Impact Analysis for On-Demand Data Warehousing Evolution

Duong Thi Anh Hoang

Supervised by: Prof. A Min Tjoa

Institute of Software Technology and Interactive Systems,
Vienna University of Technology
Vienna, Austria
{htaduong, amin}@ifs.tuwien.ac.at

Abstract. On-demand data warehousing systems (DWHs) naturally imply a high quality of customized solutions for each individual, where customized business requirements can be facilitated. Therefore, on-demand DWHs have to increasingly deal with the autonomous data changes and schema changes due to continually evolving users' requirements. In fact, DWHs have to be update according to different types of evolution of information sources to reflect the real world subject to analysis. The thesis outlines our investigation focusing on a methodology that can cope with the variety of business requirements in the context of on-demand DWHs. We discuss our plans to address these problems bridging the semantic gap between the requirements, the components and the architecture design of current DWH systems. Overcoming these challenges will facilitate effective tracing of requirements which implies the impacts upon the architectural design and thus increase usability of DWHs for end users.

Keywords: Data warehousing system, Impact analysis, Traceability.

1 Introduction

From our perspective, on-demand DWH systems require constant and rapid evolution. Consequently, changes to DHWs models are inevitable to reflect the current state of DWH requirements. On a database level, data from source systems passes through various databases and transformation processes before being provided to end-users by means of analysis and reporting tools [4]. In this context, small changes in analytical requirements can lead to major unintended impacts for the whole DWHs. Therefore, to effectively determine these unintended impacts caused by requirement changes in the DWH context, impact analysis must be applied, making use of the traceability of data, i.e. where it originates, what its various fields mean and what transformations are needed to perform required analysis.

However, in complex DWHs which consist of many layers, elements, and resources and their complex relationships, it becomes costly and labor intensive to correctly analyze change impacts. The impact of schema changes has been well researched by the database research community, under the motto of schema evolution [3]. In comparison, the amount of research into the impact of DWH schema changes is much less investigated and can be mainly classified into three different approaches: schema evolution, schema versioning and view maintenance [11]. In the common recommended industrial practice for managing DWHs, the impacts of requirements changes upon DWH models must often be estimated manually by application experts. Moreover, most current research in this direction focuses on describing the evolution process of DWH architectures, and lack consistency analysis of dynamic evolution of DWHs in general.

Hence, we are in need of techniques and tools that provide more effective support for impact propagation within a DWHs. Change propagation may not be fully automated, since there are decisions that requires human expertise. However, it is possible to provide support in tracing and managing dependencies, determining what parts of the DWHs architecture are affected by a given change.

In this thesis we consider these challenges in more detail and sketch the ways in which the implications for managing and propagating changes in the context of data warehousing systems can be addressed. Overcoming these challenges will facilitate DWHs to effectively adapt a broader range of change impacts which occur in the DWH layers. To the best of our knowledge, this is a novel research approach that examines the issue of schema changes in a on-demand data warehouse environment.

Section 2 gives an overview over the problems we address. Section 3 introduces our approach to overcome the presented challenges. Section 4 summarizes state-of-the-art related to our research. Finally, Section 5 provides an overview about the next steps for completing the thesis.

2 Problem statement

A data warehouse is a repository integrating information from heterogeneous and autonomous sources for the sake of efficiently implementing decision-support or OLAP queries. This is provided by a multi-dimensional schema that captures the user needs in terms of data content, data constraints, and views of data.

2.1 Challenges of change management in context of DWHs

User requirements can change over a period of time and this causes the need for schemata to be redesigned from scratch. Not changing the schema with respect to changing business contexts can result in wrong predictions and even information losses. Let's suppose that we have *Sale* is the fact table and captured based on the dimensions *Customer, Sales Person, Warehouse, Time* and *Product*. A product marketing strategy is estimated based on the *monthly* sales. After a period of time, analysts want to see the sales information based on *seasons*. This new information has to be propagated to the schema by introducing a new dimension level called *Season*. Not creating this dimension level would lead to wrong product marketing campaign.

Changing requirements in DWHs have to be continuously monitored and captured for better decision support. Unfortunately redesigning a schema is an expensive and difficult process due to various challenges and short-comings, including:

Challenge 1. Auditing and tracing of information and operations are difficult to achieve within many current data warehouse systems. At a time when many organizations are subject to a broad range of regulatory compliance issues, difficulties to conduct audits and demonstrate traceability can be a significant risk.

Challenge 2. Another significant challenge for DWH evolution is the inability to deal with change and the impact change can bring. Enterprise data warehouses should be sustainable over a period of years. Meanwhile, it is quite certain that within a longer period of time, some components of a DWH will change, and that this change will not be completely reflected in the data warehouse.

The problem statement for this research project can be considered as a general problem for DWHs and based on the specified objectives it can be defined as follows:

What is a suitable model for the process of Impact Analysis, based on the current literature, research and practice of others in the field of data warehousing systems? How can an Impact analysis model be developed that supports the process of analyzing the impact of changes in a data warehouse environments?

This general problem can be solved by answering several knowledge problems, which can be formulated in the following research questions:

- How to investigate interoperable dependencies between the diverse and heterogeneous analytical requirements and DWH components, thus efficiently manage impact traceability [14] covering the whole range from analytical requirements to the various DWH layers in DWH architecture design?
- How to develop an effective methodology to provide an impact analysis of the relevance of DWH models with respect to the changing requirements and impacted DWH artifacts?
- How to develop methodologies for incremental impact analysis of the changing analytic requirements in an efficient way? Using the DWH quality metrics, how can a theoretically optimal ranking algorithm can evaluate the performance of the developed impact analysis model?

2.2 Research contribution

To solve the presented challenges, the aim of this research is the development of a requirements-based methodology supporting the impact analysis of DWHs evolution, which also has been outlined in [12]. More specifically, our approach strives to further develop the concept of dependency between requirements engineering and architectural modeling, and provide it with a formally defined semantics.

The dependency concept is hereby clearly positioned with respect to existing concepts for dynamic impact analysis aspects.

- Firstly, it offers an integrated view of the entire DWHs which allow us to, for example, analyze the effect at a business level of a change that takes place at a technical level.

- Secondly, it is supported by an (semi-)automatic mechanism of change propagation based on inconsistency management by extending common change propagation frameworks.

Moreover, the expected contributions of this research will also provide support for impact calculation where the results of dependency tracing can be used to predict the effect of database schema changes, and investigate how impact calculation can be practically and efficiently implemented. Hence, consistency and correctness of changing DWH models are maintained by designing and enforcing constraints over the impact model. An impact analysis framework is designed to support dynamic aspects of DWHs, emphasizing on analytic requirements and the transition between DWH layers. Verifying the feasibility of the transition to an implementation architecture will serve as a validation of the research results.

3 Proposed approach

The core of DWH architecture is laid on the relevant dependencies between components, layers and relations. In terms of impact analysis, these dependencies will help to define the modifications which correspond to the requirement changes, which leads to the specification of a sound impact set. It is crucial that the impact analysis process in DWH system focuses on minimizing changes with an impact on the extraction and interpretation of data, which should be fulfilled in different development phases.

- The use of semantic technologies is considered as the main instrument to identify the traceability among different layers and the mapping between the requirements to the generic architecture. Consequently, impact analysis requires *a knowledge-based impact traceability framework* as common semantics that creates a shared understanding between requirements descriptions and architectural designs. Based on this foundation, the dependency links can be established between the changed elements and the impacted elements.
- The impact analysis process also requires *a formal model that captures the conceptual modeling features of Dependency analysis and Traceability analysis.* Enhanced with constraint propagation mechanisms, the defined formal models provide a sound approach to investigate and calculate the correctness and corresponding consistency of dynamic evolution in the context of DWHs.

The method of a case study is proposed as one way of acquiring empirical data. The case study will provide data of the actual use of impact analysis methods and applications in real settings. We will attempt to apply the defined impact analysis framework in multi-national enterprise data warehouses, which went through mergers and acquisitions and whose operational IT landscapes cannot be harmonized easily [10]. In this case study, the impact analysis process will take advantage of a semantic business information model on top of logical DWH data models, thus enables the harmonization and reusability in complex data warehouse environments.

4 State of the art

Impact analysis relies on techniques and strategies that date back a long time. The maturation of software engineering among software organizations has led to a need to understand how changes affect other software objects than source code. We can find various research approaches and published papers which propose methodologies for software impact analysis. In this view, we distinguish the following categories of methods: implementation-based, model-based methods and requirement-driven impact analysis methods [1]. Researches on requirement-driven impact analysis methods can be found in [7], [9]. However, existing Requirement-Driven Impact Analysis methods have deficiencies, e.g. lacking a mechanism to trace relationship for impact analysis [14]. To date, these issues remain, and there arises a need to provide a mechanism relating requirements and various types of impacted items.

Within the DWH context, to the best of our knowledge, in the state-of-the-art research and practice, most of the impact analysis techniques only focus on changes at the code/implementation level [2], [6], [13]. The relations between requirements, multidimensional design and underlying data sources were focused in [10] which applied a goal-oriented approach to requirement analysis for data warehouses based on the TROPOS methodology; or as in [5] an approach is introduced where conceptual multidimensional models capturing the various user requirements can be obtained. However, these approaches also mainly studied within the context of a static set of requirements. Moreover, determining how to relate requirements to other DWH components is still not further addressed. Furthermore, in current approaches, a comprehensive relationship between requirements and architectural designs is still an open research topic, mainly due to a semantic gap of how the requirements are being realized to the architectures and their compositions in [8].

5 Conclusions and Future research

Reusing DWH models is an important concept DWHs development and management for reducing the modeling time, the designers' workload and the risk to make errors. In this thesis we propose a framework for impact analysis processes which addresses the challenges presented above and enables effective combination of changing analytical requirements and the architecture design into a unified impact analysis process. Our approach can be considered as the foundation to define the semantic needs in an impact analysis framework for the design and development of more effective DWH systems [12].

Our work is still in an early stage and clearly, a lot of work remains to be done for the completion of the thesis. In the near future, we plan to formally define impact analysis principles. Further dependency principles could be considered as well. Moreover, a number of changing operations in DWHs have not yet been explored or are still under definition. Thus, the requirement changes over the architectural design can be implemented with formal semantics, enabling a step further in the development of an impact traceability framework.

Acknowledgment

This work is supported by a Technology Grant for South East Asia of the ***Austrian Council for Research and Technology Development*** and ***ASEA UNINET***.

References

1. Chen, C.-Y. & Chen, P.-C., 2009. A holistic approach to managing software change impact. *Journal of Systems and Software*, 82(12), pp.2051-2067.
2. Cui, Y. & Widom, J., 2003. Lineage tracing for general data warehouse transformations. *The VLDB Journal The International Journal on Very Large Data Bases*, 12(1), pp.41-58.
3. Curino, C., Moon, H. & Zaniolo, C., 2008. Graceful database schema evolution: the prism workbench. *Proceedings of the VLDB Endowment*, 1(1), pp.761–772.
4. Giorgini, P., Rizzi, S. & Garzetti, M., 2005. Goal-oriented requirement analysis for data warehouse design. *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP - DOLAP*, pp.47 - 56.
5. Golfarelli, M., 2009. From User Requirements to Conceptual Design in Data Warehouse Design–a Survey. *Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction*, p.1–16.
6. Hassani, K. et al., 2009. An Approach to Tracking Data Lineage in Mediator Based Information Integration Systems. *2009 International Conference on Information Management and Engineering*, pp.579-583.
7. Imtiaz, Salma, Ikram, N. & Imtiaz, Saima, 2008. Impact Analysis from Multiple Perspectives: Evaluation of Traceability Techniques. *2008 The Third International Conference on Software Engineering Advances*, pp.457-464.
8. Khan, S.S., Greenwood, P. & Garcia, A., 2008. On the Impact of Evolving Requirements-Architecture Dependencies: An Exploratory Study. *Advanced Information Systems Engineering: 20th International Conference, Caise 2008*, pp.243 - 257.
9. Li, Y. et al., 2008. Requirement-centric traceability for change impact analysis: a case study. *Proceedings of the Software process, 2008 international conference on Making globally distributed software development a success story; ICSP'08*, 5007, p.100–111.
10. Mazón, J.-N., Trujillo, J. & Lechtenbörger, J., 2007. Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms. *Data & Knowledge Engineering*, 63(3), pp.725-751.
11. Oueslati, W. & Akaichi, J., 2010. A Survey on Data Warehouse Evolution. *Journal of Database Management*, 2(4), pp.11-24.
12. Priebe, T., Reisser, A. & Hoang, D.T.A., 2011. Reinventing the Wheel?! Why Harmonization and Reuse Fail in Complex Data Warehouse Environments and a Proposed Solution to the Problem. *Proceedings of the 10th International Conference on Wirtschaftsinformatik (WI 2011)*, pp.766-775.
13. Reisser, A. & Priebe, T., 2009. Utilizing Semantic Web Technologies for Efficient Data Lineage and Impact Analyses in Data Warehouse Environments. *Proceedings of 2009 20th International Workshop on Database and Expert Systems Application*, pp.59-63.
14. von Knethen, a, 2002. Change-oriented requirements traceability. Support for evolution of embedded systems. *Proceedings of the International Conference on Software Maintenance (ICSM'02)*, pp.482-485.