Requirements Analysis Method For Extracting-Transformation-Loading (Etl) In Data Warehouse Systems

Azman Taa, Mohd Syazwan Abdullah, Norita Md. Norwawi

College of Arts and Sciences
Universiti Utara Malaysia, 06010 Sintok, Kedah
Tel : 04-9284600, Fax : 04-9284753
E-mail : {azman, syazwan, norita}@uum.edu.my

ABSTRACT

The data warehouse (DW) system design involves several tasks such as defining the DW schemas and the ETL processes specifications, and these have been extensively studied and practiced for many years. The problems in heterogeneous data integration are still far from being resolved due to the complexity of ETL processes and the fundamental problems of data conflicts in information sharing environments. The understanding of an early phase of DW development is essential in properly tackling the complexity of ETL processes. The method to analyses the DW requirements from the abstract level (e.g. goal, sub-goal, stakeholder, dependency) toward the specification of ETL processes (e.g. extracting, filtering, conversion) are important in order to manage the complexity of the ETL processes design (e.g. semantic heterogeneity problems). However, current approaches that are based on existing software requirement approach still have limitations on translating the business semantics for DW requirements toward the ETL processes specifications. Moreover, the understanding of goal in the perspective of the organization and decision makers are important to ensure the semantic of DW requirements can be properly determined, organized, and implemented by the ETL processes. Therefore, the proposed method will utilize the ontology with goal-driven approach in analyzing the requirements of ETL processes.

Keywords
ETL Processes, Data Warehouse, Requirement Analysis, Ontology, Business Intelligence

1.0 INTRODUCTION

Data warehouse (DW) is a system for gathering, storing, and processing huge amount of data with analytical tools to present complex and meaningful information for decision makers. The growing interest in DW system is driven by the need for having a good system for decision-making and such system is crucial for organization to exploit the ever-growing amount of data. These data are collected and stored in centralized databases in order to sustain competitiveness in businesses (Inmon, 2002). However, the DW system is dependent on the process of extract-transform-loading (Kimball & Caserta, 2004). In other words, the success of DW system is dependent on the performance of the ETL processes that are considered as DW operational processes. Currently, the ETL processes software almost supported by the industry in DW development tools. However, there still remain open issues that need to be tackled especially in requirement, modeling, and designing the ETL processes due to the non-standardization of methods and procedures imposed by the DW providers. Moreover, the complexity of ETL processes derived from the DW requirement process need to be managed properly in order to ensure the information needs are satisfied by the DW system (Giorgini et al., 2008).

Therefore, the design of DW system should be based on systematic requirement analysis method for ETL processes due to the limitations and linkages of these methods in modeling and designing the DW system (Simitsis, 2004; Mazon et al., 2007). Clearly, these limitations have contributed to the failure of DW projects. This paper is structured as follows: related work is described in the section 2. Our approach for requirement analysis method for DW system is defined and presented in section 3. Section 4 explains our approach on conceptual design and describes an example by using case study in section 5. Finally, we present our conclusions and plan for the future work in section 6.

2.0 RELATED LITERATURE

The designing of ETL processes is essential for helping the developer to design and maintain the DW system from the early stages of system development. Due to the heterogeneity problems, the tasks to manage and develop the ETL processes become difficult, tedious and complex. The emergence of ontology as the main artifacts of semantic web technology has been used as a solution in semantics heterogeneity problems in information sharing environments (Skoutas & Simitsis, 2007). The ontology has been used to tackle the semantics heterogeneity problems in database integration, especially in DW system environments. Moreover, the database schemas can be modeled as an ontology model with respect of the complexity in ontology construction.

Therefore, an effort to simplify these tasks is important through the ETL tools that support the multimode and multipurpose data integration platform together with the ontology. Furthermore, the creation of ETL specifications and managing their changes in ETL tools are of the highest importance in maintaining the DW systems. However, without a proper method and systematic design of ETL processes from the early stages of DW development, the ETL specifications will be unmanageable and worsen the DW maintenance tasks.

A good quality of software design requires unambiguous, complete, verifiable, consistency and usable user requirements that support data analysis and decision-making processes (Bruckner et al., 2001). However, the work of capturing and analyzing the business or user requirements (refer as DW requirements) is not an easy task because it involves various levels of users, departments and organizations. The tasks involve gathering all the requirements, realities, business rules and constraints affecting the ETL processes into one place. In general, the research efforts on developing software requirements and DW requirements according to the requirements engineering guidelines have been carried out by researchers (Bruckner et al., 2001). In short, their approaches on DW requirements can be classified into process-driven (Kimball & Caserta, 2004), supply-driven/data-driven (Inmon, 2002) and demand-driven/ requirement-driven (Giorgini et al., 2008). Additionally, these approaches have been enhanced by using goal-driven (Giorgini et al., 2008), and model-driven (Mazon et al., 2007).

Indeed, all the related works focused on the requirement of DW schemas. Thus, the specification of ETL processes should be determined from the analyses of DW requirements was not given attention by the researchers. However, few works focused on resolving the heterogeneity problems in DW systems, especially in designing the ETL processes. An outstanding work on this has been carried out by Simitsis (2004) and further enhance on ETL processes design using ontology by Skoutas and Simitsis (2007). However, these approaches have not focused on the analyses of user requirements that important for designing the ETL processes. As a result, the ETL processes were not defined according to the high-level objectives of the stakeholders and decision-makers.

This paper aims to present the systematic methods in designing the ETL processes based on the i* framework and Tropos methodology that emphasis on high-level requirements. Furthermore, both supply and demand driven approaches will be adopted to support the complex requirements of DW systems.

3.0 GOAL AND ONTOLOGY-DRIVEN FOR ETL PROCESSES REQUIREMENTS

Requirement analysis of ETL processes focuses on the transformation of informal statements of user requirements into formal expression of ETL processes specifications. The informal statements of user requirements can be derived from two main approaches namely supply/data driven and demand/ user driven. These approaches are practical in the real implementation of DW, which are the user requirements elicited and analysed from the organization and decision-maker perspective (Giorgini et al., 2008). We argue, an effort to analyse the DW requirement from the abstract definition of user requirements (e.g. goal, sub-goal, stakeholder) toward the detail specifications of ETL processes (e.g. extracting, filtering, conversion) are important in order to handle the complexity (e.g. semantic heterogeneity problems, ETL specifications) of ETL design and ensure the successful of DW system. Thus, a proper and systematic transformation analysis is required on the early phase of requirement analysis to overcome these problems.

In software engineering literature, it is widely accepted that the early requirement analysis will significantly reduces the possibility misunderstanding user requirements. The higher understanding among stakeholders possibly increases the agreeable about terms and definitions used during the ETL processes execution. Therefore, our requirement analysis approach is centered on the organizational and decisional modeling, and focuses on the transformational analysis for defining ETL processes specifications. By adapting the approach used by Giorgini et al. (2008) on requirement engineering of DW systems, the model of our approach is presented in Figure 1.

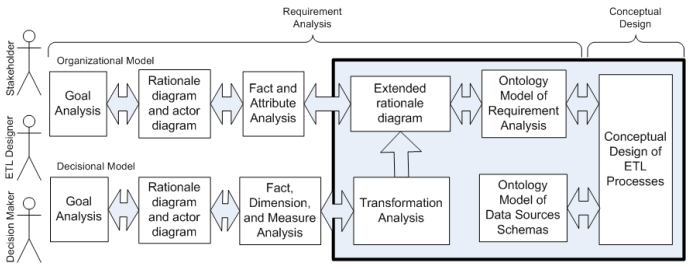

Figure 1: The Requirement Analysis Method of ETL Processes

The model presented in Figure 1 is built from the two different perspectives of requirement analysis: i) organizational modeling that centered on stakeholders, and ii) decisional modeling that related to decision makers. These perspectives include the proposed transformation analysis, which describe the ETL processes. The organizational modeling is used to identify the goal that is related to the DW components such as facts, dimension, measures, and referred

as is analysis and the goal/sub-goal must be satisfied by the organizational and decisional modeling. The decisional modeling is directly focused to the information needs by decision makers and is referred *to be* analysis. Thus, decisions need to be established by the information which is provided by the transformational analysis process that directly produces the information needed by the DW schemas.

The new concept of ontology model for ETL processes is introduced in order to tackle the semantic heterogeneity problems. The analysis tasks of user requirements will be modeled using goal-oriented and ontology approaches. Again, from the goal-oriented approach used by Giorgini et al. (2008), we extend the existing method as shown in Figure 1. It is shows the method in requirement analysis of ETL processes from the elicitation and understanding of the organization goal up to defining the transformation activities for supporting the DW development. At present, the analysis phases are described into three perspectives: i) organization modeling; ii) decisional modeling; and iii) ontology modeling. These three modeling approaches will produce the specifications for designing the DW schemas and ETL processes. Our extended approach in the given model is highlighted in shaded area.

The detail explanation of our approach is based on the Tropos methodology that was developed from the well-accepted i* conceptual framework of software development (Bresciani et al., 2003). By using Tropos methodology, we can analyse the user requirements through actors diagram that present models for actors and show how the actors are depending on each other. The rationale diagram will present the rules of relationship between actors for desired goals to be achieved.

3.1 Organizational, Decision and Ontology Model in Requirement Analysis

Designing the ETL processes to overcome the semantic heterogeneity problems require adequate understanding of the user requirements in order to ensure an appropriate mapping between the data sources to targets is achieved. This can be done through the ontology model, which is developed between the requirement analysis process and conceptual design as shown in Figure 1. The requirement analysis process produces the glossaries of DW components (i.e facts, attributes, dimensions, measures, functions) and will be linked to the appropriate data sources through the ontology mapping mechanism. These tasks comprise three main steps: (i) ontology construction - to develop the rationale ontology for both DW and data sources glossaries; (ii) ontology mapping - to develop the rationale linking between DW and data sources glossaries; and (iii) ETL specifications construction - to develop the rationale specifications of ETL processes.

Finally, the ontology model will be used to design the conceptual model of ETL processes. The detail implementation of these steps will differ from Skoutas & Simitsis (2007) work as our work focus on mapping the application domain to requirement ontology for defining the ETL processes specifications. The organizational, decisional, and ontology model determines the appropriate data (tables and attributes from the sources to the targets) and functions of ETL processes. All glossaries for facts, dimensions, attributes, measures, and functions will be used for building the conceptual design of ETL processes. Since these glossaries represent the data sources from heterogeneous environments, the semantic heterogeneity problems will occur in determining the appropriate data sources toward the agreeable glossaries. More importantly, the agreeable glossaries should present the actual semantics of user requirements.

3.2 Transformation Analysis

Researchers were not focusing on the analysis of transformation activities (i.e. ETL processes) in the early phase of requirement analysis in the previous approaches. Most efforts propose the analysis to be performed in detailed design (Kimball & Caserta, 2004; Giorgini et al., 2008). However, we argue an analysis of transformation activities of DW should be done from the early phases of requirement engineering in order to ensure the correctness of implementation based on the given requirements. Thus, the developer needs to provide the information required by the decision-maker by analyzing the transformation activities as needed. In our approach, the developer needs to associate the measures, facts, dimensions, attributes as previously identified to define the necessary transformation activities that need to be performed.

In Tropos methodology, the transformation analysis can be considered as plan modeling that supporting to the goal modeling. The plan modeling is based on reasoning techniques such as means-end analysis, contribution analysis, and AND/OR decomposition that is also used in analyzing the goal modeling. The notion of Plan modeling is to present an abstract level of doing something that will be applied to model the transformation activities for the specific decision-goal. This model will be the foundation of the conceptual design of ETL processes. As mentioned, the definition of transformation: a finite set of input and output activities are determined by understanding the schemas of data sources. This can be done at the conceptual design stage, where the schemas of data sources will be analysed. However, at requirement analysis stage, we can roughly determine the set of functions, constraints, and rules as required by the transformation activities. Based on the extended rational diagram of organization and decision modeling, we can model the transformation activities through the plan modeling

approach. By using the means-end analysis, we can determine the appropriate plan and appropriate constraints can be determined by using contribution analysis to support the decision-goal.

4.0 CONCEPTUAL DESIGN OF ETL PROCESSES

The aim of requirement analysis is to define the decisional information from the perspective of organizational and decision-makers. Thus, the components of DW schemas need to be defined in the analysis diagrams. The components of DW schemas (i.e. fact, dimension, measure) as represented in specific notations help explain its roles and analysis activities respectively. Generally, these analysis activities implemented sequentially, since the outputs will be inputs for the next analysis. End of these analyses, the glossaries of facts, dimensions, and measures will be used to proceed on the conceptual design of DW systems. However, these tasks are not enough to implement the DW systems, since the details activities of ETL processes are not defined accordingly. The transformational analysis activities need to be carried out in order to determine the ETL processes specification, and starts to execute the back room activities of DW systems.

In conceptual design of ETL processes, all the DW glossaries produced from the organizational and decisional model are restated into the ontology model, which is restructuring the DW requirements toward the ETL processes specifications. According to Simitsis (2004), a conceptual design of ETL processes will guarantee the schemas matching of data sources and targets that founded the user requirements and business goals. Thus, in our approach, we proposed the conceptual design methods as the following steps:

- merging on the application ontology, which is developed from the data sources domain
- mapping the application ontology with requirement ontology, which is required to establish the relationship between DW components (i.e. fact, dimension, measure, action) and data sources
- exploring the hierarchies' structure, where an ETL conceptual schema is generated by navigating the data sources schemas and manipulate the user requirements, rules, and constraints.
- refining the ETL conceptual schema to fully meet the business goal and user requirements expectations.

In the next section, we explain these requirement analysis methods in the example case study. These methods carried out prior to the execution of conceptual design.

5.0 EXAMPLES CASE STUDY

The examples discussed are based on our previous work on a modeling business intelligence model in academic domain. However, requirement analysis work was not properly tackled and the goal-oriented paradigm was disregarded. The requirements elicitation process is based on structured interviews with the identified stakeholder (i.e. Director of Academic Affairs Department (AAD), System Analyst) and study on current system documentations, which focus on goal-oriented business processes. Based on the results of interview, the university goals are identified and explored in details focusing on the goals of AAD in supporting the university main goal. The university goals can be shown in Figure 2.

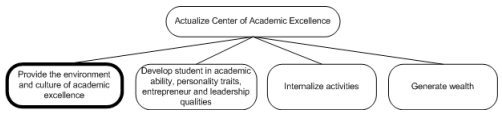

Figure 2: University Goals

To simplify the process, examples will focus on the subject area of student affairs. Based on Figure 2, the sub-goal has provided the environment and culture of academic excellence and it relevancy with the business tasks of AAD. Thus, the next tasks of requirement analysis will be focused on this sub-goal. The scenario of student affairs that require the information from the DW systems, which support it goal can be described as follows: the AAD depends on student for achieving the excellent student and depends on lecturer for the goal of culture of academic excellence. Moreover, the lecturer depends on student for the goal of providing excellent teaching and learning. The analysis task starts with modeling the requirements in the perspective of organization (i.e. the AAD) and followed by the perspective of decision modeling. Due to the space limitation, the rationale diagrams produced during the analysis process is not shown.

The examples focused on the rationale diagram of the university actors that having the goal provide excellent student. The goal decomposed into sub-goals which are: manage promotion, manage admission, manage time-table, manage lecturer, and manage infrastructure by AND decomposition approach. Further analysis decomposes the goal manage student register into another six sub-goals: a record student matric, a record register name, record course taken, record student gender, record student race, and record date register. For the goal manage student class, it decomposed into manage passed results and manage dropped results. By using AND decomposition approach, both goals of manage passed results and manage dropped results were decomposed into a record student matric, record course taken, record grade and record CGPA.

In academic domain, the AAD Director (AADD) is one of the main decision makers that require the information about

student registers and performances in each of academic session. The focus goal associate to AADD (i.e. analyse student register, analyse student performance) is decomposed into sub-goals (i.e. analyse total register, analyse total unregister, analyse student excellence, analyse student passed, analyse student dropped). The transformation analysis will be based on the goals defined by the AADD.

In transformation analysis, the relevant plans are connected to the decision goal. The plans are presented as an abstract level of ETL processes, which is implemented in the implementation phase. By using means-end and contribution analysis, abstract of ETL can be determined properly. There are no activities to determine the appropriate data source schemas toward DW structure at this level. However, as the transformation analysis is carried out, the facts, dimensions, attributes, measures, and abstract processes of ETL (actions) can be used to design the ETL processes as required by goal to be fulfilled. All the tasks in the transformation analysis require a clear understanding of decision makers' need in order to define the suitable transformation activities on ETL processes. The part of plans for student performances goal is presented in Figure 3.

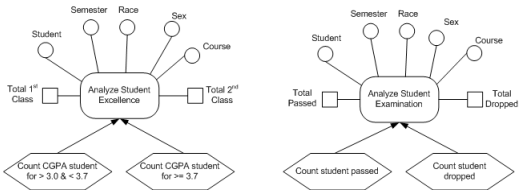

Figure 3: Transformation analysis for Student Performances

In Figure 3, the goal of Analyse Total Student Excellence is based on the facts of student performances. In order to provide information for the goal, appropriate plans are decomposed into two: Count CGPA student between 3.0 and 3.7 and Count CGPA student for 3.7 and above. These plans are proposed to achieve the goal of Analyse Total Student Excellence. The rest of the examples was analysed each of the goals. The rationale diagram for AADD will be completed when each of the decision-goal contains plan that support the information required by the decision makers.

5.1 Ontology Model of DW Requirements

The DW components or glossaries produced need to be modeled accordingly in order to support the conceptual design of ETL processes. All the glossaries are based on extended rational diagrams of organizational and decisional modeling will be presented in ontology structure as shown in Figure 4. Based on the ontology model defined by Skoutas and Simitsis (2007), the constructed ontology can be modeled as O=(C,P,A) whereas, O=Ontology, C=set of classes representing the concepts of the domain, P=set of properties representing attributes of the concepts, A=set of axioms used in defining the relationship between classes. In Figure 4, three classes have been identified as *Register, Examination,* and *Excellence* by using Protégé-2000 tool. Each of the classes contains properties such as student, semester, race, sex, program, total register, under graduate, post graduate, 1st class, and 2nd class. The properties represent the dimension, measure, attribute, and action components. An axiom determines the relationship between classes such as *has_measure, has_attribute, has_process*. These axioms would describe the functions (e.g. conversion, filtering, and aggregation), constraints (e.g. "students must be Malaysian"), and rules (e.g. all students mean under and post graduate) of the ETL processes.

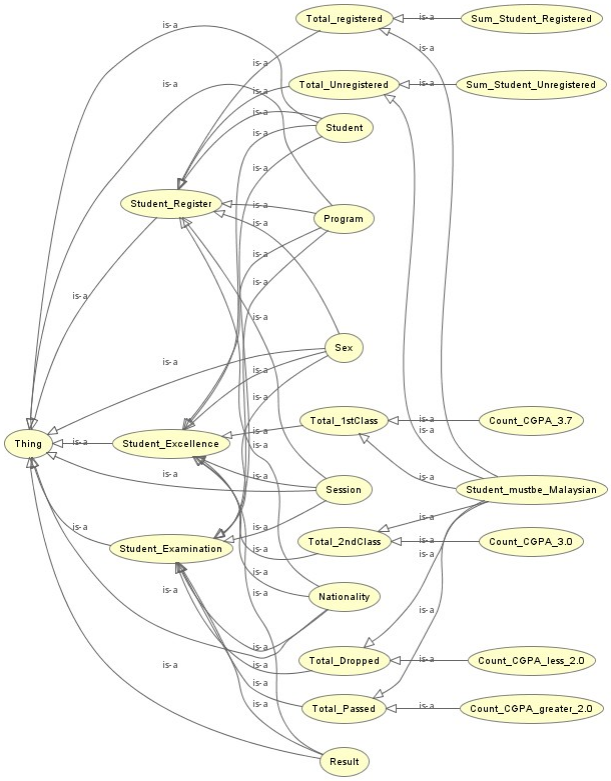

Figure 4: Ontology model for Requirement Analysis

A snippet of OWL code of the above ontology model presented in Figure 5.

```
<owl:Class rdf:ID="Examination"/>
<owl:Restriction>
  <owl:onProperty rdf:resource="#has_measure_student passed"/>
  <owl:onProperty rdf:resource="#has_measure_student dropped"/>
  <owl:onProperty rdf:resource="#has_process_student passed"/>
  <owl:onProperty rdf:resource="#has_process_student dropped"/>
  <owl:allValuesFrom rdf:resource="#Student_must_be_Malaysian"/>
</owl:Restriction>
<owl:Class rdf:ID="Register"/>
<owl:Restriction>
```

```
  <owl:onProperty rdf:resource="#has_measure_student_register"/>
  <owl:onProperty rdf:resource="#has_process_student registered"/>
  <owl:allValuesFrom rdf:resource="#Student_must_be_Malaysian"/>
</owl:Restriction>
```

Figure 5. A snippet of OWL codes of the Ontology model

5.2 Ontology Model of Data Sources

The data sources model is required for further analysis on user requirements. This model will be used to understand the application domain and identify appropriate attributes to be mapped to the data targets. Importantly, prior to design the conceptual model of the ETL processes, ontology model of requirement analysis must be mapped to the heterogeneous data sources model. The ontology model of data sources (as presented in Figure 6) needs to be developed in order to establish the appropriate mapping among concepts or classes between the requirement analysis and the heterogeneous data sources. The ontology models defined from two separate applications that are Academic Student Information System (ASIS) and Graduate Student Information System (GAIS). The concepts Student, Program, Result, and Session were introduced to define the agreeable semantics among the data sources (i.e. ASIS and GAIS). Consequently, the semantics mapping among the appropriate data sources to target DW can be established during the definition of ETL specifications, since the semantic heterogeneity problems have been handled prior to the ETL execution.

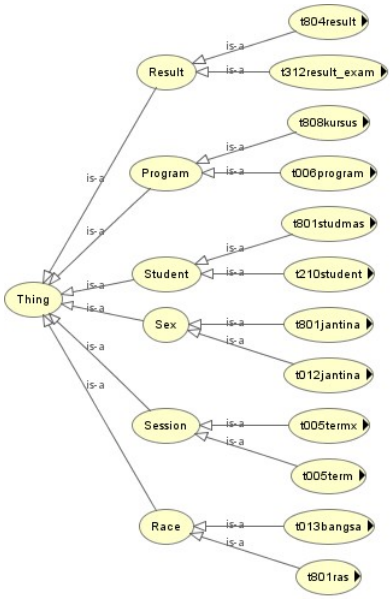

Figure 6. Ontology for Application Domain

The integration of both data sources will be based on ontology structure, which clarifying the semantic heterogeneity during the mapping process. For examples, the concept of *Student* is referred to the schemas *t210student* of ASIS and *t801studmas* of GAIS which represents the meaning of student studying in UUM for both under and post graduate program. In conceptual design of ETL processes, the requirement ontology (as shown in Figure 4) will be mapped with the application domain ontology (as shown in Figure 6) in order to define the ETL processes specifications. By using ontology reasoning, the ETL processes specifications can be produced according to axioms and rules defined on each of relationship between the concepts (Skoutas & Simitsis, 2007). For examples, the ETL processes specifications of facts *Analyse Examination* (Figure 4) can be defined from the mapping by following the specific rules. The possible ETL processes from the above mapping can be presented in Table 1.

Table 1. *ETL processes specifications examples*

Data Source to be selected		ETL Processes to be specified	Data Target to be viewed		
Concept	Field		Concept	Field	Type
Student	f210matric, f804nosem	Merge(f312 term, f804nosem)	Dw_Examination	Student	Dimension
Result	f312pmk, f804cgpa	Count(f312 pmk>=2, f804cgpa>=2)	Dw_Result	Result	Fact

5.0 CONCLUSIONS AND FUTURE WORK

Current work is focused on detailing the whole approach of proposed requirement analysis and design method, especially in transformation analysis and relation to the ontology of requirement glossaries, DW schemas, and data sources schemas. Further works will explore the method on designing the conceptual of ETL processes. The conceptual design will be based on the earlier tasks of requirement analysis. We believe that the adoption of our method can help developer to clearly define the ETL processes prior to the detail design of DW systems. Our methods will be implemented and validated in different domains to prove that the ETL processes can be defined and constructed in the early stages of DW development.

REFERENCES

Bresciani, P., Giorgini, P., Giunchiglia, F., Mylopoulos, J., & Perini, A. (2003). Tropos: An Agent-Oriented Software Development Methodology. Kluwer Academic Publishers, 1-40.

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

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.

Mazon, J.-N., Pardillo, J., & Trujillo, J. (2007). A Model-Driven Goal-Oriented Requirement Engineering Approach for Data Warehouses. Paper presented at the RIGiM, Auckland, New Zealand.

Simitsis, A. (2004). Modeling and Optimization of Extraction-Transformation-Loading (ETL) Processes in Data Warehouse Environments. Unpublished PhD, National Technical University of Athens, Athens.

Skoutas, D., & Simitsis, A. (2007). Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data. Semantic Web & Information Systems, 3(4), 1-24.