Proceedings of the 3rd International Conference on Computing and Informatics, ICOCI 2011,8-9 June, 2011 Bandung, Indonesia | *Paper No. 061*

ETL PROCESSES SPECIFICATIONS GENERATION THROUGH GOAL-ONTOLOGY APPROACH

Azman Ta'a, Mohd. Syazwan Abdullah and Norita Md. Norwawi

Universiti Utara Malaysia, Malaysia, {azman, syazwan}@uum.edu.my, norita@usim.edu.my

ABSTRACT. The common design-related problems for extract, transform, load (ETL) processes are far away from being resolved due to the variation and ambiguity of user requirements and the complexity of ETL operations. These were the fundamental issues of data conflicts in heterogeneous information sharing environments. Current approaches are based on existing software requirement methods that have limitations on reconciliation of the user semantics toward the modeling of the DW. This will prolong the process to generate the ETL processes specifications accordingly. The solution in this paper is focused on the requirement analysis method for designing the ETL processes. The method - RAMEPs (Requirement Analysis Method for ETL Processes) was developed to support the design of ETL processes by analyzing and producing the DW requirements in perspectives of organization, decision-makers, and developers. The ETL processes are modeled and designed by capturing DW schemas and data sources integration and transformation. The validation of RAMEPs emphasizes on the correctness of the goal-oriented and ontology requirement model, and was validated by using compliant tools that support both these approaches. The correctness of RAMEPs was evaluated in three real case studies of Student Affairs System, Gas Utility System, and Graduate Entrepreneur System. These case studies were used to illustrate how the RAMEPs method was implemented in generating the ETL processes specifications.

Keywords: requirement analysis, ontology requirement model

INTRODUCTION

DW is a system for gathering, storing, processing, and providing huge amounts of data with analytical tools to present complex and meaningful information for decision makers (Ta'a et al., 2010). These data are collected, stored, and accessed in centralized databases in order to sustain competitiveness in businesses (Inmon, 2002). However, the DW system requires the ETL processes to provide the data (Kimball & Caserta, 2004). Specifically, the success of DW system is highly dependent on the ETL processes specifications. There are many issues in requirement, modeling, and designing of the ETL processes due to the non-standardization of methods imposed by the providers through their own DW tools. Moreover, the design tasks need to tackle the complexity of ETL processes from the early phases of DW system development. An early phase is important to ensure the appropriateness of information for the DW systems (Giorgini et al., 2008).

GOAL-ONTOLOGY FOR ETL PROCESSES REQUIREMENTS

Requirement analysis of ETL processes focuses on the transformation of informal statements of user requirements into a formal expression of ETL processes specifications. The

informal statements are derived from the requirement of stakeholders and analyzed from the organization and decision-maker perspectives (Giorgini et al., 2008). We argue that analyzing the DW requirements from the abstract of user requirements toward the detail of ETL processes is important in tackling the complexity of DW system design. It is widely accepted that the early requirement analysis significantly reduces the possibility misunderstanding of user requirements (Yu, 1995). The better understanding amongst stakeholders, the higher are the chances on agreeing on terms and definitions used during the ETL processes execution.

Therefore, our requirement analysis method for ETL processes (RAMEPs) is centered on the organizational and decisional modeling and focuses on the transformation model from the perspective of a developer. By adapting the approach used by Giorgini et al. (2008), the RAMEPs model is presented in Figure 1. The extended works in RAMEPs model are highlighted in the shaded area. The organizational modeling is used to identify the goals that are related to facts, and attributes. The decisional modeling is focused on the information needs by decision makers and related to facts, dimension, and measures. The developer modeling is defined the related actions for the data sources and business rules given. The detail of RAMEPs method is presented in Ta'a et al. (2010).

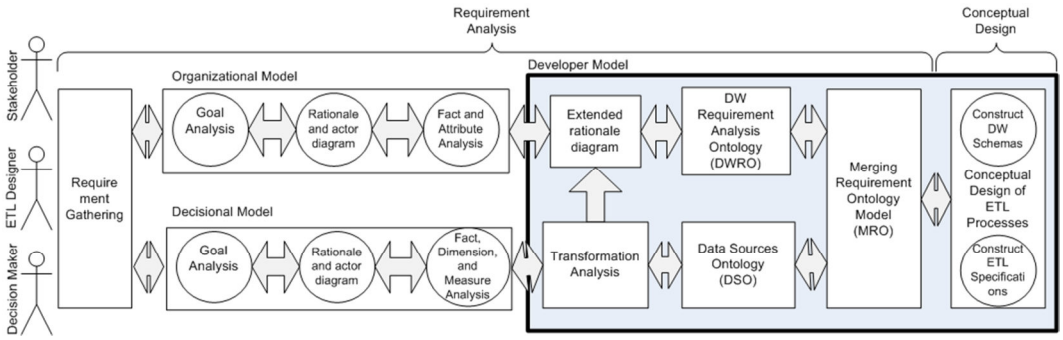

Figure 1. The RAMEPs

THE VALIDATION AND EVALUATION OF RAMEPS

The aim of RAMEPs is to support the design of ETL processes by analyzing and producing DW requirements as required by the decision-maker and organization. Through RAMEPs, the ETL processes are modeled and designed by capturing two important facts: i) DW schemas and ii) data sources integration and transformation activities. Since the RAMEPs is based on the goal-oriented and ontology approach, the validation process is emphasized on the correctness of both approaches. Consequently, the correctness of RAMEPs is not enough until it can be evaluated in the real design of ETL processes. To validate the correctness and ensuring the satisfaction of the RAMEPs, the appropriate goal-oriented and ontology compliant tools are required for capturing and analyzing the DW requirements. The compliant goal-oriented tools must be able to accommodate the elements of organizational, decisional, and developer into the modeling functionalities. Moreover, the compliant ontology tools must be able to capture and present the DW requirements and data sources in an ontology model. The evaluation is conducted for ensuring the usefulness of RAMEPs for designing the ETL processes and was implemented in the real DW project case studies.

Model Checking Process

Generally, model checkers are used to verify the correctness of software systems at design stage. The correctness of a software system is verified according to their system's properties that must be a model-checked. System properties in RAMEPs are DW components (i.e. facts, dimensions, measures, business rules, measures) as defined from the goal-oriented analysis. The method proposed by Ogawa et al. (2008) was adopted to validate the DW components by using compliant tools (DW-Tool and Protégé-OWL). This method was chosen because it uses goal oriented requirement analysis for formal presentation of the software properties.

Moreover, the validation of properties is focused on the sufficiency of design against requirements, which is similar to our objectives. However, our approach is based on Tropos model that emphasized on the goal and resources that describing the DW characteristics. The model checking process and tools are illustrated in Figure 2.

In the checking method, the compliant tools are used to ensure the DW components are properly captured from one model to the next model. For examples, the goals, facts, and attributes in organizational modeling correctly support the goals, facts, dimensions, and measures in the decisional modeling. These DW components in decisional modeling correctly support the developer modeling. The complete DW requirements are modeled as ontology and rechecked for their correctness as ontology structure by using ontology reasoner called Pallet. Since the DW-Tool (Giorgini et al., 2008) not support data transformation analysis as required for ETL processes, a transformation analysis (TA-Tool) was developed and provides the data transformation diagram in developer modeling.

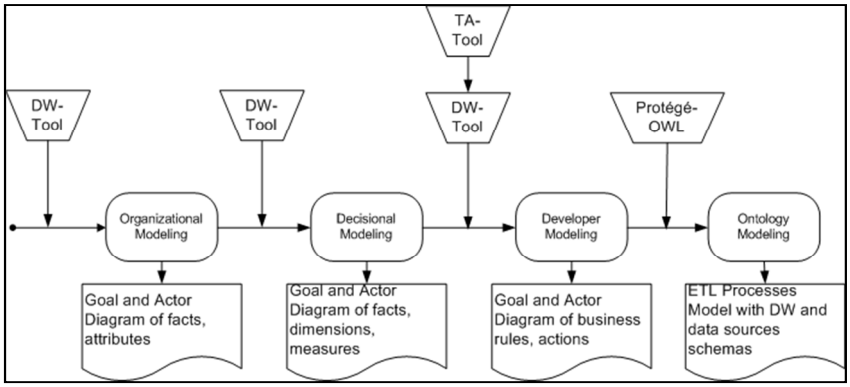

Figure 2. Model Checking Process and Compliant Tools

Model Checking Examples

In organizational modeling phase, the goal cannot be updated in the fact definition. Furthermore, the facts also cannot be updated in the dimension definition. The DW-Tool will ensure the goals only can be inserted or updated within the goal definition area. The gray area of goal description explains the checking mechanism of the model correctness. This principle applies for all phases of modeling to avoid inconsistency among the diagram produced by the DW-Tool. Meanwhile, the TA-Tool is used as an interface for inserting new actions and business rules for data transformation analysis. The TA-Tools is developed by using Visual Basic and make use XML-based data for goal diagram to be manipulated. The goal diagram in DW-Tool is stored in XML and helping developer to added new diagram as proposed for the data transformation analysis.

The Case Study

The case study discussed in this paper focuses on Gas Malaysia (M) utility company. This company promotes, constructs, and operates the Natural Gas Distribution System within Peninsular Malaysia. The company mission in providing the cleanest, safest, cost effective, and reliable energy solutions were motivated them to provide innovative energy solutions to the nation. The requirements' gathering was carried out with the company stakeholders and focuses on the information needed. These requirements are focused on the billing area, which is comprised of billing transaction activities. The billing system is implemented in the Utility Billing Information System (UBIS), which focus is on residential customers and supported by the external application systems namely JDE System and Call Center System (CCS).

i) Organization Modeling

The main goal of the company is Innovative Value for Energy Solutions Provider. This main goal is supported by four sub-goals that need to be fulfilled for achieving the main goal.

To simplify the evaluation process, the case study is focused on the Cost Effective Energy Solution that related to the billing area. The analyses task commence by modeling the DW requirements in the perspective of organization (i.e. the billing department). The stakeholders involved in billing area were identified and represented them by using an actor model. An Actor model explains about dependencies among actors (i.e. billing department, customer, billing operator, call center department). The next step is to analysis fact. Fact analysis aims to identify all the relevant facts for the billing area. The facts are explaining about the information required within goal structure in the billing area. Thus, the analysis is carried out by identifying the facts for each goal from top to down of the goal hierarchy.

ii) Decisional Modeling

There are four phases in decisional modeling: goal analysis, fact analysis, dimension analysis, and measure analysis. All four analyses are connected to each other and aimed to identify and define the DW components. The requirement analysis shifted to decisional modeling, which focuses on the decision-maker perspectives. The tasks are surrounded with the goals for the decision maker to define the requirements. The analysis process starts with identifying actors in goal analysis, and extends it to the fact, dimension, and measure. In this case study, a Billing Manager (BM) was selected as an actor for the decision maker. In previous approaches, the requirement analysis process ends at this stage. The knowledge of facts, dimensions, attributes, and measures will be used in further design of DW and ETL processes. However, the extended analysis on data transformation related to defining facts, dimensions, and measures need to be carried out to ensure the successful implementation of DW system. Therefore, the analysis on data transformation activities is explained next.

iii) Developer Modeling

In business rule analysis, the developer needs to identify the constraint applicable to the ETL processes according to the user requirements. The ETL processes will populate the data sources according to the constraints given. In the case study, the business rules were identified for facts of Sale Volume and Revenue and Billing and Customer Status. According to the analysis, list of business rules is presented in Table 1.

Table 1. List of Business Rule

Facts	Measures	Business Rules
Sale Volume and Revenue	Count Total Customers	Only for spot billing and prepaid billing mode
Billing and Customer Status	Count Total Consumption	Only for spot billing and prepaid billing mode
	Total Customers	• Only for residential customer • Only for spot billing and prepaid billing mode
	Total Billing	Only for spot billing

Based on the business rules given, the transformation analysis can be carried out for conceptually presenting the actions to be taken for providing the DW. The transformation analysis emphasized on the achievement of the ETL processes model for the user requirements and required business rules to absorb the complexity of the data sources. Based on the extended goal diagram of the BM, the plans of actions for Total Customers and Total Consumption for Sale Volume and Revenue goal are presented in Figure 3. The integration of UBIS and JDE data sources will be based on the ontology structure, which clarify the semantic of data sources by define the concepts, classes, properties, and relationship. The ontology mapping between DWRO and DSO is shown in Table 2. The merged requirement ontology (MRO) is reconstructed and rechecked by using Pallet as shown Figure 4.

Table 2. Partial DWRO and DSO mapping for Sale Volume and Revenue

DWRO	DSO	The Mapping
Fact (Sale Volume and Revenue)	UBIS, JDE	Concept: Sale Volume, Sale Revenue
Dimension (account number, customer type, supply type, gas consume, cost billing, billing mode)	Concept: Mode Billing (tbillmode, -) Concept: Customer Type (tbConsType, CommType) Concept: Customer Profile (tbConsumer, Customer) Concept: Supply Type (tbSuppType, SupplyType) …	Billing Mode ↔ Concept: Mode Billing Customer Type ↔ Concept: Customer Type Customer, Account number * ↔ Concept: Customer Profile Supply Type ↔ Concept: Supply Type …

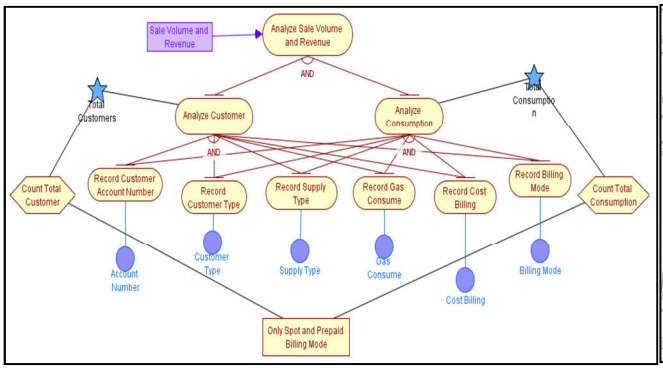

Figure 3. Transformation analysis

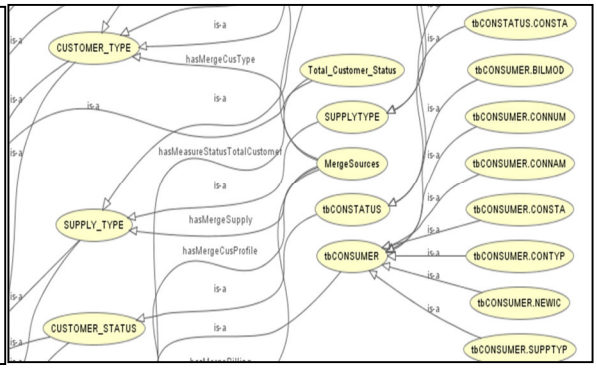

Figure 4. The MRO

iv) Constructing the ETL Processes Specifications

Using the MRO as knowledge representation of DW requirements and ETL operations of billing area, the generation of ETL processes specifications are done automatically. This task can be realized by manipulating the semantic annotation of user requirements and the data sources in MRO. The manipulation process propose set of ETL processes specifications that transform the data sources to the DW schemas as determined in the goal-oriented analysis approach. The generic data transformations used in this case study are EXTRACT, MERGE, FILTER, CONVERT, AGGREGATE, and LOADER. As presented in MRO, the knowledge about information as required and their related data sources have been defined according to RDF/OWL based language. Thus, the MRO is processed according to the algorithm defined by Ta'a et al. (2010) to identify and proposed a set of ETL processes specifications. The reasoning power is based on an inference mechanism in ontology that deals with a wide range of elaborative processing of information. Ontology reasoning is applied on classes and their related properties to derive the ETL processes specifications according to the generic data transformation tasks. To generate the ETL processes specifications automatically, a prototype application for reading, and manipulating the MRO was developed by using Java. The MRO structure that is represented by RDF/OWL language is manipulated through Jena 2 Framework that running on the Eclipse platform.

RESULTS

The results have shown that the ETL processes specifications can be derived from the early stages of user requirements. The list of ETL processes represents data transformation of

utility billing for producing the information Sale Volume and Revenue and Billing and Customer Status. The ETL processes specifications can further be translated into SQL statements or applied to any ETL tools for DW system implementation. However, it is out of this paper scope. The sequence of ETL processes executions were following the results as produced in the generation process. Therefore, the execution order may not necessarily be following the sequences of the ETL processes list. The best practices still depend to the developer efforts, experiences, and knowledge.

EXPERT REVIEWS

The expert reviews were conducted to help clarify the strengths and weaknesses of RAMEPs by using DW scenario of the case study. This method known as an exemplar and is used for evaluating the methodology, especially for requirement engineering approach (Cysneiros et al., 2004). A set of a questionnaire together with the case study was given to six DW developers, which three of them are from government agencies, and others are from DW companies. Their experiences are within the ranges of three to seventeen years in developing and implementing the DW systems in various organizations. The set of questionnaires were designed and accommodate within the scope of RAMEPs. As shown in Figure 5, the experts agreed that the RAMEPs can be implemented by using proper tools, but it will take time to implement in the real environments because of the complexity of DW model and requires more time for learning the method. Nevertheless, the RAMEPs approach enables DW developers to model the DW system from the beginning to the generation of ETL processes.

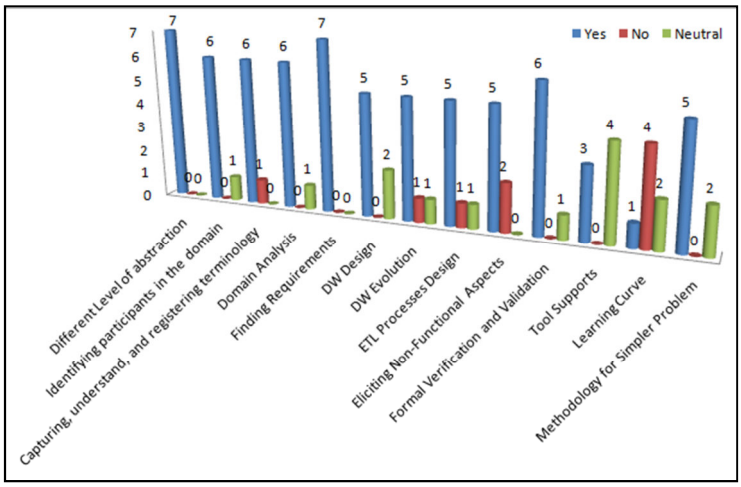

Figure 5. Expert Review Feedbacks

CONCLUSION

The RAMEPs approach has proven that the ETL processes specifications can be derived from the early stages of DW system development. The methodology used in analyzing the user requirements has been validated by DW-Tool and Protégé-OWL successfully. The evaluation approach was carried out by implementing the RAMEPs into various domains of case studies. This will give the multi views of information in DW systems. The DW experts have reviewed the RAMEPs and positively support the method to be implemented in the real environment. It is believed that the adoption of this method can help developers to clearly define the ETL processes prior to the detail design and accelerates the implementation of DW systems. Furthermore, the ontology helps a developer to resolve semantic heterogeneity problems during data integration and generate the ETL processes specifications.

REFERENCES

Inmon, W. H. (2002). *Building the Data Warehouse - Third Edition*: John Wiley & Sons, Inc.

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit. Practical Technique for Extracting, Cleaning, Conforming and Delivering Data*: Wiley Publishing, Inc., Indianapolis.

Giorgini, P., Rizzi, S., & Garzetti, M. (2008). GRAnD: A Goal-Oriented Approach to Requirement Analysis in Data Warehouses. *Decision Support Systems, 45*, 4-21.

Alexiev, V., Breu, M., Bruijn, J. d., Fensel, D., Lara, R., & Lausen, H. (2005). *Information Integration with Ontologies: Experiences from an Industrial Showcase*: John Wiley & Son Ltd.

Yu, E. (1995). *Modeling Strategic Relationships for Process Reengineering*. Unpublished Ph. D, University of Toronto.

Ta'a, A., Abdullah, M. S., & Norwawi, N. M. (2010). RAMEPs: A Goal-Ontology Approach To Analyse The Requirements For Data Warehouse Systems. *WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS*, 7(2), 295-309.

Ogawa, H., Kumeno, F., & Honiden, S. (2008). *Model Checking Process with Goal Oriented Requirements Analysis*. 15th Asia-Pacific Software Engineering Conference.

Cysneiros, L. M., Werneck, V., & Yu, E. (2004). *Evaluating Methodologies: A Requirements Engineering Approach Through the Use of an Exemplar*. 7th Workshop on Autonomous Agents.