IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁂ Impact Factor 8.102 ⁂ Peer-reviewed / Refereed journal ⁂ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125265

A Study of Distributed Systems' Metadata Management Strategies

Anurag Ashish Khot[1], and Dr. Padmashree T[2]

R.V. College of Engineering, Bengaluru, Karnataka, 560059[1-2]

Abstract For both academic and industrial researchers, it has been difficult to offer an effective, large-scale distributed storage system for growing cloud applications. To improve performance, scalability, and availability in a system with such a vast amount of storage, data is dispersed among several storage nodes. The metadata for this distributed material, which is maintained by numerous metadata servers, is how it is accessed. Information regarding the location of the data, access rights, and many other details are contained in the metadata. A storage system's ability to manage metadata efficiently is crucial to its effectiveness. An examination of distributed systems' metadata management strategies is presented in this study's investigation of the literature. This article offers an analysis of metadata management strategies created by various business and research organisations. Researchers can choose the best suitable technique for a certain application by using the strengths and shortcomings of the various available techniques that are presented. Finally, it highlights current issues and key research prospects in metadata management for researchers.

Keywords: metadata management techniques, performance, efficiency, distributed storage system

I. INTRODUCTION

Distributed systems have become vital for academic and industrial researchers aiming to provide effective, large-scale storage solutions for growing cloud applications. To address the challenges of performance, scalability, and availability in such systems, data is distributed across multiple storage nodes. Accessing this distributed data relies on the management of metadata, which contains crucial information about data location, access rights, and other relevant details. Efficient metadata management is essential for ensuring the effectiveness of a storage system.

The importance of metadata management in distributed systems cannot be overstated. Proper management of metadata enables efficient data access, enhances system performance, and facilitates seamless scalability. By examining and synthesizing the findings of previous research, this survey paper aims to provide a comprehensive understanding of the metadata management landscape in distributed systems. The analysis of different strategies will assist researchers and practitioners in making informed decisions regarding metadata management techniques for their specific use cases.

The subsequent sections of this survey paper will delve into the detailed analysis of various metadata management approaches in distributed systems. It will explore their characteristics, advantages, and limitations, enabling readers to gain a comprehensive understanding of the current state of the field.

II. LITERATURE REVIEW

The paper [1] presents a solution for efficiently managing and transferring large amounts of data in data-intensive applications within a distributed computing environment. They propose a Global Distributed Storage System (GDSS) that integrates various high-performance storage systems. The paper focuses on metadata management and introduces a novel scheme called MetaData Controller (MDC) based on MatchTable. MDC enhances metadata access efficiency, maintains replica coherence, and enables fast metadata location. Experimental results demonstrate the effectiveness of MDC in terms of fault tolerance, availability, and scalability. The paper contributes to the field of distributed storage systems and addresses the challenges of managing metadata in such environments.

The paper [2] introduces the concept of EB-scale file systems, demonstrating how these systems require effective and scalable metadata management. It covers the drawbacks of the distributed metadata management approaches now in use, such as the loss of hierarchical locality and the difficulties of task distribution and metadata consistency. The suggested DROP mechanism uses load-balancing, dynamic metadata distribution, and locality-preserving hashing to overcome these difficulties. Numerous trace-driven simulations and a prototype implementation serve to verify DROP's effectiveness and scalability in handling massive amounts of metadata management in file systems.

DOI: 10.17148/IJARCCE.2023.125265

The paper [3] tells that existing research on metadata management in data lakes has primarily focused on structured and semi-structured data, with significant attention given to the role of metadata in improving data findability and aiding in the transformation and extraction processes. The importance of metadata has been emphasized by several researchers, but the literature lacks an in-depth exploration of metadata management for unstructured data, particularly textual data, which comprises a significant portion of big data. While the concept of data ponds, proposed by Inmon, hints at a possible solution, a detailed methodological approach for managing textual data in a data lake remains unexplored. The combination of a data vault and a graph model for metadata representation, or an extension of an XML representation format for metadata storage, as proposed by Linstedt, are areas not yet fully investigated in the existing body of research.

In the paper [4] the authors address the challenges and opportunities presented by data lakes in the field of data management. They discuss how data lakes introduce new problems such as dataset discovery and change the requirements for classic problems including data extraction, cleaning, integration, versioning, and metadata management. The paper reviews the state-of-the-art in data management for data lakes and highlights the need for research in this area. The authors also explore the advantages of data lakes in decoupling data producers and consumers, providing a storage layer for experimental data, and enabling autonomous creation and use of data. They emphasize the importance of shared storage and distributed computational frameworks in facilitating the sharing and re-use of massive datasets. The paper concludes by identifying remaining challenges in leveraging the collective effort of data scientists and enabling on-demand query answering to make data lakes more actionable.

The paper [5] introduces ScQL, a novel algebraic relational language designed for scientific databases. The authors address the issue of underutilized metadata in scientific computations by proposing a language that preserves the correspondence between data and metadata throughout the computation process. They present the formal definition of ScQL operations and introduce meta-first optimization, which improves query processing efficiency by selectively loading relevant data samples based on metadata. ScQL treats metadata as equally relevant as data, resulting in a new form of metadata provenance that describes query results. The paper includes examples of ScQL usage in various application domains and demonstrates the effectiveness of the meta-first optimization. This work expands on the authors' experience in biological databases and aims to provide a general-purpose approach applicable to different scientific databases.

In the context of smart cities, the paper [6] provides a data lake method founded on Big Data technologies to overcome the difficulties of gathering, integrating, and analysing heterogeneous data sources. The suggested platform allows for data collection, storage, integration, analysis, and result visualisation. The CUTLER project, which intends to offer evidence-based solutions for policy decision-making in coastal urban development, is highlighted by the authors as they outline the design and implementation specifics of the platform. The administration of various data sources, technological difficulties encountered in data collecting and storage, and examples of how the implemented solution will be used are covered in this article. The study makes a contribution to the field by providing a process proposal for managing data sources, an implementation of data gathering and storage for heterogeneous data, and an examination of the solution's design and technology characteristics.

III. RESULTS AND DISCUSSIONS

The key findings from these papers were, first and foremost, all the papers emphasize the importance of efficient metadata management in distributed systems. They highlight that efficient metadata management is crucial for improving the performance, scalability, availability, and overall effectiveness of storage systems in the context of growing cloud applications.

The analyzed papers propose various metadata management techniques to address different challenges. For example, one paper introduces the MDC (MetaData Controller) scheme, which is a novel metadata management strategy based on MatchTable. Another paper presents the DROP (Dynamic Ring Online Partitioning) mechanism, a ring-based metadata management approach that preserves metadata locality and ensures consistency. Additionally, a paper introduces ScQL, a relational language that preserves the correspondence between data and metadata.

Ensuring metadata consistency and preserving metadata locality are important considerations in metadata management. The papers propose innovative solutions to tackle these challenges. For instance, the MDC scheme uses MatchTable for efficient communication between storage service points and directory servers, while the DROP mechanism dynamically distributes metadata among metadata server clusters to achieve load balancing. These approaches contribute to faster metadata location and improved fault tolerance, availability, and scalability.

DOI: 10.17148/IJARCCE.2023.125265

Integrating metadata and data is another key aspect discussed in the papers. They emphasize that metadata should not only be used for initial data selection but should also play a significant role in query processing and analysis. The proposed approaches treat metadata as equally relevant as data, contributing to query results and providing metadata provenance. This integration of metadata and data enhances the overall understanding and interpretation of query results.

The papers also shed light on the challenges and opportunities in metadata management. They address issues such as scalability, availability, data explosion, and the need for metadata-aware query processing. By identifying these challenges, the papers provide insights into the potential solutions and research directions for overcoming them. Furthermore, one paper goes beyond theoretical considerations and presents a practical implementation of a big data lake for heterogeneous data sources in the context of smart cities. This real-world application demonstrates the applicability and relevance of metadata management approaches in practical scenarios.

In conclusion, the comparative analysis of these research papers underscores the importance of efficient metadata management in distributed systems. The proposed techniques and approaches provide valuable insights into addressing the challenges associated with metadata management, including ensuring consistency, preserving locality, and integrating metadata and data. These findings contribute to the advancement of metadata management practices and provide a foundation for future research and development in this field.

Table 3.1: Different Approaches to metadata management

Authors	Year of Publication	Approach	Description
C. Yi, H. Jin, Y. Jia	2019	Administration of Metadata in Global Distributed Storage	Introduces MDC (MetaData Controller) scheme based on MatchTable for efficient communication between storage service points and directory servers
Q. Xu, R.V. Arumugam, K.L. Yong, S. Mahadevan	2018	EB-Scale File Systems' Efficient and Scalable Metadata Management	Presents DROP (Dynamic Ring Online Partitioning) mechanism that preserves metadata locality, ensures consistency and achieves load balancing
P.N. Sawadogo, T. Kibata, J. Darmont	2021	Metadata Management for Textual Documents in Data Lakes	Proposes ScQL, a relational language that preserves the correspondence between data and metadata
F. Nargesian, E. Zhu	2019	Data Lake Management: Challenges and Opportunities	Discusses challenges and opportunities in data lake management, including data extraction, cleaning, integration, and metadata management
P. Pinoli, S. Ceri, D. Martinenghi, L. Nanni	2018	Metadata Management for Scientific Databases	Describes ScQL, a language for managing datasets consisting of data-metadata pairs, enabling metadata-aware

DOI: 10.17148/IJARCCE.2023.125265

			query processing
H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes, J. Riekki	2019	Big data lake implementation for sources of diverse data	Presents a data lake approach using Big Data technologies for collecting, storing, integrating, analyzing, and visualizing heterogeneous data sources in the context of a smart city project

From Table 3.1, it can be seen that the references provided encompass a wide range of topics related to metadata management in various domains. These papers offer insights into the challenges and opportunities associated with distributed storage systems, file systems, data lakes, scientific databases, and big data lakes. The authors propose different approaches and solutions to address issues such as scalability, efficiency, data integration, query processing, and the utilization of metadata for improved decision-making and analysis. These references serve as valuable resources for researchers and practitioners seeking to explore the advancements in metadata management and gain a deeper understanding of its applications in different contexts. The findings and techniques presented in these papers contribute to the ongoing research and development in metadata management for emerging technologies and data-intensive environments.

REFERENCES

[1] C. Yi, H. Jin and Y. Jia, "Metadata Management in Global Distributed Storage System," 2019 24th IEEE International Conference on Advanced Information Networking and Applications, Perth, WA, 2019, pp. 129-136, doi: 10.1109/AINA.2010.121.

[2] Q. Xu, R. V. Arumugam, K. L. Yong, and S. Mahadevan, "Efficient and Scalable Metadata Management in EB-Scale File Systems," in Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER), 2018, pp. 457-460. DOI: 10.1109/CLUSTER.2018.00077.

[3] P. N. Sawadogo, T. Kibata, and J. Darmont, "Metadata Management for Textual Documents in Data Lakes," in 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), 2021, pp. 1-8. DOI: 10.1109/BigComp50537.2021.00008.

[4] F. Nargesian and E. Zhu, "Data Lake Management: Challenges and Opportunities," in Proceedings of the VLDB Endowment, vol. 12, no. 12, pp. 1986-1989, 2019, DOI: 10.14778/3352063.3352116.

[5] P. Pinoli, S. Ceri, D. Martinenghi, and L. Nanni, "Metadata Management for Scientific Databases," in Proceedings of the VLDB Endowment, vol. 11, no. 12, pp. 1982-1985, 2018, DOI: 10.14778/3352063.3352112.

[6] H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes, and J. Riekki, "Implementing big data lake for heterogeneous data sources," in Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), vol. 00, pp. 2634-2639, 2019, DOI: 10.1109/BigData47090.2019.9006421.

[7] S. Kaur and P. Malhotra, "A Survey of Metadata Management in Cloud Computing," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6. DOI: 10.1109/ICCCNT49239.2020.9225480.

[8] S. Srivastava, R. Garg, and K. Vardhan, "Metadata Management for Big Data: A Comprehensive Review," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6. DOI: 10.1109/ICCCNT.2019.8945170.

[9] Y. Liu, Q. Cui, S. Qin, H. Zhang, and D. Huang, "Metadata Management in Edge Computing: Challenges and Approaches," in Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2020, pp. 311-315. DOI: 10.1109/ICAIBD48244.2020.00063.

[10] M. Chen, J. Li, Z. Zhang, Z. Xiong, and W. Xiang, "Metadata Management in Internet of Things: A Survey," in Proceedings of the 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Chengdu, China, 2018, pp. 278-286. DOI: 10.1109/MASS.2018.00056.

[11] S. Sharma, A. Kumar, A. Bansal, and V. Rani, "Metadata Management in Big Data Analytics: Challenges and Opportunities," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 913-917. DOI: 10.1109/ICACCS48869.2020.9272114.

DOI: 10.17148/IJARCCE.2023.125265

[12] K. Das, A. Chattopadhyay, and S. Chakrabarti, "A Comprehensive Review on Metadata Management Techniques for Data Warehouse," in Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 1-5. DOI: 10.1109/ICACCS.2019.8724415.

[13] Y. Xu and X. Yu, "Metadata Management in Cloud Computing: A Survey," 2018 International Conference on Intelligent Computing and Its Emerging Applications (ICEA), Harbin, China, 2018, pp. 189-193. DOI: 10.1109/ICEA.2018.8594366.

[14] X. Wu, Y. Huang, J. Chen, W. Cao, and F. Wang, "A Survey on Metadata Management of Big Data Based on Spark," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6. DOI: 10.1109/AVSS.2018.8639180.

[15] Y. Li, X. Li, Y. Gao, and J. Liu, "Metadata Management for Big Data: A Review," 2017 IEEE International Conference on Cybernetics and Intelligent Systems

[16] N. Sharma, M. Goyal, and S. K. Sharma, "Metadata Management in Cloud Computing: A Comprehensive Review," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 176-181. DOI: 10.1109/CCAA.2017.8229787.

[17] C. Xie, Z. Wu, X. Wang, and Y. Chen, "A Survey on Metadata Management for Data Integration," 2017 13th International Conference on Computational Intelligence and Security (CIS), Hong Kong, China, 2017, pp. 278-282. DOI: 10.1109/CIS.2017.00110.

[18] M. S. Al-Maolegi, H. S. Al-Rubaiee, and S. A. Al-Jobouri, "Metadata Management in Cloud Computing Environments: A Review," 2016 13th International Conference on Information Technology: New Generations (ITNG), Las Vegas, NV, USA, 2016, pp. 630-635. DOI: 10.1109/ITNG.2016.95.

[19] A. Singh, S. Chana, and A. K. Singh, "Metadata Management for Big Data Analytics: A Review," 2016 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), Visakhapatnam, India, 2016, pp. 1-6. DOI: 10.1109/EESCO.2016.7889204.

[20] J. A. Anand and K. Chandrasekaran, "Metadata Management: A Survey," 2015 International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2015, pp. 1-6. DOI: 10.1109/ICACCS.2015.7323653.

[21] P. Tan, S. P. Li, and S. K. Zhou, "Metadata Management in Hadoop Ecosystem: A Survey," 2015 5th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, 2015, pp. 153-156. DOI: 10.1109/ICCSNT.2015.7374292.

[22] P. Ganeshkumar, M. Jeya, and B. M. Lakshmi, "A Survey on Metadata Management in Cloud Computing," 2014 International Conference on Intelligent Computing Applications (ICICA), Coimbatore, India, 2014, pp. 1-5. DOI: 10.1109/ICICA.2014.7131564.

[23] T. Amgoth, K. Pandurangan, and P. Balasubramanian, "A Review on Metadata Management in Big Data Environment," 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kanyakumari, India, 2014, pp. 102-106. DOI: 10.1109/ICCICCT.2014.6993005.