marlabs
driving digital agility

WHITEPAPER

Building a Data Lakehouse in the Age of AI

A Blueprint for Business **Resilience** and **Speed**

TABLE OF CONTENTS

OVERVIEW

Executive Summary

As enterprises race to harness the transformative potential of AI, their ability to manage, govern, and activate data at scale has become a defining competitive factor. Yet, legacy data architectures—fragmented, rigid, and costly—are fundamentally misaligned with the demands of AI/ML workloads, real-time analytics, and rapid innovation cycles. The future belongs to organizations that can turn data into a system of intelligence—fueling faster decisions, smarter products, and more adaptive operations.

This whitepaper presents a strategic deep dive into the evolution of modern Data Lakehouses that unify the Data Lakes' flexibility with the performance and governance of Data Warehouses. It explores how the Lakehouse architecture empowers enterprises to streamline data pipelines, enable real-time operational insights, and drive innovation through integrated AI/ML capabilities.

We also take a closer look under the hood to examine the architecture's key layers, from scalable object storage and open processing engines to built-in ML tooling and security frameworks. Furthermore, we will also discuss real-world examples that illustrate the tangible business value unlocked across industries.

SETTING THE SCENE

Introduction

The rapid surge in enterprise data and explosive maturation of artificial intelligence (AI) technologies are outpacing the capabilities of traditional data architectures. While racing to harness generative AI (GenAI), real-time analytics or predictive modeling, many businesses today find themselves trapped by legacy data platform constraints. Modernization of these aging systems with incremental band-aid upgrades or mere 'lifts and shifts' can prove costly—missed market opportunities, regulatory failures, and operational fragility.

The need of the hour? A fundamental reinvention of how we store and access data. Enter the Data Lakehouse: a next-generation architecture that combines the best of Data Lakes and Data Warehouses, while surpassing their limitations. It acts as a single repository for all types of data—structured, semi-structured, and unstructured. This unified approach has enabled businesses to overcome the challenge of data silos, ensuring that all users, from business analysts to data scientists, are working with the same, consistent data. And for businesses that aim to embed AI across their operations, the Data Lakehouse provides the foundation needed to turn today's data deluge into tomorrow's strategic advantage.

THE PROBLEM

A Macroscopic View - When Data Debt Becomes Business Risk

Organizations today can no longer treat data as an operational byproduct. It must be leveraged as a strategic asset that should flow freely, be analyzed instantly, and continuously inform decisions. Most legacy systems were designed in an era where AI-powered analysis and instantaneous feedback loops were not the norm. Today, as we push toward real-time responsiveness and predictive insights, the gap between what businesses need and what their data estates can deliver has never been wider. Let's look at how some of the legacy bottlenecks manifest across industries—proving modernization isn't optional, but existential.

Siloed data

When information is locked across different domains, internal and external, it creates fractured data ecosystems. This fragmentation results in higher storage costs, increased difficulty of governance, and higher manual intervention. For instance, in manufacturing, if the sensor data is isolated from ERP quality logs, teams can find themselves diagnosing quality issues weeks after they occur, rather than detecting them in real time.

Slow decision latency

A direct consequence of disparate data is the extension of ETL cycles and batch pipelines. This results in the generation of outdated insights that hamper critical business decision-making and market responsiveness.

Financial enterprises must be wary of this, particularly when relying on batch-loaded anti-fraud dashboards. Due to the delay, teams may not spot suspicious transactions until hours after they've occurred, placing customers and reputation at risk.

Performance constraints

Data Warehouses struggle to scale cost-effectively, while Data Lakes lack the optimized query performance required for fast, real-time analytics. As the data volume grows, organizations find themselves resorting to spinning up oversized clusters to handle peak loads, resulting in runaway costs.

Retail businesses can be susceptible to this challenge, such as when their teams process multi-terabyte clickstream and sales data during promotional events. This often escalates cloud bills that can spike three to four times.

Non-support for AI workloads

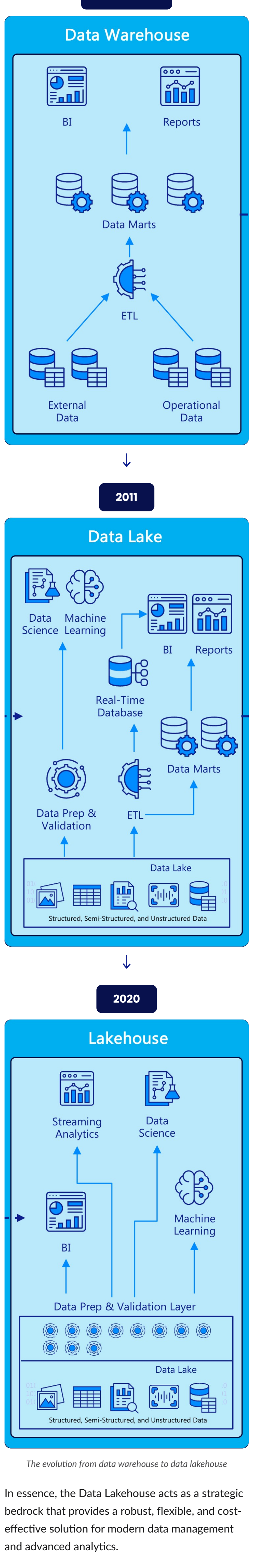
Modern AI workflows require flexible ingestion, schema evolution, feature store support, and rapid retraining capabilities. Legacy platforms—built for static, structured data—aren't equipped for these demands.

For example, in healthcare, AI models are assisting medical professionals to accurately and quickly diagnose complex diseases and conditions. But integrating clinical records and imaging onto legacy systems is labor-intensive and slow, delaying predictive diagnostics that could otherwise improve patient outcomes.

THE ANSWER

The Data Lakehouse – Redefining Enterprise Data Management

So far, we've discussed how traditional storage solutions fall short in terms of modern enterprise needs. To become truly data-driven in real-time, businesses require a platform that combines flexibility, performance, and governance at scale. The Data Lakehouse answers this call—a fundamental unified re-architecture combining the advantages of the Data Lake and Warehouse, while dissolving the tradeoffs of both.



The evolution from data warehouse to data lakehouse

In essence, the Data Lakehouse acts as a strategic bedrock that provides a robust, flexible, and cost-effective solution for modern data management and advanced analytics.

An architectural overview

At its core, the Lakehouse consists of five key layers that build upon each other to create an environment where raw, semi-structured, and structured data coexist.

The Data Lakehouse Architecture

Sequential flow through five core layers from data collection to analysis

1 Ingestion Layer

The entry point for all data. Collects raw data from databases, IoT devices, logs, and applications in both batch and real-time streams.

Pushes data into the Storage Layer

2 Storage Layer

Foundation layer providing cost-effective, scalable storage for vast amounts of data in open formats. Combines data lake flexibility with data warehouse reliability.

Receives data from Ingestion Layer

Indexed by Metadata Layer

3 Metadata Layer

Critical unifying layer providing a centralized catalog of all data assets. Manages schemas, versions, and statistics while enabling ACID transactions and time travel capabilities.

Catalogs Storage data

Provides query-able index for API Layer

4 API Layer

Standardized interface for tools and engines to access, process, and query data. Decouples storages from compute, enabling different applications to work with consistent data using familiar APIs like SQL.

Queries Metadata

Exposes data to Consumption Layer

5 Consumption Layer

User-facing layer where data value is realized. Data analysts, scientists, and business users access data for BI, reporting, advanced analytics, and machine learning.

Connects to API Layer for queries, models, and visualizations

The open table format

Data Lakehouses which are open source-enabled rely on open table formats (Delta Lake, Apache Iceberg, or Hudi) that maintain ACID transactions, schema evolution, and fine-grained versioning. Essentially, this brings database-like features and reliability alongside the flexible and cost-effective storage of a data lake.

As a result, businesses can ensure high data reliability by tracking every update, delete, or merge, thereby empowering data stewards with unified governance controls. With built-in lineage and auditability, compliance requirements are met more easily, and trust in analytics outcomes is significantly higher.

Decoupling of storage and compute

Traditional storage solutions have tightly integrated compute and storage, forcing rigid sizing and costly over-provisioning to handle peak workloads. In contrast, the Lakehouse separates these layers: storage remains an elastic, cost-effective object store, while compute engines spin up and down independently. Due to this, businesses gain the following benefits:

- Pay only for actual compute seconds used
- Scale resources precisely for batch, streaming, or interactive workloads
- Run BI and AI-driven queries alongside real-time inference without contention

Native analytics and machine learning

By providing a single source of truth and seamless integration with popular analytics and machine learning frameworks, the Lakehouse accelerates the full data lifecycle—from exploration and feature engineering to model training and deployment. Data scientists can pull from the same tables that BI analysts use, reducing handoffs and inconsistencies. The result is a streamlined path from raw data to production-grade models, enabling enterprises to operationalize AI at scale and unlock real-time, predictive insights.

INNOVATION

Driving AI and GenAI Innovation with the Data Lakehouse

AI applications are no longer confined to innovation labs. As the technology continues to be embedded across core business functions and products, the ability to move fast—from data ingestion to model deployment—has become a critical imperative. The Data Lakehouse plays a central role in this transformation by offering a unified platform that simplifies training, deployment, and governance, dramatically accelerating AI lifecycles.

Streamlining of data pipelines

Modern AI models, especially large language models (LLMs), thrive on high-volume, high-quality training data. The Lakehouse simplifies the creation of production-grade datasets through built-in support for scalable ingestion, schema evolution, and version-controlled transformations. This means that ETL processes can be managed natively where data movement is minimal, and both batch and streaming ingestion are supported. Thus, teams can create reliable, reproducible training pipelines without complex data duplication or brittle orchestration.

Key benefits include:

- Faster access to real-time and historical data
- Reduced data duplication and manual intervention
- Version control and reproducibility of training datasets

Fueling GenAI use cases

GenAI applications such as intelligent chatbots and summarization tools depend on the ability to retrieve, reason over, and generate responses based on enterprise knowledge. Legacy platforms crumble under GenAI's unstructured data demands, causing model hallucinations and compliance risks. The Data Lakehouse overcomes this challenge by allowing structured, semi-structured, and unstructured data to coexist under one governance model. This enables the holistic indexation of enterprise knowledge followed by retrieval of relevant chunks at inference time using techniques like Retrieval-Augmented Generation (RAG).

Furthermore, built-in cataloging and access control features ensure privacy and security of sensitive enterprise data. Assets such as support tickets, policy documents, or technical manuals can be securely integrated into GenAI implementations without introducing compliance risks.

Driving intuitive, intelligent search

As organizations build more AI applications, the demand for semantic search—retrieving data based on meaning rather than keyword matching, continues to grow. The Lakehouse enables this with native storage of vector embeddings alongside transactional and analytical data. This unlocks the ability to perform similarity search and contextual retrieval at scale—powering everything from next-gen recommendation engines to AI copilots that understand user intent.

Enhancing model development velocity

A major bottleneck in scaling AI is operational friction—data scattered across systems, inconsistent environments, and fragmented teams. The Lakehouse breaks down these silos by providing a single, governed workspace for data engineers, analysts, and ML teams. Thanks to its native acceleration capabilities, organizations can speed up their time-to-market for AI lifecycles:

MLflow Registry

The MLflow Registry provides a centralized repository for managing the complete lifecycle of machine learning models, enabling one-click promotion from experimental validation to production deployment environments. This streamlines MLOps workflows by ensuring version control, lineage tracking, and seamless model serving.

In-built Feature Store

Integrated feature stores enable the creation and reuse of highly curated, point-in-time correct features across various machine learning models, fostering consistency and accelerating feature engineering processes. This critical component ensures data integrity and reduces redundant data preparation efforts for scalable AI development.

Unity Catalog Governance

Unity Catalog enforces a unified governance framework across all data and ML assets within the Lakehouse, automating compliance, access control, and auditing for models and their underlying data. This robust catalog ensures data security, lineage transparency, and regulatory adherence, foundational for trusted AI systems.

SPEED

Unlocking Real-time Analytics and Faster Decision-making

The ability to act on fresh data with agility can spell the difference between winning and waiting. The Data Lakehouse delivers true streaming ingestion-whether it's clickstream events, IoT telemetry, or financial transactions—into a unified store where updates are immediately available for analysis. By eliminating batch windows and sync delays, organizations gain a continuous pulse on their operations.

Once data lands, Lakehouse engines leverage advanced indexing, caching, and vectorized execution to deliver sub-second query performance—even over billions of rows. This instantaneous responsiveness means analysts and business users can explore current trends, run ad hoc scenarios, or power live dashboards without resorting to pre-aggregated summaries.

Beyond pure reporting, Lakehouses support operational analytics by embedding queries directly into transactional workflows. Customer-facing applications can tap into up-to-the-second insights—such as inventory availability or dynamic pricing—without the latency and complexity of ETL handoffs. The result is smarter, proactive decision-making at the edge of the business.

A Marlabs Case Study: A Leading US Telecom Enterprise Gains 360° Customer View and Reduces Churn by 8%

A large American telecommunications operator struggled to efficiently manage their data infrastructure. Their teams faced multiple issues including delayed data integrations due to multi-vendor systems reliance, lack of real-time user activity visibility, and unoptimized plans driving customer churn.

Marlabs helped them overcome these challenges by designing and implementing a Data Lakehouse on AWS. The solution enabled the telecom enterprise to enhance various aspects of their operations by:

- Establishing a unified and holistic view of their customers
- Optimizing their data management pipelines with seamless integration and real-time processing
- Minimizing data quality issues

Ultimately, by implementing the Data Lakehouse, they were able to better understand their customers and drive greater satisfaction and loyalty.

CONSIDERATIONS

Understanding the Trade-offs

While the Data Lakehouse brings a centralized approach to data management, it is not without its complexities. Understanding these nuances is crucial for successful implementation, so let's delve into some of them.

The sophistication in ensuring unified governance at scale

Combining diverse data types under one roof demands a governance framework that's versatile. In practice, defining and enforcing consistent policies for access control, data quality, lineage, and metadata across object stores and query engines often requires custom tooling or extensive configuration. As the user base, data domains, and compliance requirements grow, the governance landscape can become fragmented, leading to both operational overhead and steeper learning curves for data stewards.

The constant need for performance tuning to support hybrid workloads

One of the Data Lakehouse's selling points is handling both BI-style queries and ML-driven exploration in the same system. Yet optimizing for that full spectrum can be tricky. BI workloads typically involve analytical queries on historical, structured data, often requiring fast aggregations and reporting. AI/ML workloads, on the other hand, might involve iterative processing of large datasets, feature engineering, model training, and real-time inference, often leveraging unstructured or semi-structured data.

Calibrating a single system to perform optimally for such diverse demands is a delicate balancing act. This can include careful resource allocation, workload management, and potentially even different compute clusters or configurations for different types of operations within the same lakehouse environment.

The risk of vendor lock-ins

Open standards like Delta Lake and Apache Iceberg form the foundation of many lakehouse solutions. But many vendors layer proprietary enhancements on top such as advanced indexing, performance acceleration or built-in governance dashboards. While these features can speed up time-to-value, they also raise the cost and complexity of a migration later. Once the pipelines are built and dashboards closely tied to a vendor's SDK or UDF framework, moving to another platform can feel like untangling a web of dependencies.

The unplanned escalation of costs

Running micro-batches or stream-processing jobs, coupled with large-scale feature engineering and model training, can quickly drive up compute and storage bills. Furthermore, uncontrolled auto-scaling, excessive data retention in hot storage, or poorly optimized UDFs in Python/Scala can ramp up expenses. To keep budgets in check, teams must invest in rigorous usage monitoring, quota management, and often build automated shutdown or tier-ing strategies for idle resources.

Comparing the Data Lakehouse to Alternative Modern Data Architectures

While the lakehouse stands tall as a data management powerhouse, it's helpful to see how it stacks up to other prominent solutions. Here's a quick look at different common alternatives, highlighting where the lakehouse shines—and where it might play second fiddle.

Cloud-Native Data Warehouse with GenAI Hooks

DEFINITION

- A fully managed, elastic SQL engine augmented with built-in generative AI services for automated insights and natural-language querying.

PROS VS DATA LAKEHOUSE

- Instant AI/ML insights (summaries, anomaly detection)
- Best-in-class analytics performance

CONS VS DATA LAKEHOUSE

- Tied to vendor's data formats & GenAI models
- Unstructured or streaming data often needs pre-staging, adding latency

Data Mesh Architectures

DEFINITION

- A federated approach that treats data as domain-owned products, enabling decentralized ownership, self-service infrastructure, and cross-domain interoperability.

PROS VS DATA LAKEHOUSE

- Empowers domain experts on data quality & semantics
- Reduces central ingestion bottlenecks

CONS VS DATA LAKEHOUSE

- Query federation across systems can suffer on performance & consistency
- Overhead of catalogs, APIs, cross-domain SLAs

Data Fabric Overlays for Decentralized Governance

DEFINITION

- A metadata-centric layer that provides unified cataloging, policy enforcement, and lineage tracking across existing data platforms without relocating data.

PROS VS DATA LAKEHOUSE

- Leverages best-of-breed storage & compute engines
- Single pane for governance over multiple platforms

CONS VS DATA LAKEHOUSE

- No native storage—real-time enforcement relies on connectors that may lag
- Additional licensing & architectural complexity for the overlay layer

Building an AI-ready Data Lakehouse – A Snapshot

The Data Lakehouse isn't just a convergence of lakes and warehouses—it is an architectural rethink designed to handle the complexity and velocity needs of today's enterprise data. Underneath its simplicity lies a layered system that balances flexibility with performance, openness with control, and experimentation with compliance. Let's take a closer look at the inner workings of the Lakehouse to understand how it powers real-time, AI-ready, and regulation-aware data capabilities at scale.

The Storage Layer

The Lakehouse relies on a storage layer built for scalability and efficiency. Cloud-native object stores like Amazon S3 and ADLS offer virtually unlimited capacity and pay-as-you-go flexibility. With open formats such as Apache Parquet and Delta Lake, all data stay accessible, queryable, and free from vendor lock-in.

Processing Engines

Support for diverse analytic capabilities is enabled through multiple engines. Apache Spark handles large-scale batch processing, while Apache Flink enables low-latency, real-time analysis. SQL accelerators allow analysts and BI tools to query data directly in the lake, delivering warehouse-like performance without duplication or data movement.

The AI/ML Layer

A key Lakehouse feature is the native support for the full ML lifecycle. This layer includes model registries for versioning and governance, feature stores for consistent training and inference, and experiment tracking for iterative development. Integration with LLM frameworks allows for the building and fine-tuning of LLMs on private data.

In-built Governance

Governance is central to the Lakehouse, ensuring data integrity, security, and compliance. Role-based access control limits users to authorized data, lineage tracking records data flow for transparency, and encryption protects sensitive information both at rest and in transit.

OUTLOOK

The Way Forward with Marlabs

The enterprise data landscape is at a decisive inflection point. As AI rapidly goes from being experimental to foundational, organizations will be compelled to rethink their data strategies from the ground up. The next wave of competitive advantage will not be determined merely by who has the most data, but by who can activate it with speed, trust, and intelligence.

More importantly, the future will demand not just scalable infrastructure, but intelligent infrastructure—architectures that can self-optimize, auto-govern, and dynamically adapt to new data, workloads, and compliance mandates. The Lakehouse is not just keeping pace with this vision, it is actively shaping it.

Marlabs can be your partner of choice, proactively empowering you with the right tools and solutions to be at the forefront of the changing tides. We bring a deep expertise in transforming and modernizing data ecosystems for AI, at scale. Our proprietary data strategy framework, Modern Analytics Platform (MAP), is designed to help businesses plan and build a holistic roadmap towards enterprise data-driven decision intelligence. No matter where you are in your data journey, our pre-built accelerators and personalized solutions enable you to drive business success with speed and confidence.

About Marlabs

Marlabs designs and develops digital solutions with data at the center. We leverage our deep data expertise and cutting-edge technology to empower businesses with actionable insights and achieve improved digital outcomes.

Marlabs' data-first approach intersects with AI and analytics, digital product engineering, and advisory services to build and scale digital solutions. We work with leading companies around the world to make operations sleeker, keep customers closer, transform data into decisions, boost legacy system performance, and seize novel opportunities in new digital revenue streams.

Marlabs is headquartered in New Jersey, with offices in the US, Germany, Canada, Brazil, Bulgaria, and India.

Marlabs Inc.(Global Headquarters) One Corporate Place South, 3rd Floor, Piscataway NJ - 08854-6116,
Tel: +1 (732) 694 1000 Fax: +1 (732) 465 0100, Email: contact@marlabs.com.

https://www.facebook.com/marlabsinc/
https://www.linkedin.com/company/marlabs
https://x.com/marlabs?lang=en
https://www.youtube.com/@Marlabsinc

marlabs driving digital agility