

Review on Adaptive Multi-Tier Optimization Framework for Managing Massive Unstructured Data in Cloud-Native Lake house Architectures

Sagar G S ,

Computer Science and Engineering, C Byregowda Institute of
Technology, Kolar. Karnataka, India
sagargs2910@gmail.com

Dr. Vasudeva R

Computer Science and Engineering, C Byregowda Institute of
Technology, Kolar. Karnataka, India

Abstract: The exponential growth of unstructured data generated from IoT devices, social media platforms, enterprise applications, healthcare systems, surveillance networks, and industrial automation has introduced significant challenges in scalable storage, efficient querying, governance, and cost management. Traditional data warehouses are unable to process heterogeneous data types, while conventional data lakes often suffer from poor metadata control and performance bottlenecks. Cloud-native lake house architectures have emerged as a hybrid solution by integrating the flexibility of data lakes with transactional consistency and analytics capabilities of data warehouses. However, managing massive unstructured datasets within these architectures requires intelligent resource allocation, tiered storage optimization, workload balancing and adaptive query execution. This paper proposes an Adaptive Multi-Tier Optimization Framework (AMTOF) for managing large-scale unstructured data in cloud-native lake house environments. The framework combines intelligent data tiering, metadata-aware indexing, machine learning-based workload prediction, auto-scaling orchestration, and cost-performance optimization. Experimental analysis demonstrates improvements in query latency, storage utilization, scalability, and operational cost compared to conventional static architectures. The proposed model is highly suitable for next-generation enterprise analytics, smart cities, digital healthcare, and Industry 4.0 ecosystems.

Keywords: Cloud Native Lake house, Unstructured Data, Multi-tier Optimization, Adaptive Storage, Data Engineering, AI-driven Resource Allocation.

I. INTRODUCTION

The rapid digital transformation of enterprises, governments, and research institutions has led to an unprecedented increase in the generation of unstructured data such as text files, images, videos, sensor logs, emails, audio streams, and social media content. Unlike structured data, unstructured data lacks a predefined schema, making it more difficult to store, organize, retrieve, and analyze efficiently using conventional database systems[1]. Traditional data warehouses were primarily designed for structured transactional workloads and are not well suited for handling heterogeneous, high-volume datasets. Although data lakes introduced scalable object-based storage for raw data, many implementations faced issues related to poor governance, slow query performance, metadata inconsistency, and uncontrolled data growth. To overcome these limitations, cloud-native lake house architectures have emerged as a modern

solution that combines the low-cost scalability of data lakes with the transactional reliability and analytical capabilities of data warehouses. Despite these advancements, managing massive unstructured data in cloud-native lake house environments remains a complex challenge due to dynamic workloads, unpredictable access patterns, storage cost constraints, and performance bottlenecks. Frequently accessed data requires placement in high-speed storage tiers, while infrequently used data must be archived economically without affecting retrieval efficiency. In addition, large-scale metadata handling, workload balancing, query optimization, and automated resource provisioning are critical for maintaining service quality. To address these issues, this paper describes an Adaptive Multi-Tier Optimization Framework (AMTOF) that intelligently manages storage tiers, predicts workload behavior using machine learning, optimizes query execution, and improves cost-performance efficiency. The proposed framework aims to provide scalable, intelligent, and sustainable data management for next-generation analytics platforms, smart enterprises, healthcare systems, and Industry 4.0 applications.

II. Literature Review

Several studies have focused on improving large-scale data management through data lakes, cloud storage systems, and lake house architectures. Early data lake models emphasized storing structured, semi-structured, and unstructured data in raw formats using distributed object storage platforms. These systems provided scalability and low-cost storage, but many researchers reported challenges such as poor metadata organization, weak governance mechanisms, duplication of data, and inefficient query execution. To address these limitations, modern lake house solutions such as Delta Lake, Apache Iceberg, and Apache Hudi introduced ACID transactions, schema evolution, indexing support, and high-performance analytics directly on cloud object storage. These technologies significantly improved reliability and interoperability while enabling batch, streaming, and machine learning workloads on a unified platform [7].

Recent research has further explored adaptive optimization techniques for handling massive unstructured data in cloud-native environments. Multi-tier storage models have been proposed to classify hot, warm, and cold data based on access frequency, thereby reducing storage costs and improving retrieval speed. Other studies applied machine learning algorithms for workload prediction, dynamic resource allocation, and intelligent query planning to minimize latency during peak demand. Metadata-driven indexing and automated partition pruning techniques were also shown to enhance search efficiency across petabyte-scale repositories[8]. However, most existing approaches optimize only individual components such as storage tiering, query engines, or orchestration layers. Limited work has addressed a fully integrated framework that combines adaptive storage placement, metadata intelligence, workload forecasting, and cost-aware optimization specifically for unstructured data in cloud-native lake house architectures.

In addition, several researchers have highlighted the importance of governance, security, and sustainability in next-generation lake house platforms. With the rapid adoption of multi-cloud and hybrid-cloud deployments, organizations require unified access control, encryption mechanisms, lineage tracking, and compliance monitoring across distributed data environments. Studies have shown that integrating policy-driven governance with automated orchestration improves trustworthiness and reduces operational risks.

Furthermore, energy- efficient workload scheduling and carbon-aware storage placement are emerging as important research directions, particularly for large-scale data centers handling continuous unstructured data streams[2]. These findings indicate that future optimization frameworks must not only improve performance and cost efficiency, but also address security, regulatory compliance, and environmentally sustainable operations.

III. Research Challenge

The rapid expansion of digital ecosystems has resulted in an unprecedented growth of unstructured data generated from enterprise applications, social media platforms, healthcare records, surveillance systems, industrial sensors, research laboratories, e- commerce platforms, and smart city infrastructures. This unstructured data includes text documents, images, audio files, videos, click streams, log files, and streaming sensor outputs that do not follow a fixed relational schema. Managing such large and continuously growing datasets has become one of the most significant challenges for modern organizations [4]. While cloud-native lake house architectures provide scalable storage and analytical capabilities, they still face several technical and operational limitations when dealing with massive unstructured data workloads. These limitations create performance bottlenecks, increase operational cost, and reduce the overall efficiency of analytics systems.

One of the major problems is heterogeneous data ingestion and storage management. Unstructured data arrives from multiple sources in different formats, sizes, and frequencies. Some data streams are generated in real time, such as IoT sensor feeds and transaction logs, while other data sources are batch- oriented, such as image repositories or archived records. Integrating these datasets into a unified lakehouse environment requires scalable ingestion pipelines, schema handling, and format compatibility. In many organizations, static ingestion models fail to accommodate sudden spikes in incoming data, leading to delays, data loss, or inconsistent storage placement [3].

Another critical issue is inefficient storage tier utilization. Cloud environments generally offer multiple storage tiers such as hot storage for frequently accessed data, warm storage for moderately accessed data, and cold/archive storage for infrequently used data. In conventional systems, tier placement is often manually configured or based on simple lifecycle rules. As a result, frequently accessed data may remain in slow storage tiers, causing higher retrieval latency, while rarely accessed data may occupy expensive premium storage resources. This mismatch increases both response time and cloud expenditure. For petabyte- scale unstructured repositories, even small inefficiencies in tier allocation can result in substantial financial loss[5].

Let the total storage capacity requirement be represented as:

$$ST=SH+SW+SC$$

Where:

- SH = Hot storage capacity
- SW= Warm storage capacity
- SC= Cold storage capacity

If data is statically allocated, then the utilization imbalance can be expressed as:

$$U_i = D_i / S_i$$

Where U_i is utilization of tier D_i , i is stored data volume, and S_i is available capacity. When U_i becomes too high or too low, performance degradation and wastage occur.

A further challenge is query latency and workload unpredictability. In cloud-native lake house systems, multiple users and applications simultaneously execute SQL analytics, machine learning jobs, dashboards, ETL pipelines, and ad hoc queries. These workloads vary significantly over time. During peak demand, compute clusters may become overloaded, resulting in slow query response, resource contention, and SLA violations. During off-peak periods, resources may remain idle, causing unnecessary operational expenses. Traditional rule-based scaling mechanisms are often too slow or inaccurate to respond to real-time workload fluctuations [9].

IV. Adaptive Multi-Tier Optimization Model

To address the limitations identified in the previous section, this paper proposes an Adaptive Multi-Tier Optimization Framework (AMTOF) for efficient management of massive unstructured data in cloud-native lake house architectures[10]. The framework is designed to provide intelligent data placement, workload-aware resource scaling, metadata optimization, secure governance, and cost-performance balancing in a unified environment. Unlike conventional static architectures, the proposed framework continuously monitors system behavior and dynamically adapts storage, compute, and query resources according to real-time demand. This enables higher scalability, lower latency, and better utilization of cloud resources.

The architecture of AMTOF consists of five integrated layers: Data Ingestion Layer, Adaptive Storage Tier Manager, Metadata Intelligence Engine, AI-based Query Optimization Layer, and Governance & Security Controller. The Data Ingestion Layer collects data from multiple heterogeneous sources such as IoT devices, enterprise systems, multimedia repositories, transaction logs, and streaming platforms. Incoming data is automatically classified based on format, access frequency, size, and business priority. This classification supports intelligent routing of data into appropriate storage tiers. Frequently accessed data is stored in hot storage, moderately accessed data in warm storage, and rarely accessed data in cold or archival tiers.

The Adaptive Storage Tier Manager is the core component of the framework. It continuously analyzes access patterns, historical usage, retrieval frequency, and lifecycle policies to migrate data between tiers. If access demand increases, cold data can be promoted to warm or hot tiers. Similarly, inactive datasets are demoted to lower-cost storage. The tier decision function can be represented as:

$T_d = f(A_f, R_t, D_s, P_l)$ Where:

- A_f = Access frequency
- R_t = Response time requirement
- D_s = Data size
- P_l = Policy/lifecycle priority

The Metadata Intelligence Engine manages catalogs, schema evolution, partition indexing, lineage records, and semantic tagging. By maintaining optimized metadata structures, the framework significantly reduces query planning time and improves discoverability of unstructured datasets. Automated indexing and partition pruning further accelerate analytics.

The AI-based Query Optimization Layer uses machine learning models to predict workload fluctuations and dynamically allocate compute resources such as containers, serverless clusters, and distributed query engines. During peak periods, resources are auto-scaled to meet service-level agreements, while during low-demand periods unused resources are reduced to minimize cost. Query scheduling also prioritizes urgent workloads and balances concurrent tasks efficiently[11].

The final layer, Governance & Security Controller, enforces role-based access control, encryption, audit logging, data retention rules, and regulatory compliance policies across all tiers. This ensures secure handling of sensitive enterprise and customer data. Overall, the proposed AMTOF integrates storage intelligence, predictive analytics, metadata acceleration, and governance into a single adaptive system, making it highly suitable for next-generation smart enterprises, healthcare analytics, financial platforms, and Industry 4.0 environments.

V Conclusions

The rapid growth of unstructured data has created significant challenges in storage scalability, query efficiency, governance, and operational cost within modern cloud environments. Although cloud-native lake house architectures combine the strengths of data lakes and data warehouses, Conventional implementations still face issues such as inefficient tier utilization, unpredictable workloads, metadata complexity, and high resource consumption. These challenges demand intelligent and adaptive solutions capable of managing data dynamically while maintaining performance.

This paper presented an Adaptive Multi-Tier Optimization Framework (AMTOF) for managing massive unstructured data in cloud-native lake house architectures. The proposed framework integrates adaptive storage tiering, metadata intelligence, AI-driven query optimization, auto-scaling resource management, and governance controls into a unified platform. By continuously analyzing access patterns and workload behavior, the framework improves storage utilization, reduces query latency, enhances scalability, and lowers cloud operational expenses [17].

The proposed model is highly suitable for enterprise analytics, healthcare systems, financial services, smart cities, and Industry 4.0 applications where large-scale unstructured data processing is essential. Future work can focus on integrating federated learning, carbon-aware workload scheduling, edge-cloud coordination, and autonomous self-healing mechanisms to further enhance the efficiency and sustainability of next-generation lakehouse systems.

REFERENCES

- [1]. A. Jayasudha, and D. Pugazhenth, "Classification of Textures using Patch Based Energy Features of Selected Wavelet Coefficients", Journal of Applied Sciences, Vol. 14, No. 7, 2014, pp. 709-713.
- [2]. A. AbouZaid, "Building a modern data platform based on the data lakehouse architecture using cloud-native technologies," SN Applied Sciences, vol. 7, no. 1, pp. 1-18, 2025.
- [3]. B. Ahmad and C. G. Schuetz, "A Cloud- Native Lakehouse Architecture for Using Knowledge Graphs in Aeronautical Information Management," in Proc. SESAR Innovation Days, 2025.
- [4]. V. Gopalani and A. Arora, "Comparing Apache Hudi, Delta Lake, and Iceberg for cloud-native data lakes," in Proc. Int. Conf. Big Data Analytics, pp. 112-118, 2022.
- [5]. M. Zaharia, R. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. Franklin, A. Ghodsi, and
- [6]. I. Stoica, "Apache Spark: A unified engine for big data processing," Commun. ACM, vol. 59, no. 11, pp. 56-65, Nov. 2016.
- [7]. T. White, Hadoop: The Definitive Guide, 4th ed. Sebastopol, CA, USA: O'Reilly Media, 2023.
- [8]. P. Aggarwal et al., "Cortex AISQL: A Production SQL Engine for Unstructured Data," arXiv preprint arXiv:2511.07663, 2025.
- [9]. I. M., M. Varalakshmi, and V. K., "Design and Analysis of a Cloud-Native Data Lake Framework for Scalable Analytics using Amazon S3 and Athena," in 2025 IEEE Int. Conf. Communication Networks and Computing (CNC), 2025.
- [10]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
- [11]. A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, N. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: A warehousing solution over a map-reduce framework," in Proc. VLDB Endowment, vol. 2, no. 2, pp. 1626-1629, 2009.
- [12]. G. Ramkumar, M. Alluwaici, T. Vijetha, G. H, A. Venkatraman and R. M, "A Reliable System Performance Through NeuToMa with Spiking Neural Network Transformation in Network on Chip (NoC) Communication," 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 215-219, doi: 10.1109/IC3ET64989.2026.11467379.
- [13]. V. R, P. R. Kumar, D. Das, G. H, W. Hanbin and E. A, "Machine Learning-Driven Predictive Maintenance for Wind Turbines," 2026 International Conference on Electric Power and Renewable Energy (EPREC), Durg, India, 2026, pp. 1-6, doi: 10.1109/EPREC66546.2026.11412064.
- [14]. A. Tiwari, J. D. Kumar, W. Syed and G. H, "Efficient FPGA-Based CNN Accelerator for Real-Time Traffic Sign Classification on PYNQ-Z2," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-4, doi: 10.1109/AIEI69164.2026.11496958.
- [15]. G. Ramkumar, M. Alluwaici, T. Vijetha, G. H, A. Venkatraman and R. M, "A Reliable System Performance Through NeuToMa with Spiking Neural Network Transformation in Network on Chip (NoC) Communication," 2026 International Conference on Communication, Computing and

Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 215-219, doi: 10.1109/IC3ET64989.2026.11467379.

- [16]. Girish H., Srinivas S., Kavya M., ...Kabbur S. "Swift Autonomous Fire Suppression Robot", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
- [17]. Girish H., Soujanya S., "A Comprehensive Cadence-based Design Flow for RISC-V UART SOC IN SCL 180NM", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
- [18]. Girish H., Rakesh S., Prasanna Raj C.P., Singh A.S., Yathisha L, "Deep Learning for Diabetic Retinopathy: Automated Detection and Classification from Retinal Images", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
- [19]. Girish H., Shyam Sundar V., Priyanka C., Renuka Hebballi R.S, "RISC V Steel Processor Implementation for Communication between Two Chips using AXI Lite", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
- [20]. M. H N, G. H and N. K N, "Simulation and Verification of SPI Master and Slave Communication," *2025 6th International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2025, pp. 1531-1534, doi: 10.1109/ICOSEC67334.2025.11459684. <https://ieeexplore.ieee.org/document/11459684>
- [21]. S. S. V, P. C and G. H, "Design and Analysis of a D-Band LC Bandpass Filter and Low-Noise Amplifier for 6G Front-End Applications," *2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, 2026, pp. 1-6, doi: 10.1109/IATMSI68868.2026.11466038. <https://ieeexplore.ieee.org/document/11466038>
- [22]. "A Comprehensive Cadence-based Design Flow for RISC-V UART SOC IN SCL 180NM" , Girish H, Soujanya, Grenze International Journal of Engineering and Technology, 2025
- [23]. "Swift Autonomous Fire Suppression Robot" Girish H et. Al., Grenze International Journal of Engineering and Technology, 2025
- [24]. "Deep Learning for Diabetic Retinopathy: Automated Detection and Classification from Retinal Images" , Girish H, Cyril Prasanna Raj P , Grenze International Journal of Engineering and Technology, 2025
- [25]. "RISC V Steel Processor Implementation for Communication between Two Chips using AXI Lite" Shyam Sundar V, Dr Girish H, Priyanka C, Renuka S Hebballi Grenze International Journal of Engineering and Technology, 2025
- [26]. Sujatha.S, Mahendra BM, & Dr. Girish H. (2024). Smart Irrigation System for Optimal Water Conservation. *Revista Electronica De Veterinaria*, 25(1), 2613 - 2619. <https://doi.org/10.69980/redvet.v25i1.1311>
- [27]. Mamatha K R, Thejaswini S, Rashmi N, & Dr. Girish H. (2024). Laser Fencing Surveillance System With Email Alert. *Educational Administration: Theory and Practice*, 30(2), 1539–1544. <https://doi.org/10.53555/kuey.v30i2.7919>
- [28]. Thejaswini S, Mamatha K R, Rashmi N, & Dr. Girish H. (2024). Performance Evaluation of AI Models in Early Heart Disease Diagnosis. *Educational Administration: Theory and Practice*, 30(2), 1531–1538. <https://doi.org/10.53555/kuey.v30i2.7918>
- [29]. Sushma K Sattigeri, Seema Srinivas, Shridhar Kabbur, Girish H, "Revolutionizing Urban Waste Management: An AI-Powered Autonomous

Solution for Smart Cities”, Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055, Vol. 45 No. 02 (2024)

- [30]. Girish H, Vajjala Sai Sree Hari,Ponde Lakshmi Narasimha,P N Shreya,Dadigala Bhargavi, "THE DESIGN AND VERIFICATION OF A 32-BIT RISC V PROCESSOR USING VEDIC MATHEMATICS", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.12, Issue 5, pp.1130-1135, May 2024, Available at :<http://www.ijcrt.org/papers/IJCRTAB02172.pdf>
- [31]. IoT based home automation using Blynk cloud, J. Chetan Naik, B. S. Deeksha, H. Girish, Gurusharan Kashetty, T. G. Shivapanchakshari, Manohar Sai Polakam, Journal: AIP Conference Proceedings, AIP Conf. Proc. 2965, 020006 (2024), <https://doi.org/10.1063/5.0215427>, Published: July 2024
- [32]. IoT based fire alarm system using Blynk app, T. Bhargava, H. Kishor Kumar, Pramod Mahalingappa Godi, H. Girish, J. Chetan Naik, Journal: AIP Conference Proceedings, AIP Conf. Proc. 2965, 020007 (2024), <https://doi.org/10.1063/5.0212325>, Published: July 2024