www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 12 Issue 11 November 2023, Page No. 25884-25888
ISSN: 2319-7242 DOI: 10.18535/ijecs/v12i11.4767

Data Warehouse Metadata Management: Ensuring Data Quality and Governance

Sneha Mondal[1], Anirban Pal[2], Anirban Bhar[3], Sujata Kundu[4]

[1,2] B. Tech student, Department of Information Technology, Narula Institute of Technology, Kolkata, India.
[3,4] Assistant Professor, Department of Information Technology, Narula Institute of Technology, Kolkata, India.

Abstract

In an age defined by the abundance of data, organizations face the formidable task of not only collecting but also extracting valuable insights from diverse sources of information. Data warehousing, a critical component of modern data management, emerges as a strategic solution to this challenge. This abstract provides a concise overview of the concept, significance, and primary features of data warehousing.

A data warehouse is a centralized repository that facilitates the collection, storage, and management of data from multiple sources. Unlike operational databases, which are designed for transactional processing, data warehouses are tailored for analytical purposes. They serve as a dedicated platform for data consolidation, transformation, and the provision of a consistent, accessible dataset.

The principal objective of a data warehouse is to offer a comprehensive and unified view of an organization's data, enabling decision-makers to extract valuable insights, detect trends, and make informed, data-driven decisions. As a cornerstone of business intelligence and analytics, data warehouses empower organizations to convert their data assets into a competitive advantage.

Keywords: Data Integration, Data Modeling, OLAP, Data Mart, Query Optimization, Data Quality.

1. Introduction

Data warehouse metadata refers to information that is stored in specialized metadata repositories. This information includes details about the contents of the data warehouse, such as their location and structure. It also encompasses information about the processes involved in the data warehouse, specifically regarding the refreshment of the warehouse with accurate and up-to-date data that is semantically and structurally reconciled. Furthermore, data warehouse metadata includes information about the implicit semantics of the data, in relation to a common enterprise model. This encompasses any other type of data that assists end-users in effectively utilizing the information within the warehouse. Additionally, it encompasses information about the infrastructure and physical characteristics of the components and sources of the data warehouse. Lastly, data warehouse metadata includes information related to security, authentication, and usage statistics. This information aids administrators in optimizing the operation of the data warehouse as needed.

In today's data-driven world, organizations collect and store vast amounts of information from diverse sources. This data encompasses everything from customer transactions and website interactions to supply chain details and financial records. While the accumulation of data is essential, the real challenge lies in turning this raw information into actionable insights for informed decision-making.

This is where the concept of a data warehouse comes into play. A data warehouse is a central repository for collecting, storing, and managing data from multiple sources. Unlike transactional databases, which are designed for day-to-day operations, data warehouses are optimized for analytical processing. They enable organizations to consolidate data, transform it into a consistent format, and make it readily accessible for reporting and analysis.

The primary purpose of a data warehouse is to provide decision-makers with a comprehensive and unified view of an organization's data, allowing them to gain valuable insights, detect patterns, and make informed strategic decisions. This concept has become increasingly crucial in an era where data is one of the most valuable assets a company can possess.

2. Historical Background

Data warehouses are extremely complicated systems, both to build and maintain. Multiple data sources, often located in separate operational contexts, and a large number of clients, each with their own unique needs and methods of accessing the data warehouse round out the picture. The intricacy of the infrastructure is only one part of the problem; the management of the data involved in the warehouse environment is where much of the difficulty lies. Different types of source data, each with its own format, structure, and hidden semantics, are combined in a centralized data warehouse and then disseminated to various end-users, each of whom has a unique understanding of the terminology and semantics underlying the data's structure and content. The following challenges must be well understood by administrators, designers, and application developers who work together to deliver fresh, up-to-date, consolidated, and unambiguous data from the sources to the end-users.

− the location of the data,

− the structure of each involved data source,

− the operations that take place towards the propagation, cleaning, transformation and consolidation of the data towards the central warehouse,

− any audit information concerning who has been using the warehouse and in what ways, so that its performance can be tuned,

− the way the structure (e.g., relational attributes) of each data repository is related to a common model that characterizes each module of information.

A data warehouse metadata repository is a centralized location for storing and retrieving this metadata, making it easily accessible to everyone with a stake in a data warehouse.

As is the case with the rest of the data warehousing field, the problem of the form and management of data warehouse metadata was first addressed by ad hoc solutions given by industrial vendors and consultants. Metadata were not given first-class consideration in early attempts at both academic and industrial standardization (e.g., the MDIS standard [1]) related to wrapper-mediator schemes of information integration (Information Manifold, WHIPS, Squirrel, TSIMMIS -- see for a detailed discussion of the related literature).

The European Project "Foundations of Data Warehouse Quality (DWQ)" marked the first serious attempt to tackle the issue of metadata management in data warehouses. [2, 3]. The vertical lines in Fig. 1 denote increasing layers of abstraction; for example, the data warehouse metadata repository shown in the middle layer is an approximation of the actual structure of the warehouse environment shown in the bottom layer. Additionally, [2] used the Telos programming language to address the difficulty of developing an adequate formalism for representing the repository's contents (seen in Fig. 1's upper layer).

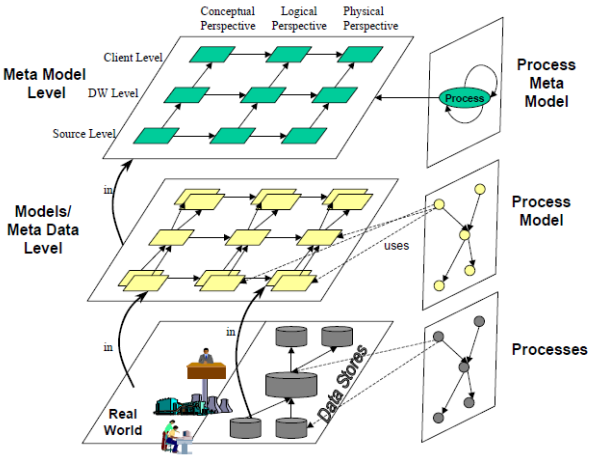

Fig.1. Role and structure of a data warehouse metadata repository [4]

Data warehouse as a concept has been the subject of a lot of research. Only works that suggest metadata for the Data Warehouse and Data Warehouse modeling are discussed in any detail.

In [5] offer a metadata strategy for Data Warehouse security, but do not go beyond technical information with business-oriented string labels and descriptions of attribute and table names.

Eight layers, including a metadata layer, are presented for Data Warehouse design in [6]. The layers here stand in for the enterprise's overarching data, communication, processing, and presentation architecture for computing at the end user level.

In [7], the logical architecture operates apart from the front- and back-end technologies used to implement it. The physical architectures are translations of the logical design into MDBMS and RDBMS, respectively.

Metadata for data warehousing was modeled using a variety of presented in [8] expanded relational ideas. A significant benefit of the extended relational model is highlighted by the comparison of the two models.

3. Metadata Creation and Cleaning

Metadata are quite important to a DW project. However, many Data Warehousing projects struggle with metadata management because of the wide variety of tools and products available for this purpose. This is due to the fact that the business world does not agree on a single standard for the definition and exchange of metadata.

Despite these challenges, due to the significance of metadata in an analytical setting, certain works have been developed to aid in the administration of metadata in DW. Metadata production and administration are not incorporated into the development process in any of the reviewed works [9].

Metadata is required to facilitate communication between different warehouse function areas and an ETL tool (Extraction, Transformation, and Load) is required to describe the warehousing process in order to provide a decisional database. In order to tell the ETL tool how to map each attribute, developers create a mapping guideline [10].

4. Metadata Management in Data Warehouses

As new problems arise in the data warehouse (DW), one of the efficiencies that has been uncovered is metadata. However, the main causes of its usefulness are the strategies used in its management. Therefore, how do different metadata management strategies influence the functioning of businesses? The purpose of this research is to examine the outcomes of various approaches to metadata management.

In the 1990s, it became clear that metadata played a crucial role. Metadata in the context of a data warehouse is defined by Palepu & Sambasiva (2012) as "everything in the data warehouse environment other than the actual

data." Metadata in a data warehouse can be divided into two broad categories: front-facing and back-facing. It's process-related, thus it directs actions like loading and extracting. Metadata staging, including specifications for data cleaning, changing policies of dimensions slowly, and so on; logs of data transformations; and DBMS system table contents are just a few examples of specifications that relate to source data(Mohammed et al., 2012, p. 44). However, front-of-house metadata delves deeply into description and is thus crucial to the efficient operation of report-writers and tools via which queries can be made. Front-desk metadata examples include join requirements, access and usage maps, and network security usage statistics.

Data warehouse administration and management cannot take place without metadata. Users and decision-makers who consume massive amounts of data from a warehouse might likewise benefit from metadata (Anand, 2014, p. 12). Metadata's worth may be demonstrated from three different angles: data management, knowledge management, and data quality management.

Metadata plays a crucial role in data management and maximizing the value of collected information. Organizing as well as "cataloguing" the data offers much efficiency when searching, control 4 over data redundancy, preservation of the integrity and on-going maintenance. Thanks to abstraction, data can be handled separately from the systems and applications that use it.

5. Establishing a Framework for Data Governance

In order for a data governance framework to be understood, accepted, and embraced within an organization, it needs to be adapted to the culture and structure of that organization and developed in a participatory manner. Only then can the organization hope to realize its full potential. The Data Governance Institute has a website that contains resources to support data governance projects. These resources include whitepapers, case studies, best practices and non-technical briefings on data-related issues, a framework to assist with thinking and communicating the concept, along with a 'how to' guide and who-what-when-where-why-how information about data governance (Data Governance Institute, n.d.a.). These resources can be found at http://www.datagovernanceinstitute. • Top-down, in which decisions are taken by senior executives and then conveyed to lower-level employees; this is one of the models for data governance.

• Bottom-up—where data-related choices are made by individuals or organizations at the local level and are communicated up. The focus here is more on the processes involved than on the policies themselves.

• Centre-out, in which decisions are made by one or more senior members of staff in a central unit that is the most conversant with the topic and who makes a suggestion about what is best for the organization.

• Silo-in refers to a situation in which representatives from several groups come to an agreement on a course of action that takes into account both the requirements of each group and the needs of the organization as a whole. In this scenario, the group in question has either been given the authority to make decisions or the ability to make suggestions.

Establishing numerous communication channels and educating stakeholders about the reasons decisions have been made is essential to achieving buy-in and compliance in any of the models that were mentioned above.

6. Conclusion

In conclusion, a data warehouse is a cornerstone of modern data management and analytics strategies. It serves as a centralized repository for organized data, facilitating efficient storage, retrieval, and analysis. Data warehouses are instrumental in providing high-quality, consistent, and historical data for informed decision-making. They empower organizations to extract valuable insights from their data, whether through complex queries, business intelligence tools, or real-time analysis.

Data warehouses also play a pivotal role in ensuring data governance, security, and compliance, which is essential in the era of stringent data regulations. With cloud-based solutions and scalable architecture, they offer cost-efficient and adaptable options, enabling organizations to keep pace with evolving data needs. Additionally, the integration of emerging technologies such as AI, machine learning, and edge computing positions data warehouses at the forefront of data-driven innovation.

In a data-centric world, data warehouses are indispensable for organizations seeking to harness the full potential of their information assets. They empower businesses to make data-driven decisions, gain

competitive advantages, and navigate the complexities of a rapidly changing data landscape, making them a fundamental component of the modern digital enterprise.

References

1. Metadata Coalition: Proposal for version 1.0 metadata interchange specification. http://www.metadata.org/standards/toc.html, July 1996.
2. M. Jarke, M.A. Jeusfeld, C. Quix, P. Vassiliadis. Architecture and quality in data warehouses. In Proc. 10th Conference on Advanced Information Systems Engineering (CAiSE '98), pp. 93-113, Pisa, Italy, June, 8-12, 1998. Lecture Notes in Computer Science, vol. 1413, Springer, 1998.
3. Foundations of Data Warehouse Quality (DWQ) homepage. Available at http://www.dblab.ece.ntua.gr/~dwq/
4. M. Jarke, M.A. Jeusfeld, C. Quix, P. Vassiliadis. Architecture and quality in data warehouses. Information Systems, vol. 24, no. 3, pp. 229-253, May 1999. Elsevier Science Ltd. ISSN 0306-4379.
5. Katic, N.; Quirchmayr, G.; Schiefer, J.; Stolba, M.; Tjoa, A M.: A Prototype Model for Data Warehouse Security Based on Metadata. Proceedings DEXA 98.
6. Ken Orr, Data Warehousing Technology, The Ken Orr Institute, A white paper, 1996.
7. M. Wu, A. P. Buchmann, Research Issues in Data Warehousing, BTW 1997: 61-82.
8. O. Mangisengi, A M. Tjoa, R. R. Wagner, Metadata for Data Warehouses Using Extended Relational Models Proc. of third IEEE Computer Society Metadata Conference, April 1999.
9. Que, The Official Client/Server Computing Guide to Data Warehousing, Que Books, 1997.
10. Thanh N. Huynh, Oscar Mangisengi, and A Min Tjoa, Metadata for Object-Relational Data Warehouse, Vienna University of Technology, Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) Stockholm, Sweden, 2000.