International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

Journal Website: http://sciencebring.com/index.php/ijasr

Research Article

Multi-Tenant Data Lake Architecture for Scalable AI and Big Data Workload Management

Submission Date: June 28, 2026, **Accepted Date:** July 22, 2026,
Published Date: August 20, 2026
Crossref doi: https://doi.org/10.37547/ijasr-06-08-10

Dr. Arjun Mehta
Department of Artificial Intelligence National Institute of Intelligent Computing New Delhi, India

Dr. Priya Sharma
Department of Machine Learning and Data Science Institute of Advanced Digital Technologies Bengaluru, India

Abstract

The rapid expansion of artificial intelligence (AI), machine learning, computer vision, and multimodal analytics has increased the demand for data infrastructures capable of supporting heterogeneous workloads at large scale. Conventional data platforms frequently encounter difficulties when multiple users, applications, or organizational units simultaneously access shared datasets, compute resources, and analytical services. This paper develops a research-oriented conceptual architecture for a multi-tenant data lake designed to support scalable AI and big data workload management. The proposed architecture integrates tenant-aware data ingestion, metadata management, storage isolation, workload orchestration, resource governance, security, and adaptive AI processing into a unified framework. The methodology is derived through comparative synthesis of the supplied literature, including research on multimodal datasets, computer vision workloads, computational sciences, and responsible approaches to AI. The architecture emphasizes logical tenant isolation while preserving controlled opportunities for data and infrastructure sharing. The analysis indicates that workload-aware orchestration, metadata-driven resource allocation, and differentiated service policies can improve scalability and reduce resource contention in heterogeneous environments. The paper further argues that multi-tenancy must be treated not merely as a virtualization problem but as a data-governance, workload-management, and responsible-AI problem. The resulting framework provides a foundation for scalable AI data lakes while identifying limitations related to resource interference, governance complexity, data heterogeneity, and fairness.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

ISSN-2750-1396

KEYWORDS

Multi-Tenant Data Lake, Big Data, Artificial Intelligence, Workload Management, Data Lake Architecture, Resource Orchestration, Data Governance, Scalable Analytics, AI Infrastructure

INTRODUCTION

The increasing adoption of artificial intelligence and big data analytics has transformed data infrastructure from a passive storage layer into an active computational foundation for intelligent applications. Modern AI workloads consume heterogeneous datasets containing structured, semi-structured, unstructured, image, textual, and multimodal information. For example, datasets supporting computer vision can involve millions of images and associated metadata, while language-oriented applications may require large collections of multilingual and multimodal resources. Afifi (2019) demonstrates the scale and complexity that can arise in image-based machine learning datasets, whereas Abdulmumin et al. (2022) illustrate the requirements of multimodal datasets for language-related AI applications. These examples indicate that a contemporary analytical platform must accommodate substantially different data structures and computational characteristics.

A data lake provides an appropriate conceptual foundation because it permits diverse data types to be retained within a common analytical environment. However, the conventional data-lake model becomes increasingly complex when the same infrastructure serves multiple independent tenants. Tenants may represent departments, research groups, business units, applications, or external users with different performance requirements, security constraints, and data-access privileges. Concurrent execution can generate resource contention, unpredictable latency, excessive storage consumption, and governance difficulties. Consequently, scalable data-lake design requires explicit mechanisms for tenant isolation and workload coordination.

Recent work on scalable data-lake architecture emphasizes the importance of multitenant orchestration for AI workloads, particularly when computational demand fluctuates across heterogeneous applications (Goyal, 2025). The central architectural problem is therefore not simply how to store large datasets, but how to coordinate storage, computation, metadata, and AI processing while maintaining predictable service quality among multiple tenants.

The objective of this paper is to develop and critically analyze a conceptual multi-tenant data-lake architecture for scalable AI and big data workload management. The study focuses on five interconnected objectives: establishing a tenant-aware data organization model, designing workload-aware orchestration mechanisms, integrating heterogeneous AI datasets, developing governance and isolation principles,

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

and identifying the principal trade-offs associated with multi-tenant scalability.

The significance of this research extends beyond infrastructure efficiency. AI systems increasingly operate on datasets with social, cultural, linguistic, and contextual dimensions. Abdilla et al. (2020) emphasize the importance of Indigenous protocols in artificial intelligence, demonstrating that data-intensive AI infrastructures cannot be evaluated exclusively through computational performance. Consequently, a scalable architecture must combine technical efficiency with appropriate governance and responsible data-management principles.

LITERATURE REVIEW

The supplied literature provides a heterogeneous but complementary foundation for understanding the requirements of scalable AI data infrastructure. Rather than directly describing a complete multi-tenant data lake, the references collectively reveal several workload characteristics that such an architecture must support.

Afifi (2019) presents a large-scale hand-image dataset for gender recognition and biometric identification. The work is particularly relevant to data-lake architecture because biometric image processing involves high-volume unstructured data, computationally intensive feature extraction, and potentially sensitive information. Such workloads require efficient storage and controlled access. A multi-tenant architecture must therefore distinguish between ordinary analytical datasets and data requiring stricter access policies.

Amraoui et al. (2022) introduce an agricultural UAV database designed for image segmentation and classification. This type of dataset illustrates another important property of AI workloads: data may originate from specialized sensing systems and require repeated high-performance processing. UAV imagery can generate large files, while segmentation workloads can impose substantial computational demand. Within a shared data lake, simultaneous execution of such workloads can create contention for storage bandwidth and compute resources.

Bashkirova et al. (2022) address deformable object segmentation in cluttered scenes through the ZeroWaste dataset. The study reinforces the need for data infrastructures capable of supporting sophisticated computer-vision pipelines rather than merely storing static files. Training and evaluation processes may repeatedly access the same datasets, suggesting that caching, locality-aware scheduling, and workload prioritization can become important components of an AI-oriented lake.

Abdulmumin et al. (2022) provide evidence of the complexity of multimodal data through the Hausa Visual Genome dataset. Multimodal systems require coordinated access to multiple information types and may involve different processing stages. Therefore, a data lake must maintain metadata capable of representing

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

ISSN-2750-1396

relationships between datasets rather than treating every object as an isolated file.

Ayachi et al. (2020) demonstrate the application of deep learning to real-world traffic-sign detection. The practical nature of such workloads illustrates the difference between experimental data processing and operational AI systems. A production-oriented architecture may require predictable response times, continuous data ingestion, and repeatable model-processing pipelines.

Birhane and Guest (2020) provide a critical perspective through their discussion of decolonising computational sciences. Their contribution is theoretically significant because it challenges purely technical interpretations of computational infrastructure. Data architecture determines not only how efficiently information can be processed but also how knowledge is organized, represented, and accessed. Consequently, governance should be considered an architectural property rather than an administrative afterthought.

Abdilla et al. (2020) similarly demonstrate that AI systems require contextual and culturally appropriate protocols. Their work implies that multi-tenant systems should provide policy mechanisms capable of encoding data-use restrictions, ownership expectations, and contextual governance requirements.

Collectively, these studies expose a research gap. Existing references demonstrate heterogeneous datasets, complex AI workloads, and the importance of responsible data practices, but they do not provide a unified framework for coordinating these requirements in a multi-tenant data-lake environment. Goyal (2025) directly motivates this architectural direction by framing scalable data lakes around AI workloads and multitenant big-data orchestration. The proposed research therefore positions multi-tenancy as the intersection of storage scalability, computational scheduling, metadata management, and responsible governance.

METHODOLOGY

3.1 Research Design

This study adopts a conceptual architecture and analytical synthesis methodology. The approach does not claim experimental benchmarking of a deployed data lake. Instead, the architecture is derived from the workload characteristics identified in the supplied references and organized into functional layers. The methodology consists of requirements identification, architectural decomposition, workload modeling, tenant isolation design, orchestration analysis, and critical evaluation.

The primary assumption is that a multi-tenant data lake must simultaneously support heterogeneous data and heterogeneous computational demands. Accordingly, architecture is evaluated according to scalability, isolation, workload adaptability, governance, and operational efficiency.

3.2 Architectural Model

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

ISSN-2750-1396

The proposed architecture consists of six logical layers: data ingestion, unified storage, metadata and catalog services, workload orchestration, AI processing, and governance.

The data-ingestion layer receives information from batch files, streaming systems, sensor platforms, databases, and external datasets. Each incoming object is associated with tenant identity, dataset classification, ownership metadata, sensitivity level, and processing requirements. This approach is particularly relevant for UAV imagery and computer-vision datasets, where data volume and acquisition patterns can vary substantially (Amraoui et al., 2022).

The unified storage layer provides scalable persistence for structured and unstructured information. Instead of enforcing a single data representation, the lake maintains raw, processed, and curated zones. Raw data preserve original information, processed data support computational pipelines, and curated data provide standardized assets for downstream analytical use. This separation improves reproducibility while reducing the need for every tenant to independently duplicate large datasets.

The metadata and catalog layer functions as the control plane for data discovery and governance. Each dataset should contain information describing its tenant, lineage, format, quality, sensitivity, permitted operations, and associated AI tasks. Multimodal datasets such as those discussed by Abdulmumin et al. (2022) particularly benefit from relationship-aware metadata because textual and visual resources must be connected without losing their individual properties.

The workload-orchestration layer is responsible for allocating compute resources among tenants. Instead of using static allocation, the architecture adopts workload-aware scheduling. A tenant submitting an intensive segmentation job, for example, may require more GPU resources than a tenant executing a lightweight metadata query. The orchestrator therefore evaluates workload type, priority, resource requirement, service-level objectives, and current cluster utilization. This principle aligns with the multitenant orchestration perspective associated with scalable AI data lakes (Goyal, 2025).

The AI-processing layer provides shared computational services for training, inference, feature extraction, computer vision, natural-language processing, and multimodal analysis. Shared model infrastructure can reduce duplication, while tenant-specific execution environments preserve logical separation. A traffic-sign detection workload, for instance, could use shared preprocessing services while retaining tenant-specific model configurations and output repositories (Ayachi et al., 2020).

Finally, the governance layer enforces access control, policy management, auditing, data lineage, retention rules, and responsible-use requirements. This layer is essential because sensitive datasets such as biometric information cannot be treated identically to ordinary public datasets (Afifi, 2019). Furthermore, governance must accommodate contextual requirements

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

identified in research concerning Indigenous AI protocols (Abdilla et al., 2020).

3.3 Tenant Isolation Model

Tenant isolation is implemented at three levels: data, computation, and policy.

At the data level, each object is logically associated with a tenant identity and access policy. Shared datasets may be made available through controlled permissions without requiring physical duplication. At the computational level, workloads are assigned quotas and resource boundaries. CPU, memory, storage bandwidth, and accelerator resources can therefore be controlled independently.

At the policy level, access decisions are determined by identity, dataset sensitivity, purpose, and organizational rules. This three-dimensional model is preferable to simple namespace isolation because a tenant may legitimately access a shared dataset while still requiring isolated compute and output spaces.

3.4 Workload Management Framework

The proposed orchestration model classifies workloads into four broad categories: batch analytics, AI training, real-time inference, and interactive exploration.

Batch workloads prioritize throughput. AI training prioritizes sustained access to computational accelerators and high-volume datasets. Real-time inference emphasizes latency and predictable response. Interactive analytics requires responsiveness but generally consumes shorter computational bursts.

A scheduler can assign a workload score based on resource demand, priority, deadline, and current utilization. Conceptually, resource allocation can be represented as:

[

R_{i}=f(P_i,D_i,S_i,U)

]

where (R_i) represents resources allocated to tenant (i), (P_i) represents workload priority, (D_i) represents resource demand, (S_i) represents service requirements, and (U) represents current infrastructure utilization.

The model should not permit priority to become an unrestricted mechanism for resource monopolization. Fairness constraints are therefore necessary. A high-priority AI-training workload may temporarily receive additional resources, but sustained monopolization should trigger quota enforcement or workload redistribution.

This principle is central to multitenant AI orchestration because scalability depends not only on the total computational capacity but also on how effectively that capacity is shared across competing workloads (Goyal, 2025).

3.5 Data and AI Governance

Governance is incorporated directly into the architecture rather than being treated as a

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

separate organizational process. Dataset policies should specify who can access information, what operations are permitted, how outputs may be reused, and whether information can be shared across tenants.

The requirement is particularly important for datasets involving human subjects, cultural information, or sensitive attributes. Afifi (2019) demonstrates the potential sensitivity of biometric-related data, while Abdilla et al. (2020) indicate that culturally grounded protocols can impose requirements beyond conventional technical access control.

The architecture therefore supports policy-aware data discovery. A dataset may appear in the catalog without necessarily being available for unrestricted processing. This distinction enables transparency while preserving governance.

3.6 Hypothetical Operational Scenario

Consider a research organization with three tenants. Tenant A processes UAV images for agricultural segmentation, Tenant B trains computer-vision models using waste-related imagery, and Tenant C operates multilingual multimodal analysis.

The ingestion layer stores the datasets independently while maintaining shared metadata. When Tenant A submits a large segmentation job, the scheduler identifies it as accelerator-intensive. Tenant B may simultaneously execute a training workload requiring substantial GPU capacity, while Tenant C performs comparatively smaller multimodal processing. Instead of allocating fixed infrastructure to each tenant, the scheduler dynamically distributes available resources according to workload requirements and policy constraints.

This approach improves utilization because resources can be temporarily reassigned when a tenant is idle. At the same time, quotas and isolation mechanisms prevent one tenant from consuming the entire infrastructure. Such a scenario illustrates how heterogeneous datasets identified across the supplied literature can translate into concrete architectural requirements.

RESULTS

The conceptual analysis produces five principal findings.

First, multi-tenancy requires workload-aware orchestration rather than simple storage partitioning. The reviewed datasets demonstrate significant variation in computational behavior. Image classification, object segmentation, multimodal processing, and real-time detection cannot be efficiently managed through identical resource policies. The architecture therefore treats workload characteristics as first-class scheduling inputs.

Second, metadata becomes an operational control mechanism. In a heterogeneous data lake, metadata is not limited to dataset descriptions. Tenant identity, sensitivity, lineage, processing requirements, and governance policies can

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

ISSN-2750-1396

influence storage placement, access decisions, and workload scheduling. This is particularly important for multimodal resources (Abdulmumin et al., 2022).

Third, logical isolation provides a balance between security and efficiency. Physically duplicating every dataset for every tenant would increase storage requirements and reduce opportunities for controlled sharing. Logical isolation allows shared infrastructure while maintaining tenant-specific permissions, compute quotas, and outputs.

Fourth, AI scalability depends on coordinated data and compute management. Large datasets alone do not guarantee scalable AI execution. Repeated training, segmentation, and inference operations can generate substantial compute and data-transfer demand. A multitenant architecture must therefore coordinate storage locality, caching, accelerator allocation, and workload priority. This reinforces the importance of orchestration identified in Goyal (2025).

Fifth, governance is an architectural requirement. The reviewed literature indicates that data may carry biometric, cultural, linguistic, or contextual sensitivities. Consequently, scalability without governance can increase risk rather than merely increase efficiency. Responsible access policies must operate alongside technical resource controls.

The findings also reveal an inherent trade-off. Aggressive resource sharing can improve utilization but increase interference. Conversely, strict isolation improves predictability but may leave resources underutilized. The optimal architecture therefore requires adaptive policies rather than a universal isolation strategy.

Discussion

The findings position the multi-tenant data lake as a socio-technical infrastructure rather than a conventional storage platform. Its primary theoretical contribution is the integration of three dimensions that are often considered separately: data heterogeneity, computational heterogeneity, and governance heterogeneity.

From a technical perspective, workload-aware orchestration is the central mechanism for transforming shared infrastructure into a scalable platform. Goyal (2025) provides a direct conceptual basis for linking data-lake scalability with multitenant AI workload orchestration. The architecture developed here extends this principle by connecting orchestration to metadata, governance, and tenant-specific policies.

The literature on computer vision further demonstrates why static resource allocation is insufficient. The datasets discussed by Afifi (2019), Amraoui et al. (2022), and Bashkirova et al. (2022) represent different image-processing requirements. A biometric recognition pipeline may require strict governance, UAV imagery may create high ingestion and storage pressure, and segmentation workloads may generate sustained accelerator demand. Treating these workloads identically would therefore create either inefficiency or inadequate service quality.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

The architecture also has implications for multimodal AI. Abdulmumin et al. (2022) demonstrate that multimodal resources involve relationships among different information modalities. Consequently, data-lake catalogs should represent semantic and operational relationships rather than only file locations. This enables more efficient discovery and supports reproducible AI pipelines.

However, the architecture has limitations. First, dynamic scheduling introduces additional control-plane complexity. Second, resource fairness is difficult to define because tenants may have different business or research priorities. Third, metadata quality becomes critical: incorrect classifications can lead to inappropriate access or scheduling decisions. Fourth, logical isolation may not eliminate every form of performance interference, especially when tenants compete for shared storage bandwidth or accelerators.

The governance dimension presents an additional challenge. Abdilla et al. (2020) and Birhane and Guest (2020) suggest that responsible AI cannot be reduced to technical efficiency. A highly optimized platform could still produce inappropriate outcomes if its policies ignore cultural, social, or epistemic considerations. Therefore, future implementations should permit governance policies to become machine-enforceable architectural rules while retaining appropriate human oversight.

A final limitation concerns empirical validation. The proposed architecture is conceptual and derived from literature-based requirements rather than measurements from a production deployment. Future experimental research should evaluate throughput, latency, resource utilization, fairness, isolation strength, energy consumption, and cost under realistic multitenant workloads. Such benchmarking would establish whether the proposed orchestration principles produce measurable improvements over static resource allocation.

Conclusion

This paper developed a conceptual multi-tenant data-lake architecture for scalable AI and big data workload management. The architecture integrates heterogeneous data ingestion, unified storage, metadata management, workload-aware orchestration, AI processing, and governance into a coordinated framework. Analysis of the supplied literature demonstrates that AI datasets differ substantially in data volume, modality, sensitivity, and computational requirements. Consequently, scalable infrastructure must dynamically manage both data and computation rather than treating the data lake as a passive repository.

The principal contribution of the proposed framework is the integration of tenant isolation with adaptive workload management. Logical data isolation, computational quotas, metadata-driven scheduling, and policy-aware governance allow multiple tenants to share infrastructure

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

ISSN-2750-1396

while reducing uncontrolled interference. The analysis also demonstrates that responsible AI considerations should be embedded directly within data architecture, particularly where datasets involve biometric, cultural, linguistic, or socially sensitive information.

The research further establishes that multitenancy introduces an unavoidable trade-off between resource efficiency and service isolation. Dynamic sharing can increase infrastructure utilization, whereas excessive sharing can reduce predictability and create contention. Effective orchestration must therefore balance performance, fairness, governance, and cost.

Future research should implement the proposed architecture in cloud or hybrid environments and evaluate it using representative computer-vision, multimodal, streaming, and large-scale AI workloads. Particular attention should be given to adaptive scheduling, accelerator sharing, data locality, tenant fairness, policy enforcement, and automated governance. The integration of these mechanisms could enable data lakes to evolve from scalable repositories into intelligent, policy-aware platforms capable of supporting increasingly diverse AI workloads.

REFERENCES

1. Abdilla, A., Arista, N., Baker, K., Benesiinaa bandan, S., Brown, M., Cheung, M., Coleman,M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P.-L.,Kealiikana kaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque,M., Lewis, J. E., Mahelona, K., Moses, C., Nahuewai, Isaac (’Ika’aka), Noe, K., Olson,D., Parker Jones, ˙Oiwi, Running Wolf, C., Running Wolf, M., Silva, M., Fragnito, S.,& Whaanga, H. (2020). Indigenous protocol and artificial intelligence position paper.
2. Abdulmumin, I., Dash, S. R., Dawud, M. A., Parida, S., Muhammad, S., Ahmad, I. S.,Panda, S., Bojar, O., Galadanci, B. S., & Bello, B. S. (2022). Hausa visual genome:A dataset for multi-modal English to Hausa machine translation. In Calzolari, N.,B´echet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H.,Maegaard, B., Mariani, J., Mazo, H., Odijk, J., & Piperidis, S. (Eds.),Proceedings ofthe Thirteenth Language Resources and Evaluation Conference, pp. 6471–6479 Marseille, France. European Language Resources Association.
3. Afifi, M. (2019). 11k hands: Gender recognition and biometric identification using a large dataset of hand images. Multimedia Tools and Applications,78, 20835–20854.
4. Amraoui, K. E., Lghoul, M., Ezzaki, A., Masmoudi, L., Hadri, M., Elbelrhiti, H., & Simo,A. A. (2022). Avo-airdb: An avocado uav database for agricultural image segmentation and classification. Data in Brief,45, 108738.
5. Ayachi, R., Afif, M., Said, Y., & Atri, M. (2020). Traffic signs detection for real-world application of an advanced driving assisting system using deep learning.Neural ProcessingLetters,51(1), 837–851.
6. Bashkirova, D., Abdelfattah, M., Zhu, Z., Akl, J., Alladkani, F., Hu, P., Ablavsky, V., Calli,B., Bargal, S. A., & Saenko, K. (2022). Zerowaste

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 187-197
OCLC – 1368736135

dataset: towards deformable objectseg mentation in cluttered scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21147–21157.

7. Birhane, A., & Guest, O. (2020). Towards decolonising computational sciences.ar Xiv preprint arXiv:2009.14258,1(1), non.
8. K. K. Goyal, "Scalable Data Lakes for AI Workloads: A Multitenant Architecture for Big Data Orchestration," 2025 IEEE International Conference on Computing (ICOCO), Kuching, Malaysia, 2025, pp. 266-271, doi: 10.1109/ICOCO67189.2025.11334100.