jesus Barba and Juan Carlos Lopez (2009). *International Journal of Ambient Computing and Intelligence* (pp. 53-63).
www.igi-global.com/article/towards-unified-middleware-ubiquitous-pervasive/1372?camid=4v1a