Metadata Management Commonalities between Data Warehouse and Big Data Systems

Hanif Lumsden

University of Maryland Global Campus

DBST 665 9040 Data Warehouse Technologies

Dr. Kuchibhotla

08/2/2022

Table of Contents

Abstract

Data warehouse metadata are bits of information stored in a data repository. Metadata is not readily available, so numerous tools used by extract-transform-load developers and IT managers, usually a metadata coordinator, can make said metadata available, which offers insight for business decision-makers. The metadata contains information on data warehouse contents, processes, semantics, and infrastructure. Architecture components and OLAP analysis is based on metadata; For big data stores, the architecture contains several types of corresponding metadata for the operation of architectural components. Research exists regarding metadata frameworks, management systems, and the evolution of handling data warehouse metadata. This research reviews the development of metadata management in data warehouse repositories. What is the current infrastructure of metadata management solutions (MMS) of data warehouses, and how does it compare to MMS in big data?

Keywords: Data warehouse, metadata, data architecture, mms, big data

Introduction

Definition and Historical Background

Metadata exists to give descriptive information regarding data regarding its format, definition, and temporal attributes (Sarda, 2001; Xiao, 2015). Metadata rationale stems from the dissemination of libraries and information and the constant growth of such since the 1960s when there was a need to organize and give context to data (Zeng & Qin, 2016). Metadata came to fruition to manage digital resources (Zeng & Qin, 2016). In the context of data warehouses, the nature of its intricacy in the facet of its design and operation, such as the extraction of data comprised of many different sources of various formats to the easy mismanagement as a result of different semantics, metadata is critical for managed data. Metadata stores data warehouse information, structure, and paths at a central reference point (Vassiliadis, 2018).

Initially, the earlier instances to address metadata and data descriptions were to provide impromptu solutions that never tended to the problem of structure (Vassiliadis, 2018). These solutions can come from the industrial sector called data dictionaries used for detailed data archiving and analysis (Xiao et al., 2015). In relational databases, the data dictionary concept forms a descriptive data definition language where tables, constraints, views, and columns, are created (Xiao et al., 2015); However, it was not until academia offered a solution for structure and metadata management in 1997 by Fabreta et al. called the data warehouse quality (DWQ) framework (Vassiliadis, 2018). Through revision, the DWQ framework became manageable. The DWQ, seen in **figure 1**, is based on data and process classification on the conceptual, logical, and physical model levels (Vassiliadis, 2018). The data schematics and data properties such as indexes, data presentation, and storage location are metadata represented in the logical and physical (Vassiliadis, 2018).

Extract-transform-load (ETL) processes are the primary source of metadata repository entries as it is responsible for the generation and usage of metadata in the data warehouse or business intelligence environment (Kimball & Ross, 2013; Vassiliadis, 2018). Metadata management ensures consistency, integrated enterprise information sharing, and defining relationships. Enterprise information sharing is centralized reference models expressed as a data warehouse view. A simplified view to extract analysis is the central role of a warehouse and the purpose for why it is created (Xiao et al., 2015). Warehouses are not treated as relational databases where the data is often manipulated through various commands, but rather, the warehouse exists for its analytical use-value. The previously mentioned data definition is not applicable in a grand sense to data warehouses where vast amounts of business data are held, so the requirements change: metadata management systems manage the description of data properties that supply relevant business information to its users (Xiao et al., 2015). MMS is further defined in data warehouses and big data systems, and its evolution is evaluated; finally, the current infrastructure of the data warehouses and big data systems are compared and assessed.

Metadata Management Review

Metadata Management in Data Warehouses

Metadata management systems (MMS) are essential in data warehouses for stakeholders to come to conclusions and make decisions (van Zyl et al., 1998). With a MMS, processes such as ETL, data models, and the entire process from system implementation to usage (Xiao et al., 2015). The data warehouse metadata management environment is used for analysis (van Zyl et al., 1998; Xiao, 2015). The extent of the analysis considers impact, operational sources, tables, history, and owners (van Zyl et al., 1998). Impact analysis of operational data source changes summarized tables, the changes, how data is changed, and data owners partitioned by specialization (van Zyl et al., 1998). An analysis is only possible with a metadata management system.

MMS offers an accessible view of data on the enterprise level to share information with stakeholders; as a result, the accuracy, integrity, and consistency are managed and facilitated, all while sustaining the constant development and upgrades to the data warehouse system (Xiao et al., 2015). Effectively implementing and using metadata management will, in turn, reduce data warehouse workload, improve the extract-transform-load operation, and improve overall efficiency (Xiao et al., 2015). MMS functions are comprised of: descriptions of data in the data warehouse; definition of generated and entered data from the warehouse; work schedule extractions on business events and data conduct; data consistency system performance detection; quality of data (Xiao et al., 2015, p. 930). MMS is partitioned between two types to be collected: technical and business.

Technical and Business Metadata

Technical metadata relates to data warehouse structure and elements such as relations, tables, data types, and storage (Sarda, 2001; Vassiliadis, 2018). It is the data warehouse system's "development, management, and maintenance" that provides solutions by connecting tools, the ETL process, physical descriptions, and the program (Xiao et al., 2015, p. 930; van Zyl et al., 1998) and is described by entities and their attributes. Business metadata, also dubbed semantic metadata, is relevant for business analysts and aids in creating an enterprise data view (Xiao et al., 2015). This sort of metadata is the essence, frames of reference, and the description of the complex networks of business and all aspects such as framework and functions (Sarda, 2001). Both technical and business metadata are mapped together (van Zyl et al., 1998). Technical metadata fundamentally pertains to the back end, while business metadata concerns the front end. Both aspects form an amalgam of what metadata is. Metadata as business praxis relies on these two aspects to operating and, as anything physical entity, changes through time.

Frameworks of Metadata Management Systems and its Creation

In MMS for data warehouses, there are three standard frameworks: centralized, distributed, and federal, as seen in **Figure 2**. A centralized framework is a singularly focused metadata repository model suitable for moderate-sized organizations or companies (Vassiliadis, 2018). Distributed frameworks manage metadata as separate but interconnected entities using local meta databases to manage data (Vassiliadis, 2018, p. 931). A federal structure synthesizes centralized and distributed methodologies composed of several independently managed meta databases (Vassiliadis, 2018). The federal method is the preferred management system framework for data warehouses (Vassiliadis, 2018). This is justified through the limited centralized and distributed systems if only relying on either one. A combination of the two makes sense for performance reasons, as fewer limitations exist.

Regarding consideration for creating an MMS, it goes back to the requirements that led to a business wanting to implement a data warehouse solution in the first place; first, the scope of the data warehouse and its domain must be defined to determine the bounds of the metadata management system (Vassiliadis, 2018). The metadata functions in the extracted data source and the tools to manage data warehouse metadata are also considered (Vassiliadis, 2018). Enterprise models will aid in homogenizing metadata management; subsequently, the enterprise system will be combined with it (Vassiliadis, 2018).

Metadata Management in Big Data

Big data is the emerging concept of the 21st century to describe data treated mathematically as continuous growth within enterprises. The attributes that describe big data are the four v's: volume, velocity, variety, and value (Dijcks, 2013, as cited in De Mauro et al., 2016). Large volume exists, and a variety of data is processed, requiring high amounts of velocity to achieve business value; furthermore, depending on the business requirements, big data requires specific technologies and analytical methods to achieve said value (De Mauro et al., 2016). In essence, metadata management systems greatly facilitate data changes in a secure manner (Patil & Thakore, 2017).

Many big data solutions exist with no elaborate metadata management support to accentuate the schemaless conditions of big data environments, considering applications that proclaim schemaless features like Hadoop (Smith et al., 2014). The user base of these solutions can be described as small or not working in an extensive network, as it is presumed to be nearly impossible to manage a large project without metadata management to aid in organization and information sharing sorting. There may exist a need to incorporate metadata management for big data to have "version histories [or] data block features" to secure logical cogency (Smith et al., 2014, p. 1). Metadata management can aid in rigid machine-learning algorithms and their widespread integration into analytical mediums, avoiding hard-coding regarding machine-learning algorithms, organizing various data sources, and sharing information (Holom et al., 2020). Hadoop and MapReduce are practical applications to manage metadata in big data systems (De Mauro et al., 2016).

Even on the big data scale, metadata is still essentially used for recording keeping and descriptive properties to speed up queries and public access to information (Edara &

Pasumansky, 2021). Where indexing is standard in relational database systems and data warehouses and relies on block deletion and compilers, big data systems using open-source applications store minor portions of the metadata system like tabular schemas into a centralized, often distributed management service and major portions such as block-level metadata for scalability reasons (Edara & Pasumansky, 2021).

Block level, or a cloud service that emulates physical hard drives, is virtualization as a service that can host metadata to manage through different approaches (Edara & Pasumansky, 2021; Xiao et al., 2006). Metadata management at the block level compensates for the I/O device requirement. Metadata contains the path and qualities of the blocks or multiple instances and records of data grouped (Xiao et al., 2006). Block-level metadata management consists of codes focusing on data allocation (Xiao et al., 2006).

Regarding the different treatment of minor and major portions, the minor portion allocations can be modified without the block data being altered, allowing for scalable mutations (Edara & Pasumansky, 2021). For this metadata management system, distributed data queries are affected since the metadata is unavailable before every block data's header or footer is scanned (Edara & Pasumansky, 2021). It can be surmised that the described availability and method will amount to many costs to run. Programs attempt to assign a centralized condition to the minor allocation at the cost of scalability; however, with metadata management being worked with as data management, a balance can be met between scalability and query performance (Edara & Pasumansky, 2021).

Future of Metadata Management

Various approaches to metadata management will remain a consideration as data grows and needs to be managed and analyzed in some way. Metadata must be integrated into some form. The importance is stressed due to the rise in "web applications" or "service-oriented architecture" (Quix, 2018, p. 2237) and the consolidation of databases or data warehouses with these implementations. The federation method successfully uses the loosely coupled metadata repository (Quix, 2018).

Due to the complexity of metadata management, there are standards to simplify metadata integration in data warehouses. The typical warehouse metamodel, seen in **figure 1**, is a standard based on unified modeling language, meta object facility, and XML metadata interchange that data warehouse merchants conform to that dictate how the operation and maintenance are managed; in addition, the ETL process is fully supported and associates a transformation with event sets to trigger transformation steps (Peyton, 2018). The future direction of metadata repositories examines metadata querying as integrated from many sources (Quix, 2018). Essentially, the aim is to have generic structures for model and map representations (Quix, 2018).

Discussion

At the core, metadata is crucial for data analysis. Analysis, or the intent of analysis, is one of the many reasons why data repositories are kept or reserved outside of record keeping and organization. The analysis brings value because it leads to profitable decision-making and management. As stated by the various authors cited in this research, data is composed of metadata (Sarda, 2001; Xiao, 2015; Zeng & Qin, 2016; Vassiliadis, 2018, van Zyl et al., 1998); where the 'meta' means self-referential, metadata refers to self-referential or self-descriptive data of the data in question. Due to the relevance of metadata in successful analysis and because data warehouses provide powerful analytical capabilities and aid in machine learning algorithms for big data systems, metadata management is critical. Indeed, metadata has come a long way since the first instances of metadata since data dictionaries were used to archive and analyze.

The current infrastructure of metadata management solutions is based on a centralized, distributed, or federated framework, with said federated framework being the most predominant as it is an amalgam of centralized and distributed features. As this is the usual framework of data warehouses, along with the typical warehouse metamodel as a standard, MMSs and how they function within the business they operate rely on these frameworks and standards. Modern solutions like Informatica allow organizations to sustain a federated data warehouse (Reza, 2021). Businesses that implement federated data warehouses will reduce execution time and lower costs (Jindal & Acharya, 2004).

Regarding big data systems, metadata management systems need to take a different approach due to the schemaless nature of the NoSQL solutions. Metadata management leads to a storage infrastructure separating the small systems into a distributed partition and utilizing block-level cloud technology for large systems. In addition, this approach to metadata requires different optimization techniques for execution costs. NoSQL applications BigQuery, Snowflake, and Redshift are other solutions with a metadata management framework like this (Edara & Pasumansky, 2019).

Based on the research, big data's MMS framework holds similarities to MMS for the data warehouse. This is based on referencing how small-scale partitions are centralized and distributed. Following the conventional federated method, data warehouses are virtual, centralized, and distributed. The use of block-level is not conventional or regularly seen in data warehouse MMS. Of course, it depends on how the data warehouse is implemented. Since block-level is the virtualization of physical storage, this is not applicable for strictly on-premise data warehouses. A cloud-based solution may consider block-level partition but will only act as a trade-off if the amount of data is significant. Businesses must consider that block-level and federated combinations may cost large execution if the complex system is not optimized.

Data warehouse and big data are differentiated in this report, but that is not to say that the two approaches cannot potentially intertwine. This has to deal with changing the requirements of data warehouses and modifying data warehouse business objectives to coordinate with big data ones. Objectives can range not only from business goals but from altering data types, ETL strategy, distributed coding, and other components to fit the big data model while keeping the data warehouse intact (Bellatreche & Chakravarthy, 2019). Businesses seek data warehouses to analyze and store the business' historical data. Big data can operate with a data warehouse as a composite arrangement. The mentioned above small and large-scale partitions may operate as stated: small partitions or relational and operational data stores will act as a data warehouse, whereas large partitions are governed by a highly scalable solution such as a NoSQL framework.

Conclusion

Metadata is defined, and metadata management systems are evaluated for data warehouses and big data. Metadata management is essential for businesses to extract use-value from data. MMS is reviewed in this report. Big data and data warehouses are different systems for different requirements, but similarities in metadata management are shared regarding the federated management method.

References

Bellatreche, L., & Chakravarthy, S. (2019). A special issue in extending data warehouses to big data analytics. Distributed and Parallel Databases, 37(3), 323-327. https://doi.org/10.1007/s10619-019-07262-1

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. Library Review, 65(3), 122-135. https://doi.org/10.1108/lr-06-2015-0061

Edara, P., & Pasumansky, M. (2021). Big Metadata: When Metadata is Big Data. Proceedings of the VLDB Endowment, 14(12), 3083-3095. https://doi.org/10.14778/3476311.3476385

Holom, R., Rafetseder, K., Kritzinger, S., & Sehrschön, H. (2020). Metadata management in a big data infrastructure. *Procedia Manufacturing*, *42*, 375-382. https://doi.org/10.1016/j.promfg.2020.02.060

Jindal, R., & Acharya, A. (2004). Federated Data Warehouse Architecture [White Paper]. Wipro Technologies. http://hosteddocs.ittoolbox.com/Federated%20data%20Warehouse%20Architecture.pdf

Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling. John Wiley & Sons.

Patil, M. A., & Thakore, D. M. (2017). A big data analytics - challenges with-in new data, metadata management & analysis platforms. *IARJSET*, *4*(4), 140-142. https://doi.org/10.17148/iarjset/nciarcse.2017.41

Peyton, L. (2018). Common Warehouse Metamodel. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of database systems (2nd ed., pp. 891-897). Springer.

Quix, C. (2018). Meta Data Repository. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of database systems (2nd ed., pp. 2233-2237). Springer

Reza, M. (2021, June 27). Migrating to cloud with code-free data ingestion. Enterprise Cloud Data Management | Informatica. https://www.informatica.com/blogs/guide-to-code-free-data-ingestion-for-your-cloud-modernization-journey.html

Sarda, N. L. (2001). Structuring business metadata in data warehouse systems for effective business support. Computing Research Repository. https://doi.org/10.48550/arXiv.cs/0110020

Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., Sexton, M., & Eckstein, A. (2014). "Big metadata": The need for principled metadata management in big data ecosystems. *Proceedings of Workshop on Data analytics in the Cloud - DanaC'14*. https://doi.org/10.1145/2627770.2627776

van Zyl, J., Vincent, M., & Mohania, M. (1998, December). *Representation of metadata in a data warehouse* [Paper presentation]. Proceedings of IEEE TENCON '98. IEEE Region 10 International Conference on Global Connectivity in Energy, Computer, Communication and Control, New Delhi, India. https://doi.org/10.1109/TENCON.1998.797089

Vassiliadis, P. (2018). Data warehouse metadata. In L. Liu & M. T. Özsu (Eds.), Encyclopedia of database systems (2nd ed., pp. 891-897). Springer. https://doi.org/10.1007/978-1-4614-8265-9_912

Xiao, B., Zhang, C., Mao, Y., & Qian, G. (2015). Review and exploration of metadata management in data warehouse. 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA). https://doi.org/10.1109/iciea.2015.7334243

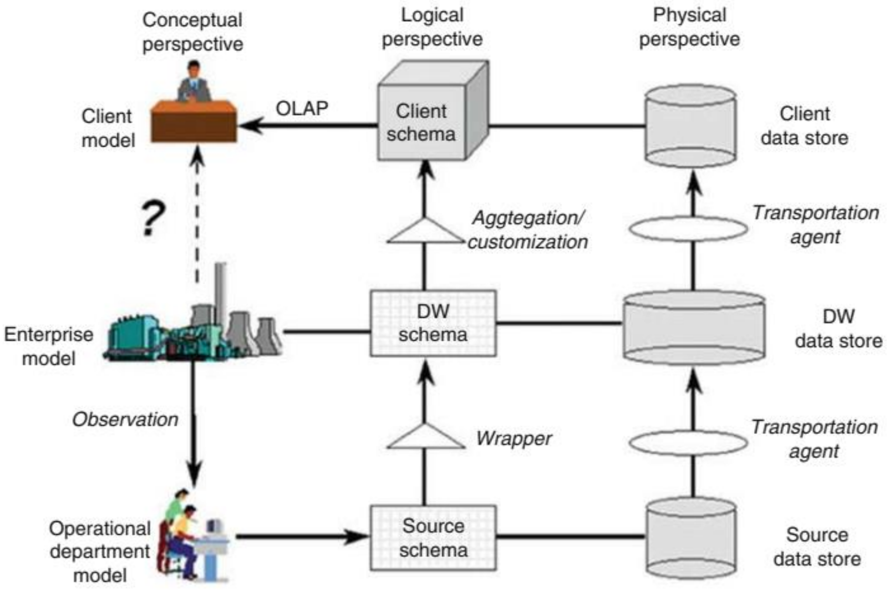
Xiao, J., Feng, D., Shi, Z., & Cheng, M. (2006). Flexible metadata management for block-level storage systems. *Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)*. https://doi.org/10.1109/snpd-sawn.2006.39

Zeng, M. L. & Qin, J. (2016). Metadata: Vol. 2nd edition. ALA Neal-Schuman.

Appendix

Figure 1

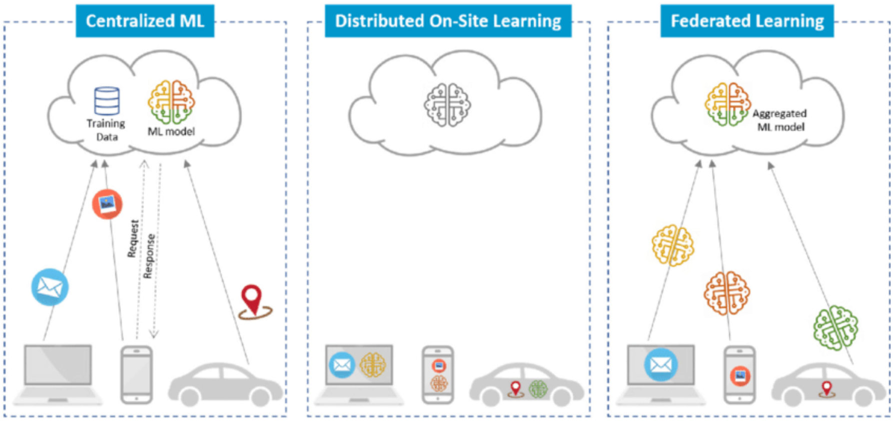
DWQ Model



Note: Figure from *Common Warehouse Metamodel (CWM) Specification, version 1.1. OMG* as cited by Vassiliadis (2018).

Figure 2

MMS framework



Note: Figure from *A Survey on Federated Learning: The Journey from Centralized to Distributed On-Site Learning and Beyond* by AbdulRahman et al. (2022).