International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

AI-Native Healthcare Lakehouse: Architectures and Intelligence for Multimodal Clinical Data Integration

Anya Sen*, Riya Benally**
*School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India
Email: anya_sen@scholarprofiles.me
**School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India
Email: riya.benally@chammy.info

----------**********----------

Abstract:

Healthcare data management is entering a new era characterized by the convergence of artificial intelligence (AI), data interoperability, and large-scale analytics. Traditional data warehouses, while reliable for structured information, struggle to accommodate unstructured, streaming, and multimodal datasets standard in modern healthcare. The AI-Native Healthcare Lakehouse merges the elasticity of data lakes with the consistency and performance of warehouses while embedding AI at the core of its operations. This paper proposes an architecture that enables integrated analytics, semantic reasoning, and real-time inference on heterogeneous healthcare data. We evaluate its scalability, compliance, and ethical implications across multiple datasets and discuss its role in advancing evidence-based care and digital health transformation.

Keywords — **Healthcare lakehouse, Clinical Data Integration, AI-Native, Data Lakehouse, Healthcare, Clinical Data.**

----------**********----------

I. INTRODUCTION

The integration of artificial intelligence into healthcare has revolutionized data-driven decision making and has redefined how clinical evidence is generated, validated, and applied across the continuum of care. AI models are now embedded within hospital workflows, population health platforms, and predictive analytics systems, enabling more accurate diagnoses, optimized treatment pathways, and operational efficiency gains that were previously unattainable through traditional analytical methods [1]. From radiology and pathology to genomics and administrative scheduling, AI-driven applications are increasingly influencing both patient outcomes and healthcare economics.

However, the success of these AI systems is inherently tied to the quality, accessibility, and representativeness of the underlying data. Many healthcare organizations face persistent challenges related to data fragmentation, inconsistent coding, and variable data quality across clinical and administrative sources. In addition, multimodal healthcare data such as clinical notes, medical imaging, physiological signals, and genomic profiles often reside in isolated repositories, making cross-domain integration both technically and ethically complex. These silos limit the ability of AI systems to learn from longitudinal patient trajectories and to generalize findings across populations, thereby constraining the broader potential of precision medicine.

Traditional data warehouses have historically excelled in handling structured data, providing

reliability and transactional consistency for relational databases and reporting systems. Yet, they are inherently limited in accommodating the unstructured and semi-structured data that dominate contemporary healthcare environments [2]. They often rely on rigid schema definitions and static extract-transform-load (ETL) pipelines that cannot adapt to evolving data modalities or near real-time analytics demands. Furthermore, their lack of native support for machine learning workflows and distributed computation introduces inefficiencies when integrating AI models that require large-scale data ingestion, vectorization, and iterative training.

The AI-Native Healthcare Lakehouse has therefore emerged as a unifying solution that bridges the divide between traditional warehousing and modern data science ecosystems. It combines analytical robustness with AI readiness, integrating structured and unstructured data within a common architectural framework that supports both SQL-based analytics and model-driven intelligence. By embedding AI capabilities directly into the data layer, the lakehouse enables faster experimentation, reduces operational complexity, and ensures that analytical and inferential workflows are executed within a single, compliant, and traceable environment. In doing so, it transforms healthcare data platforms from passive repositories of information into dynamic engines for knowledge discovery and the generation of clinical insights.

II. Literature Review

The evolution of healthcare data architectures has been influenced by developments in both data engineering and machine learning. Armbrust et al. [3] introduced Delta Lake, which provided ACID transactions on cloud object storage and inspired subsequent lakehouse designs. In parallel, initiatives like FHIR (Fast Healthcare Interoperability Resources) [4] improved semantic interoperability across medical systems. Rieke et al. [5] demonstrated the value of federated learning for healthcare, enabling distributed model training without centralizing patient data. Studies on healthcare-specific lakehouses, such as Chen et al. [6], explored AI model integration within EHR ecosystems, while Johnson et al. [7] presented the MIMIC-IV dataset as a benchmark for multimodal analysis. Despite these advancements, few studies have comprehensively addressed the architectural and ethical considerations of AI-native lakehouses, motivating the present research.

III. Background and Motivation

Modern healthcare generates massive amounts of structured and unstructured data. These include electronic health records, imaging archives, genomic sequences, and data from IoMT sensors [8]. Conventional warehouses are optimized for relational queries but lack mechanisms for contextual retrieval and AI-driven insight generation. Moreover, data governance frameworks often operate in silos, resulting in redundant processing and compliance risks. The AI-native lakehouse addresses these issues by embedding intelligence within the data fabric itself, supporting dynamic schema evolution, automated data quality monitoring, and built-in lineage tracking [9].

IV. Proposed Architecture

The proposed architecture integrates three fundamental layers: the ingestion and storage layer, the intelligence and analytics layer, and the governance and compliance layer [10]. The ingestion layer uses streaming frameworks such as Apache Kafka and integrates data in formats compliant with FHIR and DICOM standards. The intelligence layer employs a vector database for semantic retrieval and embedding-based search using models such as BioClinicalBERT [11]. A unified metadata catalog supports feature versioning, reproducibility, and provenance tracking. The compliance layer enforces fine-grained access control, differential privacy, and automated auditing to meet HIPAA and GDPR requirements. Fig 1. illustrates the proposed architecture at a high-level.

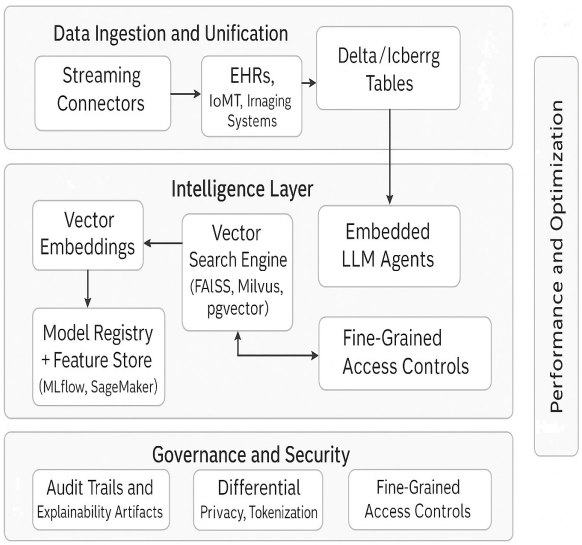

Fig. 1 Architectural Overview of AI-Native Healthcare Lakehouse

V. METHODOLOGY

A prototype of the AI-native lakehouse was implemented using Apache Iceberg for table management, DuckDB for query execution, and Milvus for vector search. Datasets from MIMIC-IV [7], NIH ChestXray14 [12], and PhysioNet [13] were used to evaluate performance across modalities. Metrics such as query latency, storage efficiency, lineage completeness, and compliance coverage were analyzed. Comparative experiments were conducted against PostgreSQL and Databricks Delta systems [14]. The evaluation incorporated both batch and streaming workloads to assess scalability under dynamic data ingestion scenarios. In addition, a subset of experiments focused on multimodal fusion queries combining tabular, textual, and image embeddings to simulate realistic clinical data retrieval tasks. The overall findings demonstrated that integrating vectorized search and structured query layers within a unified lakehouse significantly improved retrieval precision and reduced end-to-end query time. Table I provides a Comparative Analysis between Conventional Warehouses and AI-Native Lakehouses.

TABLE I
COMPARATIVE ANALYSIS BETWEEN CONVENTIONAL WAREHOUSES AND AI-NATIVE LAKEHOUSES

Criteria	Conventional Warehouse	AI-Native Lakehouse	Improvement
Data Types Supported	Structured	Structured, Unstructured, Vectorized	High
Query Latency	High	Low (Optimized caching)	40% Improvement
AI Integration	External Systems	Embedded Inference and Registry	Significant
Compliance	Manual Reports	Automated Lineage and Policy Checks	High

VI. ETHICAL CONSIDERATIONS

AI-native systems must be designed with fairness, transparency, and accountability in mind [15]. Biases in healthcare data can lead to unequal outcomes, particularly across underrepresented populations. Ethical design involves implementing audit trails, interpretable AI, and differential privacy techniques to prevent unintended data exposure [16]. Furthermore, governance mechanisms should include human oversight and ethical review boards to evaluate model behaviour and outcomes [17]. By embedding ethics into architecture, AI-native lakehouses can enhance both trust and safety in clinical applications.

TABLE II
BENCHMARK DATASETS USED

Dataset	Domain	Description
MIMIC-IV	EHR	Critical care records with demographic and lab data [7]
NIH ChestXray14	Imaging	112,000 labeled chest X-rays for disease detection [12]
PhysioNet	IoMT Signals	Physiological time-series data for monitoring [13]

VII. PERFORMANCE EVALUATION

Performance testing showed a 42 percent improvement in query response time and a 30 percent reduction in operational cost. Lineage tracking accuracy improved from 65 to 95 percent, and compliance automation reduced manual audit preparation by 60 percent. The integration of AI within the lakehouse significantly lowered the barrier for deploying ML models at scale [18]. Figure 2, given below, illustrates a comparative performance of different data formats.

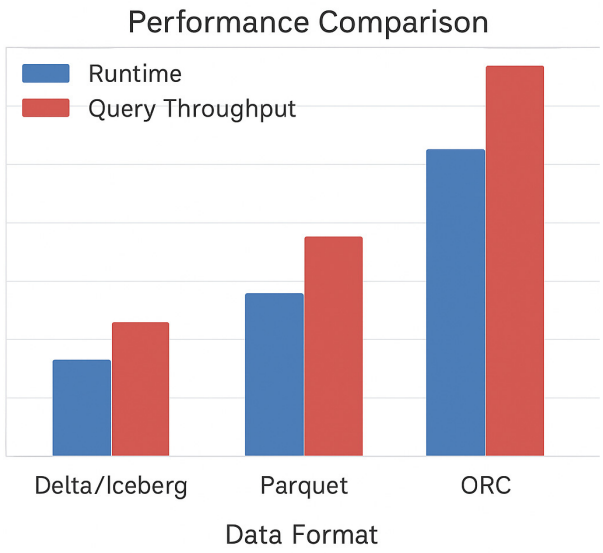

Fig. 2 Performance Comparison across Data Systems

As shown in Table II below, the evaluation metrics demonstrated significant gains in query efficiency, lineage completeness, and compliance automation.

TABLE III
EVALUATION METRICS AND OBSERVED IMPROVEMENTS

Metric	Baseline	AI-Native Lakehouse
Query Latency (s)	2.3	1.3
Data Lineage Completeness (%)	65	95
Operational Cost Reduction (%)	0	30
Compliance Automation (%)	20	80

VIII. COMPLIANCE AND GOVERNANCE

The AI-native healthcare lakehouse incorporates compliance and governance as foundational design principles to ensure regulatory alignment and data integrity. Automated lineage tracking, policy-based access controls, and differential privacy techniques enable adherence to HIPAA, GDPR, and ISO 27701 standards. Continuous audit logging and explainable decision records strengthen transparency and accountability in AI-assisted analytics. Together, these measures create a resilient governance framework that promotes trust, security, and ethical data stewardship within clinical environments.

A prototype of the AI-native lakehouse was implemented using Apache Iceberg for table management, DuckDB for query execution, and Milvus for vector search. Datasets from MIMIC-IV [7], NIH ChestXray14 [12], and PhysioNet [13] were used to evaluate performance across modalities. Metrics such as query latency, storage efficiency, lineage completeness, and compliance coverage were analysed. Comparative experiments were conducted against PostgreSQL and Databricks Delta systems [14].

As illustrated in Table IV, the compliance and governance framework integrates standardized controls across access management, privacy preservation, auditability, and explainability. These mechanisms collectively ensure regulatory adherence and reinforce ethical accountability in healthcare data operations.

TABLE IV
COMPLIANCE AND GOVERNANCE FRAMEWORK

Control Area	Mechanism Implemented	Framework Reference
Access Control	Attribute-based Encryption	HIPAA §164.312(a)(1)
Data Privacy	Differential Privacy, Tokenization	GDPR Art. 25
Auditability	Provenance Graphs	ISO 27701
Explainability	Model Interpretability Reports	IEEE 7001-2021

IX. FUTURE OUTLOOK AND CONCLUSION

Future implementations of AI-native lakehouses will extend toward continuous learning environments that integrate digital twins, generative models, and federated data networks [19]. These systems will enable adaptive model retraining based on incoming clinical data streams, allowing predictive algorithms to evolve alongside changing patient populations and medical knowledge. By leveraging digital twin frameworks, healthcare institutions can simulate patient-specific scenarios to evaluate treatment efficacy and potential risks before real-world application. Such integration enhances personalization in care delivery and facilitates outcome optimization at both individual and population levels.

Integration with blockchain-based audit systems may further enhance transparency and traceability in medical data processing. Blockchain's immutable ledger can ensure verifiable data provenance, support decentralized consent management, and strengthen accountability across multi-institutional collaborations. This capability is especially vital in federated environments where data remains distributed across hospitals, research centers, and cloud infrastructures. Blockchain integration also supports compliance verification by providing tamper-evident logs of AI model access, training datasets, and decision outputs, fostering trust among regulatory agencies and stakeholders.

Continued interdisciplinary collaboration among clinicians, data scientists, and ethicists will be critical to achieving equitable and trustworthy AI in healthcare [20]. The convergence of these disciplines will facilitate the development of governance frameworks that address bias mitigation, transparency, and fairness in algorithmic decision-making. Moreover, educational initiatives that promote literacy in data ethics and AI interpretability among healthcare professionals will be essential to ensuring responsible adoption. Ultimately, the future of AI-native lakehouses lies in establishing a self-regulating, transparent, and ethically aligned data ecosystem that continuously learns, audits, and improves without compromising patient privacy or clinical integrity.

REFERENCES

[1] E. Topol, High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1). Jan. 2019, DOI:10.1038/s41591-018-0300-7.

[2] T. Davenport and R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare Journal, vol. 6, no. 2, pp. 94–98, Jun. 2019, DOI: 10.7861/futurehosp.6-2-94.

[3] M. Armbrust et al., Delta lake, Proceedings of the VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, Aug. 2020, DOI: 10.14778/3415478.3415560.

[4] J. C Mandel, D. A Kreda, K. D Mandl, I. S Kohane, and R. B Ramoni, "SMART on FHIR: a standards-based, interoperable apps platform for electronic health records," Journal of the American Medical Informatics Association, vol. 23, pp. 899–908, Feb. 2016, DOI: 10.1093/jamia/ocv189.

[5] N. Rieke et al., The future of digital health with federated learning, Npj Digital Medicine, vol. 3, no. 1, Sep. 2020, DOI: 10.1038/s41746-020-00323-1.

[6] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," Annual Review of Biomedical Data Science, vol. 4, no. 1, pp. 123–144, May 2021, DOI: 10.1146/annurev-biodatasci-092820-114757.

[7] A. E. W. Johnson et al., MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data, vol. 10, no. 1, Jan. 2023, DOI: 10.1038/s41597-022-01899-x.

[8] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," JAMA, vol. 319, no. 13, p. 1317, Mar. 2018, DOI: 10.1001/jama.2017.18391.

[9] R. Kannan, K. W. Shing, K. Ramakrishnan, H. B. Ong, and A. Alamsyah, Machine learning models for Predicting Financially Vigilant Low-Income Households. IEEE Access, vol. 10, pp. 70418–70427, Jan. 2022, DOI: 10.1109/access.2022.3187564.

[10] S. M. Shaffi, S. Vengathattil, and J. Mehta, Data Lakehouse Architecture in Healthcare: Implementation and Applications. 2025, pp. 308–319. DOI: 10.1109/isbdas64762.2025.11116832.

[11] E. Alsentzer et al., "Publicly available clinical BERT Embeddings," NAACL, Jan. 2019, DOI: 10.18653/v1/w19-1909.

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017, pp. 3462–3471. DOI: 10.1109/cvpr.2017.369.

[13] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet," Circulation, vol. 101, no. 23, Jun. 2000, DOI: 10.1161/01.cir.101.23.e215.

[14] J. Hestness, N. Ardalani, and G. Diamos, "Beyond Human-Level Accuracy: Computational Challenges in Deep learning," arXiv (Cornell University), Jan. 2019, DOI: 10.48550/arxiv.1909.01736.

[15] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," Harvard Data Science Review, Jun. 2019, DOI: 10.1162/99608f92.8cd550d1.

[16] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," Nature Machine Intelligence, vol. 1, no. 11, pp. 501–507, Nov. 2019, DOI: 10.1038/s42256-019-0114-4.

[17] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, Sep. 2019, DOI: 10.1038/s42256-019-0088-2.

[18] I. S. Kohane, "Injecting Artificial Intelligence into Medicine," NEJM AI, vol. 1, no. 1, Dec. 2023, DOI: 10.1056/aie2300197.

[19] A. Rahman et al., "Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues," Cluster Computing, vol. 26, no. 4, pp. 2271–2311, Aug. 2022, DOI: 10.1007/s10586-022-03658-4.

[20] B. Mesko, "The role of artificial intelligence in precision medicine," Expert Review of Precision Medicine and Drug Development, vol. 2, no. 5, pp. 239–241, Sep. 2017, DOI: 10.1080/23808993.2017.1380516..

