

Chapter 1

Data Warehouse Architecture: Practices and Trends

Xuegang Huang
Danske Bank Group, Denmark

ABSTRACT

The wide adoption of business intelligence applications has let more and more organizations to build and maintain data warehouse systems. Concepts like “unified view of data” and “one version of the truth” have been the main drive of creating data warehouses. The dynamics of the business world poses the challenges of managing large volume, complex data in data warehouses while the real-time integration and master data needs are presented. This chapter summarizes the past and present patterns of typical data warehouse architectures and describes how the concept of service-oriented architecture influences the future evolvement of data warehouse architecture. The discussion takes many real world requirements in data warehouse solutions and lists considerations on how architecture patterns can solve these requirements.

INTRODUCTION

Over the past decades, the concept of data warehousing has been spread out to everywhere in the business world. Organizations have been practicing hard on achieving successful data warehouse architectures. Lessons have been learnt from those who have succeeded, as well as those who have not. Several key developments in the data warehousing industry denote the past and present of this discipline.

Specifically, starting with very few industry vendors in the 80s such as Teradata, many IT companies like Microsoft, IBM and Oracles are extending their database management systems (DBMSs) to have sufficient support on data warehouses. The early years' data warehousing theory and engineering practices have been well recorded in the publications of Inmon (Inmon, 2005) and Kimball (Kimball & Ross, 2002).

Academic research of data warehousing technologies started in the early 90s. The database research community began with a focus on incor-

DOI: 10.4018/978-1-60566-816-1.ch001

porating data from heterogeneous sources into a single database in order to provide a consistent and unified view of data. These early discussions, such as database snapshots (Adiba & Lindsay, 1980) and materialized views (Gupta & Mumick, 1995), motivated a big variety of subsequent research tracks such as OLAP (online analytical processing) databases (Chaudhuri & Dayal, 1997), data cube (Gray, Bosworth, Layman, & Pirahesh, 1996), multidimensional modeling (Agrawal, Gupta, & Sarawagi, 1997), multidimensional indexing and query optimization (Böhm, Berchtold, Kriegel, & Michel, 2000), and data warehousing for complex data types (Pedersen & Jensen, 1999). The research world has been putting recent attentions on improving the scalability of data warehouses on complex data types (Darmont, Boussaid, Ralaivao, & Aouiche, 2005) and how data warehouse can be seamlessly and efficiently integrated into the business intelligence process and applications (Furtado, 2006; Theodoratos, Ligoudistianos & Sellis, 2001).

Data warehouse architecture is a portfolio of perspectives on how different architecture pieces of a data warehouse system are connected and interacting with each other. It reflects how the academic research and industry development influence the data warehousing practices of different enterprises. For example, from a computing infrastructure perspective, data warehouse architecture has gone from past mainframe analytics to client/middleware/server environment, and now to service-oriented computing as well as the cloud computing concepts. With the rapid growth of information volume and more requirements arriving from the business side, many IT organizations of large business enterprises are facing the challenge of building an enterprise-wide data warehouse that integrates and manages various types of information that comes from different corners of the enterprises and provides the solid information for business analysis in a timely manner. Successful data warehouse architecture must be able to ensure the processing efficiency,

the information correctness, and propagation of metadata while managing over terabytes of data with a daily growth of over gigabytes.

As in the past decade, practices of data warehouse architecture have been focused on addressing classical issues such as the data integration needs, the data quality and metadata control, the data modeling requirements and the performance acceptance from both the data management and the analytical sides. Specifically, the data extraction, transformation, and loading (ETL) process has to manage large volumes of data in an efficient manner by allowing easy and fast scaling up/out hardware configurations. Extraction of metadata and reconciliation of data quality requirements must also be fulfilled through the data integration process in order to enable the data lineage across the whole data lifecycle in the warehouse. An enterprise-wide data model provides unified, consolidated view of the data which enables a consistent, logical representation of business data across different functional areas of a whole enterprise. As the data management side of data warehouse is focused on loading the data in an efficient manner while the analytical users are more interested in retrieving data in a fast and agile way, data warehouse architecture has to enable an easy way of finding the balance of both sides.

Built upon the past decade's research explorations, data warehouse software vendors are instantiating tools and engineering practices on these classical architecture topics. While vendors are rolling out more and more parallel-processing database and ETL engines, enterprise-wide metadata and data quality tools, and eagerly extending their center of excellence with vast amount of data warehousing practices, both the data warehouse industry and academic worlds are facing new challenges when novel concepts such as SOA, web 2.0, and cloud computing are spreading over the whole IT community. This chapter is devoted to addressing what challenges these new trends bring to the data warehouse architecture and how the different academic research can be used to

13 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/data-warehouse-architecture/38216

Related Content

Multidimensional Analysis of Big Data

Salman Ahmed Shaikh, Kousuke Nakabasami, Toshiyuki Amagasa and Hiroyuki Kitagawa (2019). *Emerging Perspectives in Big Data Warehousing* (pp. 198-224).

www.irma-international.org/chapter/multidimensional-analysis-of-big-data/231014

Privacy Implications of Organizational Data Mining

Hamid R. Nemati, Charmion Brathwaite and Kara Harrington (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 2856-2871).

www.irma-international.org/chapter/privacy-implications-organizational-data-mining/7807

An Experimental Replication With Data Warehouse Metrics

Manuel Serrano, Coral Calero and Mario Piattini (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 408-428).

www.irma-international.org/chapter/experimental-replication-data-warehouse-metrics/7656

Mining Clinical Trial Data

Jose Ma. J. Alvir, Javier Cabrera, Frank Caridi and Ha Nguyen (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 3675-3693).

www.irma-international.org/chapter/mining-clinical-trial-data/7857

Data Mining for Intrusion Detection

Aleksandar Lazarevic (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 2880-2887).

www.irma-international.org/chapter/data-mining-intrusion-detection/7809