Data Warehouse Schema Design

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von

Dipl.-Inform. Jens Lechtenbörger

Abstract

A data warehouse is an integrated database primarily used in organizational decision making. Although the deployment of data warehouses is current practice in modern information technology landscapes, the methodical schema design for such databases has only been studied cursorily.

In this thesis data warehouse design is performed according to a process model known from traditional database design, organized as a sequence of requirement analysis and specification, conceptual design, logical design, and physical design. The individual design phases rest upon an extension and integration of previously developed, independent warehouse design approaches.

The fact that a data warehouse differs in a number of ways from an operational database poses a need for new criteria to measure the quality of data warehouse schemata. Based on an analysis of previous data warehouse design approaches, multidimensional normal forms and independence properties, most notably update independence, are identified and made formally precise as primary yardsticks for the quality of conceptual and logical data warehouse schemata, respectively.

On the one hand, the first of the proposed multidimensional normal forms combines desirable schema properties such as faithfulness, completeness, and freedom of redundancy, suitably adapted to the warehouse context, whereas two further normal forms deal with the notion of summarizability to guarantee the correctness of analysis results and with the avoidance of null values. On the other hand, update independence of a data warehouse enables efficient maintenance of the data warehouse contents, in particular in environments where decoupled data sources are integrated.

The introduced formal criteria are addressed in a data warehouse design process starting from a requirement analysis and ending with a logical design phase that provides a relational implementation of the data warehouse.

In contrast to traditional database design, the requirement analysis and specification for data warehouses starts from data and analysis requirements as well as schemata of pre-existing operational databases, which have to be integrated in the data warehouse from a multidimensional point of view. The output of the requirement specification is structured in such a way that it supports the following design phases.

In the course of the conceptual design phase, relevant information identified during the requirement specification is arranged in terms of multidimensional fact schemata, which are expressed in a novel multidimensional data model. Additionally, an algorithmic approach utilizing functional dependencies is devised to construct fact schemata starting from the results of the requirement specification. The constructed fact schemata are shown to be in third multidimensional normal form (3MNF).

For the purposes of logical data warehouse design a schema transformation process is proposed to generate a relational implementation of the conceptual schemata by means of a set of update-independent materialized views. By taking advantage of input schemata in 3MNF, null values can be avoided within the resulting relation schemata.

To enforce update independence, a solution based on the notion of view complement is suggested, and expressions to compute complements for monotonic views are given. Importantly, these expressions are applicable to a larger class of views than those offered by earlier approaches and lead at the same time to uniformly smaller complements; furthermore, the complexity of constructing these expressions is shown to be polynomial in the size of schema information, which is in striking contrast to previous approaches, which are NP-complete. Finally, a complement-based approach towards independence of views with respect to arbitrary sets of queries and updates is presented.

Dekan:	Prof. Dr. W. Lange
Erster Gutachter:	Prof. Dr. G. Vossen
Zweiter Gutachter:	Prof. Dr. N. Spyratos

Tage der mündlichen Prüfungen: 28.6.2001 und 29.6.2001
Tag der Promotion: 29.6.2001