Formal Design of Data Warehouse and OLAP Systems

A dissertation presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
in
Information Systems

at Massey University, Palmerston North, New Zealand

Jane Qiong Zhao

2007

Abstract

A data warehouse is a single data store, where data from multiple data sources is integrated for online business analytical processing (OLAP) of an entire organisation. The rationale being single and integrated is to ensure a consistent view of the organisational business performance independent from different angels of business perspectives. Due to its wide coverage of subjects, data warehouse design is a highly complex, lengthy and error-prone process. Furthermore, the business analytical tasks change over time, which results in changes in the requirements for the OLAP systems. Thus, data warehouse and OLAP systems are rather dynamic and the design process is continuous. In this thesis, we propose a method that is integrated, formal and application-tailored to overcome the complexity problem, deal with the system dynamics, improve the quality of the system and the chance of success.

Our method comprises three important parts: the general ASMs method with types, the application tailored design framework for data warehouse and OLAP, and the schema integration method with a set of provably correct refinement rules.

By using the ASM method, we are able to model both data and operations in a uniform conceptual framework, which enables us to design an integrated approach for data warehouse and OLAP design. The freedom given by the ASM method allows us to model the system at an abstract level that is easy to understand for both users and designers. More specifically, the language allows us to use the terms from the user domain not biased by the terms used in computer systems. The pseudo-code like transition rules, which gives the simplest form of operational semantics in ASMs, give the closeness to programming languages for designers to understand. Furthermore, these rules are rooted in mathematics to assist in improving the quality of the system design.

By extending the ASMs with types, the modelling language is tailored for data warehouse with the terms that are well developed for data-intensive applications, which makes it easy to model the schema evolution as refinements in the dynamic data warehouse design.

By providing the application-tailored design framework, we break down the design complexity by business processes (also called subjects in data warehousing) and design concerns. By designing the data warehouse by subjects, our method resembles Kimball's "bottom-up"approach. However, with the schema integration method, our method resolves the stovepipe issue of the approach. By building up a data warehouse iteratively in an integrated framework, our method not only results in an integrated data warehouse, but also resolves the issues of complexity and delayed ROI (Return On Investment) in Inmon's "top-down" approach. By dealing with the user change requests in the same way as new subjects, and modelling data and operations explicitly in a three-tier architecture, namely the data sources, the data warehouse and the OLAP (online Analytical Processing), our method facilitates dynamic design with system integrity.

By introducing a notion of refinement specific to schema evolution, namely schema refinement, for capturing the notion of schema dominance in schema integration, we are able to build a set of correctness-proven refinement rules. By providing the set of refinement rules, we simplify the designers's work in correctness design verification. Nevertheless, we do not aim for a complete set due to the fact that there are many different ways for schema integration, and neither a prescribed way of integration to allow designer favored design.

Furthermore, given its flexibility in the process, our method can be extended for new emerging design issues easily.

Acknowledgement

My sincere and huge thanks go to my supervisor Prof. Klaus-Dieter Schewe for all the guidance and support that he has given me generously during my study.

My thanks also goes to my co-supervisor Associate Prof. Roland Kaschek for his critical discussions on this work.

I am grateful towards my parents for their constant love and care, and the great help in looking after my children.

Finally my thanks to my partner, Henning Köhler, and my two beautiful and lovely daughters, Angela and Nicole, for their love and support which made my life meaningful and enjoyable.

It would be impossible for the completion of this work without any of the help from the people mentioned and the people around me.

Contents

List of Figures