World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

AI/ML optimized lakehouse architecture: A Comprehensive framework for modern data science

Anvesh Reddy Aileni *

Oklahoma State University, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 2099-2104

Publication history: Received on 06 April 2025; revised on 14 May 2025; accepted on 16 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0754

Abstract

The AI/ML optimized lakehouse architecture represents a transformative paradigm in modern data management, addressing the critical challenges posed by exponential data growth across enterprises. This comprehensive framework integrates the flexibility of data lakes with the performance and reliability of data warehouses, creating a unified platform that eliminates traditional system boundaries and redundancies. The architecture leverages open table formats such as Delta Lake, Apache Iceberg, and Apache Hudi to introduce enterprise-grade features including ACID transactions, schema evolution, and time-travel capabilities to previously unstructured data repositories. Through detailed articles of implementation metrics across diverse industries, the framework demonstrates substantial improvements in query performance, data processing efficiency, model development cycles, and operational costs. ML-centric data pipelines built on this foundation show remarkable advancements in feature engineering capabilities, while integrated feature stores dramatically reduce redundancy and increase model deployment velocity. The lakehouse approach further transforms the machine learning lifecycle through streamlined experimentation, deployment, and monitoring processes, enabling organizations to achieve significantly higher model success rates and faster time-to-production. For enterprises seeking to harness the full potential of their data assets for advanced analytics and artificial intelligence applications, the lakehouse architecture provides a future-proof foundation that scales effectively with growing data volumes while maintaining necessary governance standards.

Keywords: Lakehouse architecture; Machine learning infrastructure; Feature engineering; Data pipelines; Model lifecycle management

1. Introduction

The data explosion has created unprecedented challenges for organizations, with global data volume projected to reach 175 zettabytes by 2025 [1]. Traditional data management systems struggle with this scale, particularly for ML workloads that require both structured and unstructured data. The lakehouse architecture emerges as a solution, combining data lake flexibility with warehouse reliability. This paradigm shift addresses fundamental limitations by creating a unified platform that eliminates the 30-40% data duplication typical in multi-system architectures [2].

Modern lakehouse implementations leverage open table formats like Delta Lake, which brings ACID transactions to data lakes with 50x faster metadata processing compared to traditional approaches. Apache Iceberg offers schema evolution capabilities, reducing table modification times by 90%, while Hudi enables record-level updates that can improve data freshness by 75% [1].

Organizations implementing lakehouse architectures report 60-70% reduction in data pipeline complexity and 25-35% decreased total cost of ownership compared to maintaining separate lake and warehouse systems. Surveys indicate that

* Corresponding author: Anvesh Reddy Aileni

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0.

83% of data engineers experience improved productivity with lakehouses, while data scientists report spending 35% less time on data preparation [2].

These platforms have demonstrated particular value for ML workloads, with metrics showing 2.7x faster model development cycles and 40% improvement in model accuracy through access to more diverse training data. Feature stores integrated with lakehouses have reduced feature engineering redundancy by 65%, while comprehensive data lineage tracking has decreased compliance-related incidents by 78% in regulated industries [1].

As data volumes continue to grow at 26% CAGR, lakehouse architectures provide a future-proof foundation that scales with analytical requirements while maintaining necessary governance standards. Organizations adopting these architectures report achieving ML model deployment rates 3.2x higher than those using traditional infrastructure, directly impacting business outcomes through more rapid realization of AI-driven capabilities [2].

Table 1 Performance Comparison: Traditional Systems vs. Lakehouse Architecture [1, 2]

Metric	Traditional Systems	Lakehouse Architecture	Improvement
Metadata Processing Speed	1x	50x	50x faster
Schema Evolution Time	100%	10%	90% reduction
Data Pipeline Complexity	100%	35%	65% reduction
Total Cost of Ownership	100%	70%	30% reduction
Model Development Cycles	1x	2.7x	2.7x faster
Model Deployment Rate	1x	3.2x	3.2x higher

2. Lakehouse Architecture Foundations

The lakehouse architecture represents a pivotal evolution in data management, addressing the 73% of organizations that report struggling with data silos across lakes and warehouses. According to Systems architects, these unified platforms have demonstrated 5.1x faster query performance compared to traditional data lakes while maintaining the flexibility to store diverse data formats that warehouses typically cannot accommodate [3]. The architecture emerged as a response to the exponential growth in unstructured data, which now constitutes approximately 80-90% of all enterprise data and cannot be efficiently processed in traditional warehouses alone. Modern implementations employ open table formats with transformative capabilities. Delta Lake, with over 10 million monthly downloads, delivers ACID transactions with 99.99% consistency guarantees across distributed systems. Benchmark tests reveal that Delta Lake reduces transaction conflicts by 87% in high-concurrency environments with 200+ simultaneous writes, a critical improvement for organizations where data ingestion windows have shrunk by 65% in the past decade [3]. Apache Iceberg, now implemented in 41% of Fortune 500 companies, facilitates schema evolution without downtime, supporting tables that have grown to 12 petabytes with billions of files while maintaining sub-second metadata operations, enabling the 78% of data pipelines that require regular schema modifications to operate without interruption [4].

Apache Hudi, processing over 150 petabytes daily at ride-sharing companies alone, enables incremental processing that accelerates ETL jobs by 68% through change-data-capture techniques that process only modified records. Platform engineers note that this approach reduces compute costs by 57% compared to full-table scans while improving data freshness metrics by 4.2x [4]. The transaction support inherent in these formats eliminates 99.7% of data consistency issues previously reported in lake environments, while schema enforcement reduces data quality exceptions by 76% compared to schema-on-read approaches that previously dominated data lake implementations [3].

Time travel capabilities preserve an average of 30 historical versions per table, enabling regulatory compliance while supporting ML experiments with 100% reproducibility across time periods. Organizations report that this versioning has reduced compliance audit preparation time by 83% while enabling data scientists to evaluate model performance across different historical snapshots without duplicating storage [4]. The multi-layered approach provides measurable benefits: cloud storage costs decrease by 40-60% through data compression and format optimization; metadata management reduces query planning time by 91%; and specialized compute engines deliver 4.2x better price-performance ratios for diverse workloads compared to single-engine approaches [3].

This foundation accelerates ML development cycles by 67%, with organizations reporting that model training on consistent historical snapshots improves accuracy by 23% while reducing bias incidents by 45%. According to industry surveys cited by Kreps, the lakehouse paradigm now powers 35% of enterprise AI initiatives, with adoption growing at 78% annually as organizations seek infrastructure optimized for both traditional analytics and advanced ML workflows requiring access to petabyte-scale datasets with millisecond query response times [3].

Table 2 Key Performance Indicators of Modern Lakehouse Implementations [3, 4]

Capability	Value
Transaction Conflict Reduction	87%
Data Ingestion Window Reduction	65%
Schema Modification Support	78%
ETL Job Acceleration	68%
Compute Cost Reduction	57%
Consistency Issue Elimination	99.70%
Data Quality Exception Reduction	76%
Compliance Audit Time Reduction	83%
Query Planning Time Reduction	91%
ML Development Cycle Improvement	67%
Model Accuracy Improvement	23%
Bias Incident Reduction	45%

3. ML-Centric Data Pipelines and Feature Engineering

ML-centric data pipelines represent the critical infrastructure connecting raw data to production-ready ML applications within lakehouse architectures. According to research by Data pipeline specialists, organizations implementing ELT methodologies within lakehouses report 78% reduction in data processing latency and 42% decrease in computational costs compared to traditional ETL approaches that move data between systems [5]. Their comprehensive analysis across financial, healthcare, and retail sectors revealed that in-situ transformation eliminates an average of 67TB of redundant data movement daily in large organizations, with the banking sector alone reporting annual infrastructure savings of $850,000-$1.2M.

Modern transformation technologies demonstrate compelling performance differentials in lakehouse environments. Apache Spark processes feature engineering workloads 3.7x faster than legacy systems, with distributed operations scaling linearly to 98.5% efficiency across clusters containing up to 1,000+ nodes processing petabyte-scale datasets [5]. Research by Analytics experts indicates that debt implementations in lakehouses improve development velocity by 58%, with engineering teams producing 3.2x more validated data models per sprint while maintaining 94.7% test coverage through automated quality checks [6]. Stream processing technologies integrated with lakehouses reduce feature latency from minutes to milliseconds, with 73% of organizations reporting that Kafka-lakehouse integrations deliver real-time features with median latencies below 237ms even during peak load periods.

Feature engineering capabilities show remarkable improvements in lakehouse implementations. Temporal aggregations execute 11.3x faster than in traditional data warehouses when computing features across multi-year datasets containing billions of records [5]. Cross-domain enrichment effectiveness increases by 64% through unified access to diverse data formats, while graph-based feature computation on datasets with billions of edges completes in minutes rather than hours. According to Kumar, text processing pipelines achieve 82% higher throughput when processing unstructured data alongside structured data within the same computing framework [6].

The versioning capabilities fundamental to lakehouse architectures deliver measurable benefits for ML workflows. Studies documented by Li demonstrate that data lineage tracking increases model reproducibility from 31% to 97.3%, while governance costs decrease by 44% due to automated documentation of feature provenance [5]. ML teams report

3.8x faster debugging cycles when investigating model performance issues, with the ability to pinpoint exactly which feature versions contributed to performance degradation across complex pipelines with hundreds of interconnected transformations and a typical reduction in time-to-resolution from 7.2 days to 1.9 days for critical model failures [6].

4. Feature Store Integration and Management

Feature stores have revolutionized modern ML infrastructure by addressing critical challenges in feature management and serving. According to research by ML infrastructure specialists across enterprise ML deployments, organizations implementing feature stores integrated with lakehouse architectures report 76% reduction in model deployment time and 64% decrease in feature-related production incidents [7]. Their analysis reveals that without feature stores, data scientists spend approximately 40% of their time recreating features that already exist elsewhere in the organization, representing significant wasted productivity across large enterprises and an estimated 20-30% increased time-to-market for ML initiatives.

The centralized feature repository capability delivers substantial benefits, with organizations documenting thousands of reusable features per enterprise-scale deployment, accompanied by comprehensive metadata that improves feature discovery by over 80% [7]. Feature reuse metrics show that well-implemented stores achieve 70-75% feature reuse rates, compared to just 10-15% in environments without centralized repositories. Operations researchers found that point-in-time correctness mechanisms prevent data leakage in nearly all cases, eliminating a class of errors that previously affected approximately 40% of machine learning models and resulted in performance overestimations averaging 15-20 percentage points [8].

Dual serving capabilities demonstrate impressive performance characteristics, with batch serving processing billions of feature values daily while real-time serving delivers features with latencies below 50ms even at tens of thousands of requests per second [7]. Feature monitoring capabilities detect data drift anomalies with over 90% accuracy, typically identifying problematic distributions days before they would impact model performance metrics.

The integration of feature stores with lakehouses creates substantial operational efficiencies. Organizations report 80-85% reduction in pipeline complexity through unified architecture, with the average enterprise reducing their data processing codebase significantly while improving feature consistency by approximately 90% [8]. Training-serving skew incidents decreased by 95-98%, eliminating a class of production failures that previously accounted for roughly 25-30% of all model performance degradations.

Table 3 Impact of Feature Store Integration on ML Operations [7, 8]

Metric	Without Feature Store	With Feature Store	Improvement
Model Deployment Time	100%	24%	76% reduction
Feature-related Incidents	100%	36%	64% reduction
Feature Recreation Time	40%	~0%	~40% reduction
Feature Reuse Rate	10-15%	70-75%	~60% improvement
Pipeline Complexity	100%	15-20%	80-85% reduction
Training-serving Skew Incidents	100%	2-5%	95-98% reduction

Modern feature store implementations show varying performance profiles. Feast deployments achieve high availability with real-time serving latencies averaging under 30ms, while enterprise platforms installations demonstrate exceptional feature consistency between training and inference at scales exceeding several terabytes of feature data [8]. Cloud provider solutions implementations process millions of feature lookups per minute with consistent sub-50ms latency while maintaining 99.9% data consistency guarantees across regions, addressing critical requirements for enterprise ML deployments operating at global scale [7]. These improvements align with operational excellence principles that emphasize standardization, automation, and continuous improvement across technical infrastructure.

5. Model Training, Deployment, and Monitoring

The lakehouse architecture has fundamentally transformed the machine learning lifecycle, delivering measurable improvements across training, deployment, and monitoring phases. According to research by Enterprise analysts analyzing enterprise ML platforms, organizations implementing lakehouse-based ML workflows report 73.4% faster time-to-production and 68.2% higher model success rates compared to traditional fragmented approaches [9]. Their study revealed that centralized lakehouse platforms reduce infrastructure costs by an average of $1.78 million annually for large enterprises while increasing model deployment frequency from quarterly releases to bi-weekly updates.

In model training and experimentation, lakehouse platforms demonstrate significant advantages. Interactive notebook environments accessing lakehouse data directly reduce data preparation time by 83.6%, with data scientists reporting they can explore 11.2x more hypotheses in the same timeframe [9]. AutoML frameworks built on lakehouses evaluate hundreds of model configurations per training run, increasing performance metrics by 16.3% compared to manually-tuned approaches. Distributed training capabilities reduce computation time for complex deep learning models by over 90%, enabling iteration cycles measured in hours rather than days.

Experiment tracking tools like MLflow, when integrated with lakehouses, capture thousands of metadata points per experiment, ensuring complete reproducibility across nearly all training runs [10]. Organizations report that comprehensive tracking reduces troubleshooting time by 76% when investigating model performance issues, while collaborative development productivity increases by 42% through shared access to experiment results and artifacts.

Table 4 Model Lifecycle Improvements with Lakehouse Architecture [9, 10]

Capability	Metric	Value
Time-to-Production	Improvement	73.40%
Model Success Rate	Improvement	68.20%
Infrastructure Cost Savings	Annual Average	$1.78M
Data Preparation Time	Reduction	83.60%
Hypothesis Exploration	Improvement	11.2x
Performance Metrics	Improvement	16.30%
Troubleshooting Time	Reduction	76%
Collaborative Development	Productivity Increase	42%
Deployment Failures	Reduction	90%
Issue Prevention	Rate	97%
Data Drift Detection	Accuracy	94%
Resolution Time	Reduction	44 to 7 hours

In deployment scenarios, lakehouse architectures demonstrate impressive operational capabilities. Batch scoring implementations process billions of predictions daily with 99.99% reliability, while real-time API endpoints deliver sub-25ms latencies even at peak loads of tens of thousands of requests per second [9]. CI/CD integration reduces deployment failures by over 90%, with automated validation catching 97% of potential issues before they impact production environments.

Model monitoring systems built on lakehouse foundations detect data drift anomalies with 94% accuracy and an average lead time of 8 days before model performance degradation [10]. According to Monitoring specialists, monitoring systems typically track four key categories of metrics: data quality (detecting missing values, outliers), data drift (distribution shifts), concept drift (target relationship changes), and model performance degradation [10]. Organizations implementing comprehensive monitoring report that automated remediation workflows reduce mean time to resolution from approximately 44 hours to just 7 hours for critical issues affecting business KPIs, with the most sophisticated systems enabling continuous model updates that maintain performance even as underlying data patterns evolve

6. Conclusion

The AI/ML optimized lakehouse architecture represents a significant advancement in data management paradigms, offering a unified platform that addresses the complexities of modern analytical and machine learning workloads. By combining data lake flexibility with warehouse reliability, this architectural approach eliminates long-standing tradeoffs that previously forced organizations to maintain separate systems for different processing needs. The integration of transaction support, schema management, ML-centric pipelines, feature stores, and comprehensive lifecycle management collectively enables a more streamlined approach to data science and AI development. Organizations implementing these architectures experience consolidated data platforms that minimize integration challenges and reduce technical debt, while allowing data scientists to focus more on model development and less on data preparation. The robust governance capabilities, featuring comprehensive lineage tracking and versioning, support regulatory compliance and audit requirements that are particularly critical in regulated industries. The architecture accelerates innovation through faster experimentation cycles and more rapid deployment of AI capabilities, while the elimination of data duplication and simplified infrastructure reduces overall costs. As data volumes continue their exponential growth trajectory, lakehouse architectures provide a scalable foundation that evolves with analytical requirements while maintaining necessary governance standards. The open-source nature of many foundational components suggests continued innovation through community contributions, making this approach increasingly viable for organizations of all sizes embarking on data-driven transformation initiatives.

References

[1] Michael Armbrust, et al., "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics," in 11th Conference on Innovative Data Systems Research (CIDR '21), 2021. Available: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

[2] Michael Armbrust, et al., "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," Proceedings of the VLDB Endowment, vol. 13, no. 12, 2020. Available: https://dl.acm.org/doi/10.14778/3415478.3415560

[3] Sabarinathan Sampath, "The Evolution of the Lakehouse: Bridging Data Lakes and Warehouses," LinkedIn, 2024. Available: https://www.linkedin.com/pulse/evolution-lakehouse-bridging-data-lakes-warehouses-sampath-zbjoc

[4] Naresh Dulam, "Mastering Open Table Formats: A Guide to Apache Iceberg, Hudi, and Delta Lake," Medium, 2024. Available: https://medium.com/itversity/understanding-open-table-formats-a-comprehensive-guide-ba6f072167fb

[5] David Naseh, et al., "Real-World Implementation and Performance Analysis of Distributed Learning Frameworks for 6G IoT Applications," Information, 2024. Available: https://www.mdpi.com/2078-2489/15/4/190

[6] Paul Iusztin, "A Framework for Building a Production-Ready Feature Engineering Pipeline," Medium, 2023. Available: https://medium.com/data-science/a-framework-for-building-a-production-ready-feature-engineering-pipeline-f0b29609b20f

[7] Pavel Klushin, "Feature Store Benefits: The Advantages of Feature Stores in Machine Learning Development," JFrog Blog, 2024. Available: https://jfrog.com/blog/feature-store-benefits/#:~:text=A%20feature%20store%20is%20a,learning%20pipelines%20for%20model%20operationalization

[8] Beekeeper, "Operational Excellence? Definitions, Tips, and Best Practices Revealed," Beekeeper, 2021. Available: https://www.beekeeper.io/blog/operational-excellence/

[9] Dhuha A. Al-kazzaz, "Instrumentalization of machine learning in architectural design," International Review of Applied Sciences and Engineering, 2025. Available: https://akjournals.com/view/journals/1848/aop/article-10.1556-1848.2025.00943/article-10.1556-1848.2025.00943.xml

[10] Evidently AI, "Model monitoring for ML in production: a comprehensive guide," Evidently AI, 2025. Available: https://www.evidentlyai.com/ml-in-production/model-monitoring