An Overview on the Design and Modeling of Data Warehouse

Dushyant Singh, Assistant Professor
Department of Computer and Science and Engineering, Vivekananda Global University, Jaipur
Email Id- dushyant.singh@vgu.ac.in

ABSTRACT: ***Data warehousing is a set of decision-making tools designed to help knowledge workers (executives, managers, and analysts) make better and faster decisions. Many commercial products and services are now available, and all of the major database management system vendors offer these services. Multidimensional modeling necessitates the use of specialized design methods. Despite the fact that much has been written about how to design a data warehouse, there is still no consensus on a design method. This paper is based on a broad discussion held in Dagstuhl during the Perspectives Workshop "Data Warehousing at the Crossroads," and it aims to outline some open issues in data warehouse modeling and design. Issues such as conceptual models, logical models, design methods, interoperability, and design for new architectures and applications are all taken into consideration.***

KEYWORDS: ***Data Warehouse, Design, Multidimensional Modeling, OLAP.***

INTRODUCTION

Data warehouses (DWs) are well-known for focusing on decision support rather than transaction support, and for being characterized by an OLAP workload. OLAP applications have traditionally relied on multidimensional modeling, which intuitively represents data as a cube whose cells store events that occurred in the business domain. The adoption of the multidimensional model for DWs has two advantages. On the one hand, it is similar to data analysts' thinking, so it aids users in comprehending data; on the other hand, it aids performance improvement because it's simple structure allows designers to predict users' intentions.

Non-OLTP workloads and multidimensional modeling necessitate specialized design techniques. De-normalization is the most commonly mentioned distinction between transactional database and data warehouse design, but there are several other noteworthy differences. Despite the fact that much has been written about how a DW should be designed, no consensus on a design method has yet emerged. The ability to differentiate between a conceptual design phase and a logical design phase is recognized by most methodologies[1]–[3].

This article arose as an afterthought after a lively debate during the Perspectives Workshop "Data Warehousing at the Crossroads" in Dagstuhl. While the goal of the seminar was to examine current trends in data warehousing and prepare the way for future research in the area, we will concentrate on modeling and design in particular, attempting to answer the question: "Has research on this subject come to an end?" What are your options if that isn't the case?" As a result, in this article, we review certain themes connected to DW modeling and design and highlight the problems that, in our opinion, still need additional investigation, drawing on the productive conversations that took place there among all participants.

Modeling of Concepts:

The goal of conceptual modeling is to achieve independence of implementation problems by defining the warehousing process and architecture in all of its components at a high degree of abstraction. Conceptual modeling is generally acknowledged as a fundamental basis for creating a database that is well-documented and completely meets user needs; it often depends on a graphical notation that allows both designers and users to write, interpret, and manage conceptual schemata[4]–[6].

So far, conceptual modeling for DWs has been approached from primarily two perspectives:

1) *Multidimensional Modeling:*

Existing methods may be divided into three categories: Entity-Relationship model extensions, UML extension extensions, and ad hoc models. While all models have the same fundamental expressivity in that they can all describe the basic ideas of a multidimensional model, the ability to represent more sophisticated concepts such as irregular hierarchies, many-to-many connections, and additivity varies considerably.

2) ETL Modeling:

The emphasis is on modeling the ETL process from a functional, dynamic, or static perspective. Though ETL modeling research is likely less developed than multidimensional modeling research, we think it will have a significant effect on enhancing the overall dependability of the design process and shortening its time.

i. Absence of a Standard:

Despite the fact that many conceptual models have been presented, none have been recognized as a standard, and each manufacturer proposes their own proprietary design techniques.

We suggest that the following are the major causes behind this: I despite the semantic richness of the conceptual models devised, some of the modeled properties cannot be expressed in the target logical models, so the translation from conceptual to logical is incomplete; and (ii) commercial CASE tools lag behind the research and industrial communities in terms of identifying the most relevant multidimensional properties to be modeled. We think, on the other hand, that a unified conceptual model for DWs, implemented inside advanced CASE tools, would be a useful assistance for both the academic and industry sectors.

ii. Security Modeling:

Information security is a must-have for a variety of applications. Among the many elements of security, confidentiality is especially important in the case of DWs, since business information is very sensitive and may be found by running a simple query. Unfortunately, the traditional security paradigm for transactional databases, which is based on tables, rows, and attributes, is incompatible with DWs. Two searches produced via a simple drill-down procedure, for example, may have the same table, rows, and columns, but the one made at the highest aggregate level may expose unwanted data information to the user. As a result, the traditional security model should be replaced with one that is focused on the key ideas of multidimensional modeling – such as facts, dimensions, and measures – and closely integrated with the chosen conceptual model. Furthermore, as is customary in software engineering, information security should be addressed at all phases of the development life cycle, from requirement analysis through implementation and maintenance[7], [8].

Design with mining in mind Commercial solutions from IBM and Microsoft already combine OLAP with data mining. Despite this, the academic community in general, and DW researchers in particular, have not regarded combining OLAP and data mining as a hot subject, with the noteworthy exceptions of Han's OLAM and, more recently, prediction cubes. Until now, DW design has been primarily focused on creating OLAP cubes, with little consideration given to mining needs from the outset.

Modeling in Logic:

The general goal of logical modeling is to convert conceptual schemata into logical schemata that may be optimized for and implemented on a selected target system after the conceptual modeling phase is finished. In the field of multidimensional modeling, where target database systems are usually either relational or multidimensional, significant progress has been achieved. The so-called star, constellation, and snowflake schemata are widely recognized and supported by different manufacturers in relational systems for managing data cubes. Several efficient multidimensional data structures, such as condensed cubes, dwarfs, and QC-Trees, have been suggested to handle data cubes in multidimensional implementations.

i. Semantic Gap:

There is still a semantic gap between sophisticated conceptual data models and relational or multidimensional implementations of data cubes when it comes to fact modeling. In OLAP hierarchies, for example, no commercial solutions can handle generalization/specialization connections. Furthermore, how to represent dimension restrictions or even less expressive context dependencies, both of which explain the presence of null values in dimensions in logical implementations and enable reasoning about summarizability with regard to sets of attributes, seems to be an unresolved issue. Furthermore, a systematic approach of summarizability that takes into account generic aggregate functions beyond SUM is still a work in progress.

ii. Modeling using ETL:

The conversions of conceptual ETL schemata to logical ETL schemata, as well as their optimization, are poorly understood. Indeed, although it seems to be the sole design approach that incorporates an algorithmic translation of conceptual into logical models; it looks to be the first step toward modeling and optimization of ETL operations at the logical level. Furthermore, research on DW self-maintainability and independence has shown how to build up DWs such that maintenance procedures may be simplified and made more efficient by eliminating maintenance inquiries. However, there is still a need to combine these findings with ETL modeling methods.

Interoperability and Metadata:

The wide variety of tools and software products available on the market, combined with the heterogeneity of conceptual and logical models proposed for DWs, has resulted in a wide variety of metadata modeling. In reality, tools with disparate metadata are linked together by constructing sophisticated metadata bridges, but some information is lost in the process of converting one type of metadata to another. As a result, a standard definition of metadata is required to better support DW interoperability and integration, which is especially important in the case of mergers and acquisitions, which occur frequently.

In this context, two industry standards developed by multi-vendor organizations have emerged: the Meta Data Coalition's Open Information Model (OIM) and the OMG's Common Warehouse Metamodel (CWM). MDC joined the OMG in 2000 to help create the CWM as a common metadata format. The CWM is a platform-independent metamodel definition for transferring data warehouse requirements across platforms and tools. It's based on the UML, XMI, and MOF standards, and it's basically a set of metamodels that can model an entire DW, including data sources, ETL, multidimensional cubes, relational implementations, and so on[9]–[11].

New Architectures and Applications Design:

i. Data warehousing in space:

Spatial data is heavily emphasized in spatial DWs, which may take the shape of spatial dimensions or spatial measurements. Several studies, such as this one, demonstrate the benefits of using Geographic Information Systems (GIS) features in the analysis of multidimensional data in particular fields. Other efforts, such as, combined GIS with OLAP to create more broad systems. While all existing conceptual models support basic spatial modeling (for example, most business DWs include a geographic hierarchy based on customers), location data is typically represented in an alphanumeric format. Choosing a more expressive and intuitive form for this data, on the other hand, might show patterns that would be difficult to find otherwise.

ii. Web Warehousing:

Web warehouses are data warehouses that gather data from the internet. Because of the semi-structured nature of data, the lack of control over the sources, and the frequency with which they change, the features of the Web provide additional challenges. The main challenges in this field are integrating heterogeneous web sources and automating the conceptual design process when some or all data sources are on the Internet. Some efforts have been undertaken in this approach, with the primary goal of creating a conceptual schema from XML data.

iii. BPM and Real-Time Data Warehousing:

Because DW systems offer an integrated picture of an organization, they are an excellent place to start when developing a platform for business process monitoring (BPM). BPM on top of a DW, on the other hand, has a significant impact on design and modeling, as it necessitates extended architectures that may include components not found in standard DW architectures and may be fed by non-standard data sources (such as data streams). Because BPM necessitates real-time requirements, ETL components must be rethought, making the ETL design techniques used thus far suspect.

iv. Data Warehousing on a large scale:

In the same way that distributed databases require a new phase to be added to the design method, distributed data warehousing requires a new phase to be added to the design method: the one for designing the distribution from both an architectural and physical standpoint. General decisions will be made during architectural design about which distribution paradigm (P2P, federation, grid) best suits the requirements, how to deploy the DW on the infrastructure, which communication protocols to use, and so on.

DISCUSSION

A data warehouse is defined as a "subject-oriented, integrated, time-varying, non-volatile collection of data used primarily in organizational decision-making." A data warehouse stores and feeds BI and analytics with current and historical data for the whole company. Data warehouses use a database server to pull data from an organization's databases, as well as data modeling, data lifecycle management, data source integration, and other features. Manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), and utilities (for call analysis and fraud detection) are just a few of the industries that have successfully implemented data warehousing technologies (for outcomes analysis).

CONCLUSION

Author addressed outstanding problems in the modeling and design of DWs in this article. Despite the fact that these subjects have been studied for over a decade, many significant difficulties remain. Ad hoc methods are also needed for coping with growing data warehousing applications and sophisticated business intelligence systems. Furthermore, the requirement for real-time data processing raises new issues that aren't addressed by traditional periodically-refreshed DWs. Overall, we believe that DW modeling and design research is far from dead, partly due to the need for more sophisticated techniques to solve known problems, and partly due to the new problems that arise during the adaptation of DWs to the unique requirements of today's business.

REFERENCES:

[1] S. Rizzi, A. Abelló, J. Lechtenbörger, and J. Trujillo, “Research in data warehouse modeling and design: Dead or alive?,” 2006, doi: 10.1145/1183512.1183515.

[2] L. Berrahou *et al.*, “A quality-aware spatial data warehouse for querying hydroecological data,” *Comput. Geosci.*, 2015, doi: 10.1016/j.cageo.2015.09.012.

[3] M. Goller and S. Berger, “Handling measurement function changes with Slowly Changing Measures,” *Inf. Syst.*, 2015, doi: 10.1016/j.is.2014.12.009.

[4] M. M Kirmani, “Dimensional Modeling Using Star Schema for Data Warehouse Creation,” *Orient. J. Comput. Sci. Technol.*, 2017, doi: 10.13005/ojcst/10.04.07.

[5] M. Rifaie, K. Kianmehr, R. Alhajj, and M. J. Ridley, “Data modelling for effective data warehouse architecture and design,” *Int. J. Inf. Decis. Sci.*, 2009, doi: 10.1504/IJIDS.2009.027656.

[6] R. K. Pandey, “Data Quality in Data warehouse: problems and solution,” *IOSR J. Comput. Eng.*, 2014, doi: 10.9790/0661-16141824.

[7] W. Astriani and R. Trisminingsih, “Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle,” *Procedia Environ. Sci.*, 2016, doi: 10.1016/j.proenv.2016.03.117.

[8] R. Gill and J. Singh, “A Review of Contemporary Data Quality Issues in Data Warehouse ETL Environment,” *J. Today’s Ideas - Tomorrow’s Technol.*, 2014, doi: 10.15415/jotitt.2014.22012.

[9] E. Fernández-Medina, J. Trujillo, R. Villarroel, and M. Piattini, “Access control and audit model for the multidimensional modeling of data warehouses,” *Decis. Support Syst.*, 2006, doi: 10.1016/j.dss.2005.10.008.

[10] O. Romero and A. Abelló, “A Survey of Multidimensional Modeling Methodologies,” *Int. J. Data Warehous. Min.*, 2009, doi: 10.4018/jdwm.2009040101.

[11] E. P. Putra, F. Fifilia, L. Christian, and H. Sudarma, “Modelling of Data Warehouse on Food Distribution Center and Reserves in the Ministry of Agriculture,” *ComTech Comput. Math. Eng. Appl.*, 2015, doi: 10.21512/comtech.v6i3.2251.