Core Methodologies in Data Warehouse Design and Development

James Yao, Department of Information & Operations Management, Montclair State University, Montclair, NJ, USA

John Wang, Department of Information & Operations Management, Montclair State University, Montclair, NJ, USA

Qiyang Chen, Department of Information & Operations Management, Montclair State University, Montclair, NJ, USA

Ruben Xing, Department of Information & Operations Management, Montclair State University, Montclair, NJ, USA

ABSTRACT

Data warehouse is a system which can integrate heterogeneous data sources to support the decision making process. Data warehouse design is a lengthy, time-consuming, and costly process. There has been a high failure in data warehouse development projects. Thus how to design and develop a data warehouse have become important issues for information systems designers and developers. This paper reviews and discusses some of the core data warehouse design and development methodologies in information system development. The paper presents in particular the most recent and much heated hybrid approach which is a combination of data-driven and requirement-driven approaches.

Keywords: Bottom-Up, Database Design, Data-Driven, Data Warehouse, Demand-Driven, Design Methodologies, Hybrid Approach, Multidimensional Model, Requirement-Driven, Supply-Driven, Top-Down

INTRODUCTION

Information systems were developed in early 1960s to process orders, billings, inventory controls, payrolls, and accounts payables. Soon information systems research began. Harry Stern started the “Information Systems in Management Science” column in Management Science journal to provide a forum for discussion beyond just research papers (Banker & Kauffman, 2004). Ackoff (1967) led the earliest research on management information systems for decision-making purposes and published it in Management Science. Gorry and Scott Morton (1971) first used the term ‘decision support systems’ (DSS) in a paper and constructed a framework for improving management information systems. The

DOI: 10.4018/ijrat.2013010104
topics on information systems and DSS research diversifies. One of the major topics has been on how to get systems design right.

In late 1970s, the growing success of database management systems (DBMSs) proliferated the use of databases in organizations around the world (Takecian et al., 2013). These kinds of databases are designed to handle routine business transactions. Meanwhile, with the growth of the data collected and stored, the need to analyze the data for managerial decision making increased and the drive to optimize the transactional database for analytical purpose intensified. In the early 1990s, data warehouse was coined and later developed. As an active component of DSS, which is part of today’s business intelligence systems, data warehousing became one of the most important developments in the information systems field during the mid-to-late 1990s. Since business environment has become more global, competitive, complex, and volatile, customer relationship management (CRM) and e-commerce initiatives are creating requirements for large, integrated data repositories and advanced analytical capabilities. Data warehouse is a system which can integrate heterogeneous data sources to support the decision making process (Vela, et al., 2013). By using a data warehouse, companies can make decisions about customer-specific strategies such as customer profiling, customer segmentation, and cross-selling analysis (Cunningham et al., 2006). Thus how to design and develop a data warehouse have become important issues for information systems designers and developers.

Data modeling for a data warehouse is different from operational database data modeling. An operational system, e.g., online transaction processing (OLTP), is a system that is used to run a business in real time, based on current data. An OLTP system usually adopts entity-relationship (ER) modeling and application-oriented database design (Han & Kamber, 2006). An information system, like a data warehouse, is designed to support decision making based on historical point-in-time and prediction data for complex queries or data mining applications (Hoffer, et al., 2007). A data warehouse schema is viewed as a dimensional model (Ahmad et al., 2004, Han & Kamber, 2006; Levene & Loizou, 2003). It typically adopts either a star or snowflake schema and a subject-oriented database design (Han & Kamber, 2006). The schema design is the most critical to the design of a data warehouse.

Data warehouse design is a lengthy, time-consuming, and costly process. Any wrongly calculated step can lead to a failure. A study from Gartner Group in 2005 shows that about 50% of data warehouse projects tend to fail due to problems during data warehouse design and construction (Takecian, et al., 2013). Lengthy development process is attributed to the most important cause of the failure. Often, by the time the systems become available some of the functional features are already obsolete. Therefore, researchers have placed important efforts to the study of design and development related issues and methodologies.

Many approaches and methodologies have been proposed in the design and development of data warehouses. Two major data warehouse design methodologies have been paid more attention. Inmon et al. (2000) proposed the Corporate Information Factory (CIF) architecture. This architecture, in the design of the atomic-level data marts, uses denormalized entity-relationship diagram (ERD) schema. Kimball (1996, 1997) proposed multidimensional (MD) architecture. This architecture uses star schema at atomic-level data marts. In recent years, the term data-driven and requirement-driven approaches were used. More recently hybrid model is attempted. Which methodology should an enterprise follow and what is going on in the data warehouse design and development has been brought to the attention of information systems professionals.

Nowadays, it has been widely accepted that the conceptual schema of a data warehouse must follow the MD model (Bellatreche, 2006; Han & Kamber, 2006; Romero & Abello, 2010). Some researchers call this model the data-driven design model. Nevertheless, the requirement-driven model, is also advocated, which, as another approach of data warehouse design,
Related Content

Utilizing Soft Computing Application for QOS and Security Optimization by Meta-Heuristic-Based Genetic Approach
www.irma-international.org/chapter/utilizing-soft-computing-application-for-qos-and-security-optimization-by-meta-heuristic-based-genetic-approach/180067/

Behavioral Path Planning
www.irma-international.org/chapter/behavioral-path-planning/69694/

Analysis of Direct Sensor-to-Embedded Systems Interfacing: A Comparison of Targets’ Performance
www.irma-international.org/article/analysis-direct-sensor-embedded-systems/64218/

Passive Localization of a Robot Using Multiple-View Geometry
www.irma-international.org/article/passive-localization-of-a-robot-using-multiple-view-geometry/132541/

Interactive and Collaborative Virus-Evolutionary CNC Machining Optimization Environment
www.irma-international.org/chapter/interactive-and-collaborative-virus-evolutionary-cnc-machining-optimization-environment/137696/