A Case Study on Model-Driven Data Warehouse Development

Thomas Benker and Carsten Jürck

Department of Information Systems – Systems Engineering
University of Bamberg
{thomas.benker,carsten.juerck}@uni-bamberg.de

Abstract. The development of data warehouses (DW) is still complex and costly. In recent years, a growing number of contributions suggested the application of Model-Driven Architecture (MDA) to DW development. However, most of these approaches are not evaluated in practice. Therefore, we conducted an evaluation of the Multidimensional Model-Driven Architecture (MD²A), based on a case study derived from a real-world DW project. The case study results can be summarized as follows: The suggested full-automatic transformation of conceptual schemas is not feasible under all circumstances. Necessary performance optimizations in the logical level cannot be derived from the conceptual. Additionally, we can conclude that the integration of evolution concepts into an MDA approach for DWs is desirable for two reasons: First, when conceptual schemas are changed and transformed into the logical level, previously made manual optimizations have to be conserved. Second, transformations from the logical level to code have to consider the protection of historicized data.

Keywords: data warehouse, model driven architecture, case study.

1 Introduction

Data warehouse (DW) systems are the cornerstones of corporate information management. They provide access to integrated, quality-assured, subject-oriented, and historical data for decision support (cf. [1]). However, the construction of these systems is still a complex and costly task. In recent years, more and more contributions advocated the use of the Model-Driven Architecture (MDA) for data warehouse engineering (e.g., [2-6]). Each of these approaches presents a framework for a distinct number of DW components and MDA viewpoints including the corresponding meta-models and transformations. However, most of these approaches, to the best of our knowledge, are not evaluated against requirements of real-world projects. For this reason, suggestions for future research based on practical insights are currently missing.

We present a case study based on a real-world project to evaluate the Multidimensional Model-Driven Architecture (MD²A) of Mazón et al. ([6]). We selected this framework for two reasons: First, it is one of the earliest contributions and a good sample for other MDA approaches in the context of data warehouse engineering (e.g., [2-5]). Second, it is the best described approach measured by the number of
contributions. The publications cover various data warehouse engineering topics like the unified development process and an MDA framework, including transformations and meta-models as well as extensions to special problems like spatial-data warehousing or secure XML data warehouses (cf. [7-11]).

The fundamental research question in this work is: Does the MD²A approach work and which assumptions have to be made? To answer this research question we followed a three step process: First, we created a case study, based on a two-year project in the public sector, with the goal of building a data warehouse for human resources. Second, we implemented the MD²A framework based on the Eclipse Modeling Framework. Finally, we used the prototype to implement the case study and evaluated the results against the requirements of the project.

The remainder is structured as follows: Section 2 gives a brief introduction to the design process of data warehouse systems in general. Section 3 describes the profile of the Human Capital Management (HCM) Project, which serves as an outline for our case study. It also presents the central requirements identified in the course of the project. Section 4 gives a short introduction to the MD²A approach. The application of this approach in the HCM case study is presented in section 5. Thereby, the focus is set to the conceptual design of multidimensional schemas and the derivation of logical data structures. Section 6 encompasses a discussion of the results. The paper ends with a conclusion and future research proposals (section 7).

2 The Data Warehouse Design Process

In [12] the authors distinguish, as a consensus of different approaches, four phases within a data warehouse design process: the requirements analysis, the conceptual design, the logical design, and in some approaches the physical design. We agree to these steps which are also adopted by the reviewed MD²A approach. Currently, there is a lack of agreement on the particulars that are considered within each phase. Therefore, we explicate our perspective on this process. The first phase is the requirements analysis. Its purpose is the collection and filtering of user requirements. Together with an analysis of the operational data sources, the specification of user requirements delivers the input for the conceptual design phase [13]. In this phase a multidimensional schema is created. It serves as foundation for the reconciliation between designer and user. Therefore, the conceptual design should emphasize business concepts and not on technical issues [12]. The logical schema is derived from the conceptual schema and is specific to a certain platform type (e.g., relational database). It is adjusted for the implementation on a certain target system type, considers platform specific constraints, and is optimized to reach non-functional requirements [12]. An example of a non-functional requirement could be the minimization of query response time. The logical design encompasses the translation of the multidimensional model into tables, and the improvement of the schema design such as view materialization [14] or junk dimensions and rapidly changing dimensions as presented in [15]. We consider all adoptions of the relational data structures to be in logical design, because they are largely independent from a specific data base system. All optimizations, like indices or partitions, which are very likely to be database-specific, are considered in the physical design.