8-10-2023

Database Schema as a Graph: A Methodology for Data Warehouse Design

Adir Even
IEM, adireven@bgu.ac.il

Database Schema as a Graph: A Methodology for Data Warehouse Design

TREO Talk Paper

Adir Even
Ben-Gurion University of the Negev
adireven@bgu.ac.il

Abstract

Data Warehouse (DW) is a common term for infrastructural data foundation for Business Intelligence (BI) systems. This study aims at developing a methodology for source-to-target schema conversion, based on directed-graph representation of relational database (DB) schemas. It converts the graph-based representation of a normalized DB to a DW schema that better fits analytical use (a.k.a, "Star Schema"). The methodology permits expert-user intervention for handling schema-design decisions, which often require in-depth understanding of business context business-oriented interpretation. BI systems offer infrastructure, tools and techniques for data visualization and analysis, toward data-driven decision support. BI systems have become an essential asset, as organizations growingly rely on data resources to remain competitive in highly uncertain environments. A DW, the data-infrastructure for BI systems, commonly integrates and restructures data from multiple sources, toward supporting business analysis and managerial decision support. The need to restructure data stems from the different nature of data use. Operational use typically mandates access to specific data records, while maintaining "One version to the truth", by avoiding unnecessary value duplications. Conversely, analytical use more often mandates aggregative view of a large number of data records, toward detecting possible correlations and effects.

The methodology assumes that both the data source and the DW are based on a relational DB schema. However, the different nature of data use mandates different approach toward schema design. Operational use is commonly supported by a normalized DB schema - multiple tables, each with multiple attributes that functionally dependent on the table's primary key (PK's), and some are linked by a foreign key (FK) to other tables. On the other hand, analytical use commonly relies on a "flat" DB structure that stores all attributes in a single relation without necessarily enforcing functional dependencies. Basing analytical data use on a normalized database schema might suffer from slow retrieval performance, as it may rely on multiple computationally expensive JOIN operations. The semi-normalized "Star" schema is based on a single fact table with numeric attributes (a.k.a., fact variables attributes) that reflect measurements of business activities and performance. The fact table is linked by foreign keys to multiple dimension tables, containing characteristics of relevant subjects that can be associated with business activities and may influence performance (a.k.a., dimension variables or attributes). A well-designed "Star" schema is a convenient baseline for generating flat structures along various dimension/fact variable combinations. The retrieval would typically be much faster, vs. a normalized DB schema, as it would require less JOIN operations.

The design of a normalized DB is guided by well-grounded methodologies and supported by helpful tools (e.g., the ERD - Entity-Relationship Diagram). However, the Star-Schema concept does not offer methodological conversion method, but rather commonly guided by a set of good practices and "rules of thumb" that have evolved over the years. This study aims at exploring a novel direction – representation of a relational DB schema as a directed-graph, and a conversion methodology that would permit expert-user intervention. The assumption that underlies this methodology is that some typical DW schema-design decisions cannot be directed by structure and data-type analysis alone, but rather require in-depth understanding business contexts and meaning; hence, likely to mandate expert-user intervention – e.g., adding calculated attributes and aggregations, tracking attribute-value transitions over time (a.k.a., the "slowly changing dimensions" issue), and attribution of fact-variable values. The study aims at formalizing the foundation for the proposed methodology and demonstrating it via a prototype application.