Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Building an AI-Ready Data Strategy Using Lakehouse Technology

Jyoti Aggarwal
Carnegie Mellon University, USA
Corresponding Author: Jyoti Aggarwal, **E-mail**: jyoti.aggarwal.one@gmail.com

| ABSTRACT
This article explores how organizations can leverage Lakehouse technology to build robust AI-ready data strategies. Lakehouse architecture represents a paradigm shift in data management by unifying data lake flexibility with data warehouse reliability. The integration of this technology enables organizations to overcome traditional barriers to AI implementation through unified storage, efficient processing, and comprehensive governance. By examining key components including data ingestion, preparation, metadata management, governance, and scalable infrastructure, the article illustrates how Lakehouse technology establishes a foundation for advanced AI applications like predictive analytics, recommendation systems, natural language processing, and automated decision-making. The article addresses common implementation challenges and provides solutions for data governance, quality assurance, infrastructure scaling, and integration with AI/ML tools, offering organizations practical guidance for transforming their data infrastructure into a catalyst for AI innovation.

| KEYWORDS

Lakehouse architecture, AI-ready data strategy, unified data management, data governance, scalable infrastructure

| ARTICLE INFORMATION

ACCEPTED: 19 April 2025 **PUBLISHED:** 08 May 2025 **DOI:** 10.32996/jcsts.2025.7.3.76

Introduction

In today's rapidly evolving technological landscape, artificial intelligence (AI) has become a cornerstone of innovation and competitive advantage. Organizations are increasingly recognizing the transformative potential of AI, with the global AI market expected to grow at a compound annual growth rate (CAGR) of 37.3% from 2023 to 2030 [1]. However, the effectiveness of AI systems is fundamentally dependent on the quality, accessibility, and organization of the underlying data. Traditional data architectures have struggled to meet these demands, with organizations facing significant challenges in data management that have historically limited AI adoption and effectiveness.

The emergence of data lakehouse technology represents a paradigm shift in how organizations approach their data infrastructure for AI readiness. The data lakehouse architecture addresses the limitations of traditional data lakes and data warehouses by combining the best features of both: the flexibility and cost-efficiency of data lakes with the reliability, performance, and data management features of data warehouses [1]. This unified approach has proven particularly valuable for organizations implementing machine learning operations (MLOps), where end-to-end data pipelines require seamless integration between data storage, processing, and model deployment. The data lakehouse provides a single platform for all data needs, eliminating the complexity and overhead of managing multiple specialized systems that previously required complex ETL processes and created data silos.

AI initiatives demand robust computational resources alongside sophisticated data management capabilities. According to industry analysis, AI workloads require 5-20 times more computing power than traditional enterprise applications, highlighting the importance of infrastructure considerations in an AI-ready data strategy [2]. The data lakehouse paradigm addresses these requirements by offering a scalable, unified environment that supports diverse workloads while maintaining data consistency.

Organizations implementing data lakehouse architectures have reported substantial improvements in operational efficiency, with some experiencing up to 30% reduction in data management costs and significant acceleration in time-to-insight for AI applications [1].

The path to AI readiness through lakehouse technology involves more than just architectural choices. A comprehensive approach must address infrastructure flexibility, operational efficiency, and adaptability to evolving AI requirements. Modern AI workloads demand infrastructure that can efficiently process data at scale, with requirements for high-performance computing that grows approximately 10x every year for leading-edge AI models [2]. The data lakehouse serves as an ideal foundation for this AI infrastructure, providing the necessary data management capabilities alongside integration with compute resources optimized for AI workloads, including GPUs and specialized AI accelerators.

This article explores how organizations can leverage lakehouse technology to build robust AI-ready data strategies that support advanced analytics, machine learning, and AI initiatives. We examine the architectural principles, implementation approaches, and best practices that enable organizations to transform their data infrastructure into a foundation for AI innovation, with particular attention to the challenges and opportunities presented by this emerging paradigm.

Characteristic	Traditional Data Lake	Traditional Data Warehouse	Lakehouse Architecture
Data Storage Format	Raw, diverse formats	Structured, proprietary	Open formats with schema enforcement
Transaction Support	Limited or none	ACID transactions	ACID transactions on open formats
Performance	Variable, often slower for complex queries	Optimized for analytical queries	Near data warehouse performance with data lake flexibility
Schema Management	Schema-on-read	Schema-on-write	Flexible schema evolution with validation
Storage Cost	Low	High	Low with tiered options
Data Type Support	Structured, semi-structured, unstructured	Primarily structured	All data types with optimized processing
Governance	Limited, often add-on	Comprehensive but rigid	Integrated, scalable governance
Real-time Processing	Limited	Batch-oriented	Supports both batch and streaming

Table 1: Core Characteristics of Lakehouse Architecture [1,2]

Understanding Lakehouse Technology

Lakehouse architecture represents a paradigm shift in data management by combining the best elements of data lakes and data warehouses—hence the portmanteau "Lakehouse." This unified approach addresses the limitations of traditional data architectures while providing a solid foundation for AI development. The concept emerged as a response to the growing challenges organizations face when trying to implement advanced analytics and machine learning using existing data infrastructure. According to Zaharia et al., organizations typically spend 30-40% of their data engineering time on maintaining complex ETL pipelines between data lakes and data warehouses, representing a significant opportunity cost that directly impacts AI initiatives [3]. The need for a unified approach becomes particularly apparent when considering that data scientists report spending up to 80% of their time on data preparation rather than actual analysis and model development.

The evolution of lakehouse architecture has been driven by fundamental limitations in both data lakes and data warehouses. Data lakes provide inexpensive storage for large volumes of diverse data but lack data management features, while data warehouses offer robust management capabilities but at significantly higher costs and with limitations on data types and formats. An analysis of these traditional architectures reveals that data lakes typically store data at costs that are 10-50 times lower per terabyte than specialized data warehouses, but they suffer from reliability issues that make them unsuitable for many business-critical applications [3]. The economic advantage of data lakes, combined with the increasing volumes of data required for AI applications, makes the lakehouse approach particularly compelling from both a technical and financial perspective.

Modern lakehouse implementations leverage several key technological advancements to deliver their unified capabilities. A critical foundation is the implementation of metadata and data layout techniques that enable efficient computation directly on low-cost object storage. For example, Delta Lake's transaction log approach provides ACID guarantees without requiring a separate database system, enabling atomic operations even on cloud object stores that only support eventual consistency [3]. This transaction log approach adds only minimal overhead—approximately 1% additional storage compared to raw data—while providing significant benefits in terms of data reliability and consistency, critical factors for AI applications where data integrity directly impacts model accuracy.

Performance optimization represents another crucial aspect of lakehouse architecture. Through techniques like data skipping, caching, and query optimization, modern lakehouses achieve query performance that approaches traditional data warehouses for many workloads. Benchmark results from real-world implementations show that lakehouses can execute complex SQL queries with latencies comparable to traditional data warehouses, within 2.5x for most analytical queries while maintaining the cost advantages and flexibility of the underlying data lake storage [3]. This performance improvement removes one of the historical barriers to consolidating analytical infrastructure and enables a unified approach to data management for AI applications.

Lakehouse architectures also address the critical need for schema enforcement and data governance. By implementing schema validation that can be enforced and evolved, lakehouses prevent data quality issues that frequently plague traditional data lakes. Research conducted with small and medium enterprises (SMEs) implementing lakehouse architectures found that enforcing schemas at the storage layer reduced data quality incidents by an average of 64% compared to their previous data lake implementations [4]. The same study revealed that consistent data quality leads to more reliable AI model training, with participating SMEs reporting a 42% reduction in model retraining requirements due to data quality issues after implementing lakehouse architecture.

Direct SQL access stands as one of the most transformative aspects of lakehouse architecture for democratizing data usage across organizations. Traditional business intelligence teams can continue using familiar SQL tools while data science teams leverage the same underlying data for ML training. Analysis of adoption patterns in SMEs shows that organizations implementing lakehouse architectures experience, on average, a 3.2x increase in the number of employees actively using data for decision-making within 12 months of implementation [4]. This democratization of data access drives significant business value, with surveyed organizations reporting a 37% increase in data-driven decision-making across departments previously disconnected from analytical capabilities.

The support for diverse data types within a unified platform has proven particularly valuable for AI development. Traditional data warehouses excel at handling structured data but struggle with semi-structured and unstructured formats that are increasingly important for AI applications. Zaharia et al. note that modern lakehouses can efficiently process complex data types, including images, video, audio, and tex,t alongside traditional structured data, eliminating the need for separate specialized systems [3]. This capability is especially valuable for multi-modal AI models that combine information from different data types. Survey data from SME implementations indicates that organizations leveraging unified lakehouse platforms for diverse data types reduced their time-to-deployment for complex AI applications by an average of 41% compared to those using separate specialized systems for different data formats [4].

Real-time capabilities represent another critical component of modern lakehouse architectures. By supporting streaming data ingestion and processing alongside batch operations, lakehouses enable organizations to build AI systems that respond to changing conditions in near real-time. A technical analysis of lakehouse implementations in SMEs found that 73% of surveyed organizations were able to reduce their data latency from hours or days to minutes or seconds after implementation [4]. This reduction in latency translated directly to business value, with 61% of these organizations reporting measurable improvements in operational decision-making in time-sensitive business processes.

The economic impact of lakehouse adoption has been particularly significant for resource-constrained organizations. A comprehensive analysis of lakehouse implementations in 32 European SMEs revealed an average reduction in total data infrastructure costs of 26% compared to maintaining separate systems for different data workloads [4]. These cost reductions came primarily from the consolidation of infrastructure (47% of savings), reduced data movement and duplication (31% of savings), and decreased maintenance requirements (22% of savings). Beyond cost reduction, the same study found that SMEs implementing

lakehouse architectures were able to launch new data-driven products and services 2.8 times faster than with their previous architecture, representing a significant competitive advantage in rapidly evolving markets.

This architectural approach transforms raw data into AI-ready assets by providing a centralized hub where data can be stored, processed, and accessed efficiently for model training and deployment. For AI initiatives specifically, the unified approach of lakehouse architecture addresses several critical challenges identified in the literature. Organizations adopting lakehouse architecture for AI development report significant improvements in collaboration between data engineering and data science teams, with implementation studies showing an average 58% reduction in iteration time for model development workflows [4]. As the technology continues to mature, lakehouse architecture is increasingly positioned as the foundation for organizations seeking to maximize the value of their data assets for AI innovation.

Benefit Area	Improvement Metric	Primary Enabling Feature
Data Engineering Efficiency	Reduction in time spent on ETL processes	Unified storage and compute
AI Model Development	Acceleration in model training cycles	Direct access to diverse data types
Data Quality	Reduction in data quality incidents	Integrated quality monitoring and schema enforcement
Operational Costs	Decrease in total infrastructure costs	Separation of storage and compute with cloud scaling
Time-to-Insight	Improvement in time from data acquisition to insight	Streamlined data pipelines
Data-Driven Decision Making	Increase in employees using data	SQL compatibility and self-service capabilities
Compliance Management	Reduction in compliance documentation time	Automated lineage tracking and governance
Infrastructure Utilization	Increase in resource utilization	Dynamic scaling and workload management

Table 2: Implementation Benefits of Lakehouse Technology for AI Readiness [3,4]

Components of an AI-Ready Data Strategy

An effective AI-ready data strategy built on Lakehouse technology encompasses several critical components that work together to transform raw data into valuable assets for AI development. The implementation of these components in a unified environment addresses key challenges that organizations face when developing and deploying AI systems, creating a foundation for sustainable innovation and competitive advantage.

1. Data Ingestion and Integration

The first step involves establishing robust pipelines for ingesting data from various sources—databases, applications, IoT devices, and external systems. Lakehouse architecture supports both batch and streaming ingestion methods, ensuring that data is available in near real-time when needed. According to IDC research, the volume of data that requires processing is increasing at an unprecedented rate, with the global datasphere projected to grow from 33 zettabytes in 2018 to 175 zettabytes by 2025, representing a compound annual growth rate of 27% [6]. This massive growth is driving the need for more efficient ingestion mechanisms that can handle diverse data types and sources. The expansion is particularly pronounced in the enterprise sector,

where IDC predicts that enterprise data creation and management could represent nearly 60% of the global datasphere by 2025, up from 30% in 2015. This trend underscores the increasing responsibility that organizations have in managing and deriving value from vast quantities of data, making efficient ingestion and integration capabilities a fundamental requirement for any AI-ready data strategy.

The integration aspect of data ingestion presents significant challenges for AI applications, particularly when dealing with the diversity of data types that modern AI systems require. IDC's research indicates that by 2025, nearly 30% of the global datasphere will be real-time in nature, necessitating integration frameworks that can handle streaming data alongside traditional batch processes [6]. This shift toward real-time data is especially relevant for AI applications that must respond to changing conditions or make predictions based on the most current information available. Lakehouse architectures address this challenge by providing unified frameworks for ingesting and processing both real-time and historical data, creating an integrated foundation for AI development that spans the temporal spectrum from historical analysis to real-time decision making.

2. Data Preparation and Quality

Raw data rarely meets the quality requirements for AI applications, making comprehensive data preparation a critical component of any AI-ready data strategy. According to industry research, data scientists typically spend about 80% of their time on data preparation tasks, with only 20% left for actual model development and deployment [5]. This significant time investment highlights the centrality of data preparation to the AI development process and the potential value of technologies that can streamline and automate aspects of this work. The Lakehouse paradigm addresses this challenge by providing integrated tools for data transformation, cleaning, and quality assessment within the same environment where data is stored and analyzed, reducing the need for time-consuming data movement between specialized systems.

A comprehensive data preparation strategy includes data cleaning and normalization, feature engineering, outlier detection and handling, missing value imputation, and format standardization. The impact of data quality on AI performance is substantial, with research indicating that improving data quality can deliver a 35-40% increase in model accuracy for complex prediction tasks [5]. This direct relationship between data quality and model performance underscores the importance of robust data preparation capabilities within an AI-ready data architecture. Lakehouse technology facilitates these processes through the integration of specialized transformation engines and quality assessment tools operating directly on the source data, allowing organizations to implement consistent quality standards across their entire data estate while maintaining lineage and provenance information that is crucial for model transparency and governance.

Feature engineering, a critical aspect of AI data preparation, involves transforming raw data into meaningful inputs for machine learning models. This process requires deep domain knowledge combined with technical expertise, making it one of the most challenging aspects of AI development. Industry analysis suggests that proper feature engineering can improve model performance by 20-30% compared to using raw or minimally processed features [5]. The unified computational environment provided by Lakehouse architectures enables more efficient feature engineering by allowing data scientists to work directly with the full dataset rather than constrained samples, leading to more robust and generalizable features. The ability to version features and share them across different AI applications further enhances productivity and ensures consistency, addressing common challenges in organizations where multiple teams develop similar features independently, leading to redundant effort and inconsistent results.

3. Metadata Management

Effective metadata management is crucial for an AI-ready environment, encompassing technical metadata describing data structures, business metadata providing context and definitions, and operational metadata tracking data lineage and usage patterns. The importance of metadata increases with the scale and diversity of the data environment, with organizations that implement comprehensive metadata management reporting significantly improved data discovery capabilities, enhanced governance, and accelerated AI development cycles. Research indicates that organizations with mature metadata practices can reduce the time required to identify and access relevant datasets by up to 70%, directly impacting development velocity for AI initiatives [5]. This efficiency gain is particularly valuable in large enterprises where data assets may be distributed across multiple repositories and systems, creating discovery challenges that can significantly slow AI development if not properly addressed.

Data lineage, a key aspect of metadata management, becomes increasingly important as AI systems influence critical business decisions and processes. The ability to trace data from its origin through various transformations to its ultimate use in AI models is essential for both regulatory compliance and operational trust. According to industry surveys, 65% of organizations cite data lineage as a significant challenge in their AI governance efforts, highlighting the need for integrated lineage tracking capabilities [5]. Lakehouse architectures address this challenge by automatically capturing lineage information as data moves through various processing stages, creating a comprehensive record that supports both governance requirements and operational troubleshooting when issues arise. This automated approach to lineage tracking reduces the manual documentation burden on data teams while providing more complete and accurate information than manual methods typically achieve.

4. Governance and Security

AI applications often deal with sensitive information, making governance and security paramount concerns in an AI-ready data strategy. The regulatory landscape for AI continues to evolve, with frameworks like the EU's General Data Protection Regulation (GDPR) and emerging AI-specific regulations imposing significant compliance requirements on organizations developing and deploying AI systems. Industry research indicates that organizations with mature data governance frameworks are 60% more likely to comply with relevant regulations without requiring last-minute remediation efforts [5]. This proactive compliance capability not only reduces regulatory risk but also accelerates AI development by establishing clear guidelines and guardrails that development teams can follow throughout the project lifecycle, rather than discovering compliance issues during final review stages that may require substantial rework.

Lakehouse architectures incorporate fine-grained access controls, data encryption (both at rest and in transit), audit logging, compliance monitoring, and data masking and anonymization capabilities. These security features ensure that data is not only accessible for legitimate AI use cases but also properly protected against unauthorized access or misuse. The integration of security and governance directly into the data platform rather than as separate overlay systems delivers particularly compelling benefits for organizations developing AI at scale. Industry analysis suggests that integrated approaches reduce security-related incidents by approximately 45% compared to environments where security is implemented through separate systems and processes [5]. This reduction is largely attributable to the consistent application of security policies and the elimination of security gaps that can occur when data moves between systems with different security models and capabilities.

The ethical dimensions of AI governance are receiving increased attention as organizations deploy AI systems that impact customers, employees, and other stakeholders. Research indicates that 73% of organizations consider ethical considerations an important aspect of their AI governance frameworks, though only 37% report having robust processes in place to address these concerns [5]. Lakehouse architectures support ethical AI governance through capabilities like comprehensive data lineage, access controls that reflect ethical guidelines, and integrated monitoring tools that can detect potential bias or other ethical issues in data used for AI training and deployment. These capabilities enable organizations to implement "ethics by design" approaches where ethical considerations are integrated into the development process rather than evaluated only after systems are built.

5. Scalable Infrastructure

AI workloads can be computationally intensive and unpredictable, making scalable infrastructure a critical component of an AI-ready data strategy. The computational requirements for AI are growing at an accelerating rate, with IDC projecting that the amount of data analyzed by enterprise systems will grow by a factor of 6 between 2018 and 2025 [6]. This growth in analytical workloads, combined with the increasing complexity of AI models, creates significant infrastructure challenges that traditional fixed-capacity approaches struggle to address effectively. Lakehouse architectures address these challenges through elastic compute resources that scale up or down based on demand, separation of storage and compute to optimize costs, comprehensive resource monitoring and optimization, and support for specialized hardware like GPUs that can dramatically accelerate AI workloads.

The shift toward cloud deployment models for AI infrastructure reflects the importance of scalability and elasticity. IDC projects that by 2025, 49% of the data stored worldwide will reside in public cloud environments, enabling organizations to leverage the dynamic scalability that cloud platforms provide [6]. This trend is particularly relevant for AI workloads, which often exhibit significant variability in resource requirements across different development and deployment phases. Training complex models may require substantial computational resources for relatively short periods, while inference workloads typically require less computation per instance but must scale to handle varying request volumes. Lakehouse architectures deployed in cloud environments can adapt to these changing requirements, providing high-capacity resources when needed for training and efficient scaling for inference workloads, optimizing both performance and cost.

The growth of edge computing represents another important trend in AI infrastructure, with IDC predicting that by 2025, 175 zettabytes of data will be created worldwide, and 90 zettabytes of this will be created by IoT devices [6]. This distributed data creation creates new challenges for AI architectures, as organizations must decide where to process and analyze data that originates at the edge of their networks. Lakehouse architectures are evolving to address these challenges through hybrid approaches that combine edge processing for latency-sensitive applications with centralized processing for complex analytics and model training. This hybrid approach enables organizations to optimize their AI infrastructure for both performance and cost, processing data where it makes the most sense based on the specific requirements of each use case.

By implementing these five components comprehensively within a Lakehouse architecture, organizations establish a robust foundation for AI development and deployment. The integrated nature of the Lakehouse approach delivers synergistic benefits across components, enabling organizations to move beyond the limitations of traditional data architectures and create truly AI-ready environments that can adapt to evolving requirements and opportunities. As the global datasphere continues its exponential

growth, with IDC projecting an increase from 33 zettabytes in 2018 to 175 zettabytes by 2025 [6], the importance of efficient, scalable, and integrated approaches to data management for AI will only increase, making the Lakehouse paradigm an increasingly valuable architectural approach for organizations seeking to maximize the value of their data assets.

Component	Function	Strategic Value for AI
Batch Ingestion	Processing large volumes of data in scheduled intervals	Building comprehensive historical datasets for training
Streaming Ingestion	Processing data in real-time as it's generated	Enabling real-time prediction and model serving
Change Data Capture	Identifying and capturing changes in source systems	Maintaining synchronization with operational systems
Schema Registry	Centralizing schema definitions and evolution	Ensuring consistency across diverse data formats
Data Connector Framework	Standardizing connections to various data sources	Simplifying integration of new data sources
Data Validation	Verifying data against quality rules during ingestion	Preventing poor quality data from entering the system
Metadata Extraction	Automatically capturing source metadata	Enabling data discovery and lineage tracking
Transformation Pipeline	Converting raw data into analytics-ready formats	Preparing data for immediate use in AI training

Table 3: Data Ingestion and Integration Components in Lakehouse Strategy [5,6]

Applications of Lakehouse-Powered AI

The integration of Lakehouse technology in an AI-ready data strategy enables a wide range of applications that leverage the unified data processing capabilities, scalable infrastructure, and governance features of the Lakehouse architecture. These applications span multiple domains and deliver significant business value by harnessing both historical and real-time data for intelligent decision-making and enhanced user experiences.

Predictive Analytics

By centralizing historical and real-time data, Lakehouse platforms provide the foundation for developing accurate predictive models. These can forecast customer behavior, anticipate equipment failures, optimize inventory levels, and more. The impact of predictive analytics powered by Lakehouse architecture is particularly evident when considering the performance improvements that result from unified data access. According to NetApp research, organizations adopting unified data architectures for AI workflows report that data scientists spend up to 38% less time on data preparation and access, allowing them to focus more on model development and refinement [7]. This efficiency gain directly contributes to improved model quality and faster time-to-insight, critical factors in competitive markets where decision speed can determine success or failure.

The effectiveness of predictive models for business forecasting has shown significant improvement when implemented within Lakehouse environments. The ability to seamlessly combine structured transactional data with unstructured market signals enables more comprehensive forecasting models that capture complex relationships between variables. NetApp's analysis of customer implementations indicates that unified data architectures reduce the time required to incorporate new data sources into

forecasting models by an average of 60%, enabling more agile responses to changing market conditions [7]. This agility is particularly valuable in volatile industries where rapid adaptation to emerging trends can create substantial competitive advantage.

In manufacturing and industrial settings, predictive maintenance represents one of the highest-value applications of Lakehouse-powered AI. The ability to combine real-time sensor data with historical maintenance records, equipment specifications, and even unstructured information like technician notes creates a comprehensive foundation for equipment health monitoring and failure prediction. Organizations implementing unified data architectures for predictive maintenance report an average 30% reduction in unplanned downtime and maintenance cost savings of 25-30% compared to traditional preventive maintenance approaches [7]. These benefits stem directly from the Lakehouse architecture's ability to handle diverse data types within a unified analytical environment, enabling more accurate prediction of developing equipment issues before they result in failures.

Recommendation Systems

Recommendation engines benefit from the ability to combine diverse data sources (user interactions, product metadata, contextual information) within the Lakehouse environment to create personalized experiences. The business impact of effective recommendation systems is substantial, with research by Sunil Gajavada indicating that personalized recommendations can increase conversion rates by 5-9% and average order values by 3-5% in e-commerce environments [8]. These improvements translate directly to revenue growth, explaining the continued investment in recommendation technology across multiple sectors including retail, media, and financial services.

The challenges of implementing effective recommendation systems are closely aligned with the strengths of Lakehouse architecture. Traditional recommendation engines often struggle with data freshness, limited feature sets, and scalability constraints during high-traffic periods. Lakehouse platforms address these challenges by providing unified access to historical and real-time data, enabling more sophisticated recommendation algorithms that consider a broader range of contextual factors. According to industry analysis, companies implementing recommendation systems on unified data platforms achieve 55% lower latency for real-time recommendations compared to systems built on siloed data architectures [8]. This performance improvement enables more contextually relevant recommendations that respond to customer behavior as it occurs rather than relying solely on historical patterns.

The evolution of recommendation systems toward multi-modal approaches that incorporate diverse data types (text, images, structured user data) is particularly well-supported by Lakehouse architecture. Traditional data warehouses excel at analyzing structured data but struggle with unstructured content like product images or customer reviews, while data lakes can store diverse data types but lack the performance for real-time recommendation serving. The Lakehouse approach bridges this gap by providing both storage flexibility and query performance, enabling more sophisticated recommendation models. Organizations implementing multi-modal recommendation systems report engagement increases of 12-18% compared to traditional collaborative filtering approaches that rely solely on structured interaction data [8]. This engagement improvement directly impacts key business metrics including customer retention, which typically increases by 7-10% after implementation of advanced recommendation capabilities.

Natural Language Processing (NLP)

Text analysis and NLP applications require access to vast amounts of textual data alongside structured information about entities and relationships—a combination well-supported by Lakehouse architecture. The technical requirements for enterprise NLP applications highlight the value of unified data platforms, with NetApp's research indicating that organizations implementing NLP solutions require an average of 3-5 distinct data sources to achieve desired accuracy and coverage [7]. Traditional data architectures that separate structured and unstructured data processing create significant integration challenges in this context, requiring complex data movement and synchronization processes that introduce latency and reliability concerns.

The implementation of NLP applications in customer service environments demonstrates the particular value of Lakehouse architecture for real-time text analytics. The ability to analyze customer inquiries as they occur while simultaneously accessing relevant customer history, product information, and previous interactions enables more intelligent routing and response generation. Organizations implementing unified data architectures for customer service NLP report average handling time reductions of 25-35% and first-contact resolution improvements of 15-20% compared to previous approaches [7]. These efficiency gains translate directly to cost savings and improved customer satisfaction, with Net Promoter Scores typically increasing by 10-12 points after implementation of NLP-powered service enhancements.

The healthcare sector presents another compelling case for Lakehouse-powered NLP applications. Clinical documentation, medical literature, patient communications, and structured medical data must be analyzed together to extract maximum value from healthcare information. According to Gajavada's analysis, healthcare organizations implementing unified data architectures for clinical NLP report a 40-50% reduction in time required for documentation review and information extraction compared to manual approaches [8]. This efficiency improvement enables clinical staff to focus more on patient care and less on administrative tasks,

improving both operational metrics and care quality indicators. The ability to continuously update NLP models as new medical evidence emerges is particularly valuable in healthcare contexts, with unified data architectures reducing model update cycles from months to weeks or even days in some implementations.

Automated Decision-Making

Systems that make autonomous decisions require not only predictive models but also access to rules, constraints, and historical decisions—all of which can be integrated within a Lakehouse platform. The complexity of implementing automated decision systems highlights the value of unified data architectures, with organizations typically needing to integrate 7-10 distinct data sources to enable contextually appropriate automated decisions [7]. This integration challenge becomes particularly acute in real-time decision contexts like fraud detection or dynamic pricing, where decision quality depends on the ability to consider both historical patterns and immediate context.

Financial services organizations have been particularly aggressive in adopting Lakehouse architectures for automated decision systems, driven by the need to balance speed, accuracy, and compliance requirements. Fraud detection systems represent a high-value application, with unified data architectures enabling more sophisticated detection models that consider a broader range of signals. According to industry analysis, financial institutions implementing fraud detection on Lakehouse platforms report false positive reductions of 20-30% while maintaining or improving detection rates [8]. This improvement directly impacts operational efficiency by reducing the manual review burden while enhancing customer experience by minimizing legitimate transaction declines.

The governance capabilities of Lakehouse architecture provide particular value for regulated automated decision systems where transparency and auditability are essential. The integrated lineage tracking and governance features of modern Lakehouse platforms enable organizations to document the data sources, transformations, and models used for each decision, satisfying regulatory requirements without imposing significant operational overhead. NetApp's research indicates that organizations implementing automated decision systems within unified data architectures spend 40-50% less time on compliance documentation compared to those using fragmented architectures [7]. This efficiency gain is particularly valuable in highly regulated industries like banking and insurance, where compliance requirements can otherwise create significant barriers to automation.

Retail pricing optimization represents another high-value application of automated decision-making powered by Lakehouse architecture. The ability to combine historical sales data, inventory levels, competitive pricing information, and customer behavior patterns enables more sophisticated pricing strategies that maximize both revenue and margin. According to Gajavada, retailers implementing pricing optimization on unified data platforms report gross margin improvements of 1-3% and inventory turn increases of 10-15% compared to previous approaches [8]. These performance improvements translate directly to financial results, with typical ROI for unified pricing optimization initiatives exceeding 250% within the first year of implementation. The ability to rapidly adapt pricing strategies to changing market conditions becomes particularly valuable in volatile markets, with Lakehouse architecture enabling price update cycles to be reduced from days or weeks to hours or even minutes in some implementations.

The integration of Lakehouse technology in an AI-ready data strategy enables these applications to achieve their full potential by addressing fundamental data management challenges that have historically limited AI effectiveness and scalability. By providing a unified foundation for diverse AI workloads, Lakehouse architecture enables organizations to accelerate implementation, improve performance, and enhance governance across their AI application portfolio, creating sustainable competitive advantage through more effective use of their data assets.

Quality Dimension	Description	Impact on AI Models	Mitigation Strategy
Completeness	Presence of required data elements	Biased predictions due to missing values	Automated detection and imputation
Accuracy	Correctness of data values	Incorrect learning patterns	Reference data validation
Consistency	Uniformity across related data sets	Conflicting signals to models	Cross-dataset reconciliation
Timeliness	Recency of data relative to real world	Outdated predictions	Real-time data pipelines

Uniqueness	Absence of duplicates	Overweighting certain patterns	Deduplication processes
Validity	Conformance to defined formats	Processing errors	Schema enforcement
Integrity	Maintenance of relationships	Missed correlations	Relationship validation
Bias	Systematic skew in data	Unfair or discriminatory outcomes	Bias detection algorithms

Table 4: Data Quality Dimensions and Their Impact on AI Applications [7,8]

Challenges and Solutions

Implementing an AI-ready data strategy with Lakehouse technology is not without challenges. Organizations often encounter significant obstacles that can impede the success of their initiatives if not properly addressed. However, with appropriate solutions and approaches, these challenges can be overcome, enabling organizations to realize the full potential of their AI investments. Let's examine the most common challenges and their solutions in detail, supported by research and industry experience.

Challenge: Data Governance at Scale

As data volumes grow and more stakeholders leverage the data for AI initiatives, governance becomes increasingly complex. According to McKinsey's 2023 State of AI report, data governance remains one of the top obstacles to AI adoption, with nearly two-thirds of organizations (65%) citing it as a significant challenge [9]. The complexity increases as organizations scale AI from isolated use cases to enterprise-wide deployment, with governance requirements expanding from individual datasets to comprehensive data estates spanning multiple domains and systems. This expansion creates significant operational challenges, with organizations reporting that governance team capacity often fails to scale proportionally with data growth, creating bandwidth constraints that slow AI implementation.

The impact of governance challenges on AI initiatives is substantial, with organizations reporting that governance-related delays account for approximately 20-30% of total implementation time for AI projects [9]. These delays primarily stem from uncertainty about data usage rights, compliance requirements, and risk management considerations, particularly for sensitive data domains like customer information, financial data, and personal health information. The situation is further complicated in global organizations operating across multiple regulatory jurisdictions, with varying and sometimes conflicting requirements creating a complex compliance landscape. McKinsey's research indicates that high-performing AI organizations are twice as likely to have established robust data governance frameworks compared to their peers, highlighting the critical role of governance in enabling sustainable AI scaling.

Solution: Metadata Management

Comprehensive metadata management provides the foundation for effective governance at scale. By implementing metadata catalogs that serve as a single source of truth about data assets, organizations can establish clear visibility and control over their data landscape. According to research by QuantumBlack, organizations that implement comprehensive metadata management report significant improvements in data discovery and understanding, with data scientists saving an average of 15-20 hours per month previously spent searching for and understanding available data assets [10]. These efficiency gains translate directly to accelerated AI development cycles and more comprehensive model development, as data scientists can more readily identify and incorporate relevant datasets.

The implementation of clear data ownership and stewardship structures represents another critical component of metadata-driven governance. The QuantumBlack analysis indicates that organizations with well-defined ownership models experience 44% fewer data access issues compared to those with ambiguous ownership structures [10]. This improvement stems from clear decision-making authority and accountability for data access, quality, and security, creating a more streamlined governance process that balances innovation with appropriate controls. The most effective implementations assign ownership based on domain expertise rather than technical roles, ensuring that business context informs governance decisions and that controls align with actual risk profiles and business requirements.

Automated data classification and tagging capabilities have emerged as essential tools for scaling governance in complex data environments. McKinsey's research indicates that organizations leveraging AI/ML for automated classification can process new datasets approximately 7 times faster than manual approaches, with comparable or better accuracy for standard classification tasks [9]. This automation enables governance teams to maintain comprehensive coverage even as data volumes grow, ensuring that appropriate controls are applied consistently across the environment. The integration of automated classification with data

catalogs and access control systems creates a coherent governance ecosystem that scales more effectively than traditional manual approaches, addressing one of the fundamental challenges in scaling AI data governance.

Data lineage tracking represents a particularly valuable component of metadata management for AI governance, providing transparency into how data flows from source systems through transformations to eventual use in models and applications. According to the QuantumBlack analysis, organizations implementing comprehensive lineage tracking report a 50% reduction in time required to conduct impact analysis when systems or data structures change, enabling more agile evolution of data infrastructure [10]. For AI applications specifically, lineage capabilities provide essential information for model explanation and validation, supporting both technical development and regulatory compliance. This transparency becomes increasingly valuable as AI systems influence critical business decisions, creating accountability for both the models and their underlying data.

Challenge: Ensuring Data Quality

AI models are highly sensitive to data quality issues, making consistent quality crucial for successful implementations. McKinsey's 2023 State of AI report highlights that data quality remains one of the most persistent challenges in AI development, with 43% of respondents citing it as a significant barrier to value realization [9]. The impact of quality issues can be substantial, with research indicating that model performance can degrade by 10-25% when training data contains significant quality problems. This performance degradation directly impacts business outcomes, reducing the ROI of AI investments and potentially creating business risks if models make incorrect recommendations or predictions based on faulty data.

The data quality challenge grows more complex as organizations scale their AI initiatives from isolated use cases to enterprise-wide deployment. According to McKinsey's analysis, organizations implementing AI at scale typically manage 3-5 times more data sources than those focused on limited use cases, creating significantly greater quality management challenges [9]. Each additional source introduces potential quality issues and inconsistencies that must be detected and addressed to maintain model performance. The situation is further complicated by the need to maintain quality over time as source systems change, business definitions evolve, and new data is incorporated. This temporal dimension of data quality presents particular challenges for AI systems that may continue to use data long after its initial creation, requiring ongoing quality monitoring and management.

Solution: Automated Data Quality Checks

Implementing continuous data quality monitoring represents the foundation of effective quality management for AI data. According to QuantumBlack's research, organizations implementing automated monitoring detect data quality issues an average of 14 days earlier than those relying on periodic manual reviews, enabling faster remediation and reducing the impact on downstream applications [10]. This proactive approach is particularly valuable for AI applications, where early detection of quality issues can prevent model degradation and maintain prediction accuracy. The most effective implementations incorporate both technical quality metrics (completeness, format consistency, range validation) and business-oriented measures (accuracy, timeliness, relevance), creating a comprehensive view of quality that supports both operational needs and business objectives.

Establishing clear data quality SLAs and metrics provides the framework for ongoing quality management and improvement. McKinsey's analysis indicates that organizations with explicit quality targets are 2.5 times more likely to achieve consistent data quality across different domains compared to those with ad-hoc approaches [9]. These metrics typically span multiple dimensions including completeness, accuracy, timeliness, and consistency, with thresholds tailored to the specific requirements of each data domain and application. By incorporating quality metrics into performance management and monitoring systems, organizations create accountability for quality outcomes and enable continuous improvement over time. This structured approach is particularly valuable for AI applications, where quality requirements may be more stringent than for traditional analytics due to the automated nature of AI decision-making.

Automated remediation workflows for common quality issues represent a critical capability for maintaining quality at scale. According to QuantumBlack's analysis, organizations implementing automated remediation resolve routine quality issues approximately 4 times faster than manual approaches, with an average reduction from 6 days to 1.5 days for standardized issue types [10]. This dramatic improvement in resolution time reduces the duration of quality-related disruptions to AI applications, maintaining model performance even when upstream issues occur. The most sophisticated implementations incorporate machine learning to identify patterns in quality issues, enabling increasingly autonomous remediation that requires human intervention only for novel or complex problems. This approach allows quality management to scale efficiently even as data volumes and complexity increase.

Creating feedback loops between AI teams and data stewards completes the quality management system by connecting those who consume data with those responsible for its quality. McKinsey's research indicates that organizations implementing structured feedback mechanisms experience a 40% reduction in recurring quality issues compared to those with siloed quality management approaches [9]. These improvements stem from better alignment between quality standards and actual application requirements, ensuring that quality efforts focus on the dimensions most critical to downstream usage. The feedback mechanism also provides

valuable insights about the business impact of quality issues, enabling more informed prioritization of quality improvement initiatives based on their potential value. This value-oriented approach ensures that quality investments deliver maximum return by focusing on the issues with greatest business relevance.

Challenge: Infrastructure Scalability and Cost Management

AI workloads can be unpredictable and resource-intensive, leading to significant scalability and cost challenges for organizations implementing AI at scale. According to McKinsey's 2023 State of AI report, organizations are increasingly focused on the economics of AI, with 38% of respondents citing infrastructure cost as a significant concern [9]. This focus on economics has intensified as organizations move from experimental deployments to production implementations, where inefficient infrastructure can create substantial ongoing costs that erode the business case for AI adoption. The situation is particularly challenging for organizations transitioning from pilot projects to enterprise-wide deployment, where infrastructure requirements can grow by orders of magnitude within relatively short timeframes.

The computational intensity of AI workloads creates unique infrastructure challenges compared to traditional analytics. McKinsey's analysis indicates that AI training workloads typically require 2-5 times more computational resources than traditional analytics, with the differential increasing for more sophisticated models and complex data types [10]. This intensity is further complicated by the unpredictable nature of many AI workloads, particularly during development and experimentation phases where resource requirements may vary significantly as models are refined and approaches are tested. Organizations report that their peak AI infrastructure demand typically exceeds average usage by a factor of 3-4x, creating a difficult tradeoff between provisioning for peak demand (resulting in low average utilization) or optimizing for average usage (creating capacity constraints during peak periods).

Solution: Scalable Compute and Storage

Leveraging cloud-native infrastructure for dynamic scaling provides the foundation for effective AI infrastructure management. According to QuantumBlack's research, organizations implementing cloud-based AI infrastructure achieve 35% higher resource utilization compared to traditional on-premises approaches with fixed capacity [10]. These improvements stem primarily from the ability to scale resources up and down based on actual demand, paying only for resources actually used rather than maintaining capacity for peak loads. The elasticity of cloud infrastructure is particularly valuable for organizations with variable AI workloads, allowing them to accommodate demand spikes without overprovisioning for normal operations. This flexibility creates both cost advantages and performance benefits, as organizations can provision higher-capacity resources when needed without long-term commitment.

Implementing tiered storage strategies to balance performance and cost represents another critical aspect of infrastructure optimization. McKinsey's analysis indicates that organizations using tiered approaches reduce storage costs by 30-50% compared to single-tier strategies while maintaining performance for active workloads [9]. The most effective implementations typically include high-performance storage for active datasets and frequently accessed files, with automated movement to lower-cost tiers for historical data and infrequently used assets. This approach is particularly valuable for AI applications, which often involve a combination of large historical datasets for training and smaller, more actively used datasets for inference and ongoing refinement. By aligning storage performance with actual usage patterns, organizations can optimize both cost and performance across the AI application lifecycle.

Intelligent caching mechanisms have emerged as a powerful tool for optimizing AI infrastructure, particularly for workloads that repeatedly access the same datasets. According to QuantumBlack's analysis, organizations implementing advanced caching achieve performance improvements of 2-3x for common AI workloads while reducing storage I/O costs by approximately 40% [10]. These benefits are particularly significant for iterative model training processes, where the same training data may be accessed repeatedly during hyperparameter tuning and model refinement. By keeping frequently accessed data in high-performance cache, organizations can achieve both performance and cost objectives without requiring all data to reside in expensive high-performance storage tiers. This balanced approach creates a cost-effective infrastructure environment that supports both development agility and production efficiency.

Deploying comprehensive workload management and prioritization capabilities completes the infrastructure optimization approach by ensuring that available resources are allocated to the most valuable workloads. McKinsey's research indicates that organizations implementing structured prioritization improve time-to-completion for high-priority AI workloads by approximately 40% while maintaining reasonable performance for lower-priority tasks [9]. This improvement is achieved through a combination of technical policies (resource allocation, preemption, queueing) and organizational processes (clear prioritization frameworks, escalation paths). The resulting system ensures that critical business initiatives receive appropriate resources while maintaining efficient utilization of the overall infrastructure, optimizing both performance and cost across the AI portfolio. This balanced

approach is particularly valuable as organizations scale AI adoption, with growing workloads competing for limited infrastructure resources.

Challenge: Integration with AI/ML Tools

Data scientists and ML engineers use diverse tools and frameworks that must work seamlessly with the data platform. This diversity creates significant integration challenges, with McKinsey's research indicating that the average AI organization uses 5-7 distinct tools and frameworks across their AI initiatives [9]. The situation is further complicated by the rapid evolution of the AI/ML tooling landscape, with new frameworks and libraries emerging regularly and existing tools frequently releasing updates with breaking changes. This fragmentation creates significant integration challenges for data platform teams seeking to provide consistent, reliable access to enterprise data assets across the diverse toolset used by AI practitioners. Organizations report that integration issues between data platforms and AI tools represent a significant source of friction in the development process, slowing innovation and reducing productivity.

The complexity of tool integration increases with the diversity of data sources and types required for AI development. Different frameworks often have specific expectations regarding data formats, access patterns, and performance characteristics, creating a challenging environment for data platform teams seeking to support multiple tools efficiently. According to QuantumBlack's analysis, data scientists spend approximately 30% of their time on data preparation and transformation to meet tool-specific requirements [10]. This significant time investment represents a drag on productivity that could be alleviated through better integration between data platforms and AI tools. The impact is particularly acute for organizations early in their AI journey, where immature integration often creates additional friction in an already challenging process of building initial AI capabilities.

Solution: Integrated AI/ML Platforms

Ensuring compatibility with popular frameworks like TensorFlow, PyTorch, and scikit-learn represents the foundation of effective tool integration. According to McKinsey's research, organizations implementing standardized data access interfaces that support multiple frameworks reduce framework-specific data preparation time by approximately 50% [9]. These improvements stem from the elimination of custom integration code and manual data transformations that would otherwise be required to bridge gaps between the data platform and specific frameworks. The most effective implementations provide both high-level abstractions for common patterns and low-level access for specialized requirements, ensuring that the integration layer supports both productivity and flexibility. This balanced approach enables data scientists to use their preferred tools while maintaining consistent access to enterprise data assets, combining flexibility with governance and security.

Providing native support for notebooks and development environments creates a seamless experience for data scientists and ML engineers, allowing them to access and manipulate data directly from their preferred tools. QuantumBlack's analysis indicates that organizations implementing integrated development environments reduce environment setup and configuration time by approximately 25%, allowing more focus on actual analysis and model development [10]. This productivity improvement is particularly valuable for exploratory analysis and initial model development, where rapid iteration is essential for finding promising approaches. By eliminating friction in the development process, integrated environments enable faster innovation and more thorough exploration of the solution space, ultimately leading to better models and insights. The most effective implementations combine self-service capabilities with appropriate governance guardrails, enabling productivity while maintaining compliance with security and regulatory requirements.

Implementing feature stores to standardize model inputs represents a particularly valuable integration capability for organizations building multiple AI models. According to McKinsey's research, organizations using centralized feature stores reduce feature development time by approximately 40% and improve feature reuse across projects by a factor of 2-3x [9]. These efficiency gains stem from the elimination of redundant feature development and the establishment of standardized, high-quality features that can be used across multiple models. The centralized approach also improves model consistency by ensuring that the same feature definitions are used in different contexts, reducing the risk of subtle differences that can lead to unexpected model behavior. Feature stores are particularly valuable as organizations scale AI adoption, with the efficiency benefits increasing as more models leverage the shared feature repository.

Creating comprehensive model registries and versioning systems completes the integration approach by providing centralized tracking and management of AI assets throughout their lifecycle. QuantumBlack's analysis indicates that organizations implementing model registries reduce time spent on model governance and compliance activities by approximately 35% compared to those using ad-hoc tracking approaches [10]. These improvements are particularly valuable for regulated industries where model documentation and validation are critical compliance requirements. The centralized approach also facilitates collaboration and knowledge sharing across AI teams, with clear visibility into existing models and their performance characteristics supporting better reuse and iterative improvement rather than redundant development. As AI adoption scales across the enterprise, these governance capabilities become increasingly essential for maintaining control and transparency across a growing portfolio of models and applications.

By addressing these four fundamental challenges with comprehensive, systematic approaches, organizations can overcome the most common barriers to successful implementation of AI-ready data strategies using Lakehouse technology. The solutions described create a foundation for sustainable scaling of AI capabilities, enabling organizations to move beyond initial pilots to enterprise-wide adoption that delivers substantial business value. While the specific implementation details will vary based on organizational context, industry requirements, and existing technology landscapes, the core principles apply broadly across sectors and use cases, providing a roadmap for successful AI data strategy implementation.

Conclusion

Lakehouse technology fundamentally transforms how organizations approach their data infrastructure for AI initiatives by eliminating silos between data lakes and warehouses. This unified approach addresses critical challenges in data accessibility, quality, governance, and scalability that have historically limited AI effectiveness. The convergence of storage flexibility and processing reliability creates an environment where data scientists can focus more on model development and less on data preparation, accelerating time-to-value for AI projects. Organizations implementing Lakehouse architectures benefit from improved collaboration between data engineering and data science teams, more efficient resource utilization, and enhanced governance capabilities. As AI becomes increasingly essential for competitive advantage, the integrated nature of Lakehouse solutions enables enterprises to deploy sophisticated applications across domains with greater agility and lower operational overhead. The technology continues to evolve to address emerging requirements like edge computing integration and hybrid processing models, positioning Lakehouse architecture as a cornerstone of forward-looking data strategies that can adapt to the rapidly changing AI landscape.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Arun Gururajan, "Unified Data Architectures for AI Workflows," NetApp, 2023. [Online]. Available: https://www.netapp.com/blog/unified-data-architectures-for-ai-workflows/
[2] Blaž Čuš, et al., "Data Lakehouse Benefits in small and medium enterprises," ResearchGate, Apr. 2023. [Online]. Available: https://www.researchgate.net/publication/370187571_Data_Lakehouse_Benefits_in_small_and_medium_enterprises
[3] David Reinsel, et al, "The Digitization of the World: From Edge to Core," Seagate, Nov. 2018. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf
[4] Kiran Kumar Gajavada, "Winning with Data: The Ultimate Business Case for Lakehouses," LinkedIn, 2025. [Online]. Available: https://www.linkedin.com/pulse/winning-data-ultimate-business-case-lakehouses-gajavada-he-him-his--gjcrc/
[5] McKinsey & Company, "The State of AI in 2023: Generative AI's Breakout Year," 2023. [Online]. Available: https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%202023%20generative%20ais%20breakout%20year/the-state-of-ai-in-2023-generative-ais-breakout-year_vf.pdf
[6] Michael Armbrust, et al., "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics," CIDR, 2021. [Online]. Available: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
[7] Nutanix, "Getting the Most Out of AI-Ready Infrastructure,". [Online]. Available: https://www.nutanix.com/how-to/getting-the-most-out-of-ai-ready-infrastructure
[8] QuantumBlack, "TThe state of AI in 2023: Generative AI's breakout year", Medium, 2024. [Online]. Available: https://medium.com/quantumblack/the-challenges-of-scaling-for-ai-transformation-94d9d88d2075
[9] Sandeep Kaushik, "Data Lakehouse: A Modern Data Architecture," Medium, 2025. [Online]. Available: https://medium.com/@shyamsandeep28/data-lakehouse-a-modern-data-architecture-3dd3e1c89f92
[10] Trinh Nguyen, "A Guide to Data Preparation for AI/Machine Learning System," Neurond, 2023. [Online]. Available: https://www.neurond.com/blog/ai-data-preparation