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

AI-Driven Data Mesh with AutoML for Enterprise Analytics

Lingareddy Alva
IT Spin Inc, USA
Corresponding Author: Lingareddy Alva, **E-mail**: alvalingareddy@gmail.com

| ABSTRACT
This article explores the transformative potential of AI-driven Data Mesh architectures for enterprise analytics. By reimagining traditional centralized data structures through domain-driven ownership principles, organizations can achieve unprecedented scalability and agility. The implementation leverages Databricks Unity Catalog and Delta Sharing for federated governance while maintaining domain autonomy. At its core, an AutoML-powered Data Quality Engine ensures data integrity through machine learning capabilities that detect anomalies, impute missing values, and generate explainable reports. Event-driven pipelines built with Apache Kafka and Delta Live Tables enable real-time insights and forecasting, allowing businesses to respond immediately to changing conditions. This architectural paradigm empowers enterprises to move beyond static reporting toward autonomous data-driven operations with intelligent insights and seamless cross-domain collaboration.

| KEYWORDS

Data Mesh, AutoML, Real-time Analytics, Domain-driven Architecture, Data Quality

| ARTICLE INFORMATION

ACCEPTED: 12 April 2025 **PUBLISHED:** 20 May 2025 **DOI:** 10.32996/jcsts.2025.7.4.87

1. Introduction

In today's data-driven business landscape, traditional centralized data architectures have become a significant bottleneck, limiting organizational agility, scalability, and innovation. Recent research indicates that this centralization problem compounds as data volumes grow at rates of 25-30% annually across enterprise systems [1]. The emergence of edge computing, with processing occurring at over 15 billion IoT devices by 2025, further demonstrates why funneling all information through a central data team creates unsustainable bottlenecks in modern enterprises.

As organizations generate exponentially growing volumes of data across disparate systems, decentralized architectures have become necessary. The conventional centralized approach leads to average analytics backlogs of 3-4 months, with 65% of data projects failing to deliver business value on time. This article explores how an AI-driven Data Mesh architecture can transform enterprise analytics by decentralizing data ownership while maintaining governance, leveraging cutting-edge technologies including Apache Spark, Databricks, and cloud-native services.

The Data Mesh paradigm aligns with modern distributed systems trends where computation and data storage are increasingly moved closer to data sources. As noted in research on federated data architectures, organizations implementing domain-driven ownership models have seen a 40-60% reduction in time-to-insight for analytics initiatives [2]. By treating data as a product and applying domain-oriented ownership principles, enterprises can achieve substantially improved data quality metrics while supporting containerized microservices architectures that process up to 2.5 million events per day in high-throughput environments.

2. The Limitations of Traditional Data Architectures

Traditional centralized data architectures often struggle with scaling to meet modern enterprise needs. These architectures typically rely on monolithic data warehouses or lakes managed by a single team, creating dependencies that slow down innovation and time-to-insight. Research indicates that approximately 80% of enterprises still operate with centralized data control models despite their inherent limitations, resulting in significant operational inefficiencies [3]. The centralized approach creates what experts call "organizational silos," where decision-making becomes concentrated among a small group of data gatekeepers, frequently causing delays of 6-8 weeks for even modest analytics requests.

Domain teams must wait for data engineering resources, leading to project backlogs and outdated analytics. According to enterprise architecture studies, organizations with strictly centralized data architectures experience 42% longer time-to-market for new data products compared to those with distributed models [3]. This delay translates to tangible business impacts, with nearly 67% of business leaders reporting that centralized bottlenecks directly impede strategic initiatives. Moreover, when decisions must pass through central approval channels, organizations report a 35% decrease in responsiveness to changing market conditions, particularly problematic in volatile industries.

Furthermore, centralized approaches frequently result in a disconnect between data producers and consumers, where business context is lost in translation, and domain expertise isn't effectively incorporated into data models. The architectural complexity compounds this problem—traditional enterprise data architectures typically contain between 6-8 layers of processing between source systems and business consumption [4]. This complexity is not merely theoretical; it creates practical challenges where approximately 73% of business analysts report spending over 60% of their time translating between technical data structures and business requirements.

Traditional architectures also face significant technical limitations. The average enterprise data warehouse supports only 240-300 concurrent queries before performance degradation occurs, creating bottlenecks during peak business hours [4]. Additionally, these systems typically require 3-5 times more maintenance resources than modern distributed architectures, with an average of 54% of the data engineering budget allocated to maintaining existing pipelines rather than creating new capabilities. The monolithic nature of centralized systems means that changes in one data domain frequently impact others, with organizations reporting that 40% of data pipeline modifications result in unintended downstream consequences requiring additional remediation.

Metric	Centralized Architecture
Enterprises Using This Model	80%
Business Leaders Reporting Strategic Impediments	67%
Responsiveness to Market Changes (Relative)	65%
Processing Layers Between Source and Consumption	6-8
Analysts' Time Spent on Data Translation	60%
Maximum Concurrent Queries Before Degradation	240-300

Data Engineering Budget for Maintaining Existing Pipelines	54%
Pipeline Modifications Causing Unintended Consequences	40%
Analytics Request Processing Time	6-8 weeks

Table 1: Key Operational Metrics of Enterprise Data Architecture Models [3, 4]

3. Data Mesh: A Paradigm Shift in Enterprise Data Management

The Data Mesh paradigm, first introduced by Zhamak Dehghani in 2019, reimagines data architecture by treating data as a product and assigning ownership to domain teams. This revolutionary approach addresses the limitations of centralized data platforms, which according to AWS research, experience a 70% failure rate when scaling beyond certain thresholds [5]. The core principles of Data Mesh—domain ownership, data as a product, self-serve infrastructure, and federated governance—create a framework that has demonstrated remarkable improvements in enterprise data usability.

Unlike centralized approaches, Data Mesh applies domain-driven design principles to analytical data, positioning domain teams as both producers and consumers of data products. This structural alignment creates what AWS refers to as "data domains"—autonomous units that own both operational and analytical data within their business contexts. Organizations implementing this approach have reported that domain teams can develop and deploy new data products in 30-40% less time compared to traditional centralized models, primarily due to reduced coordination overhead and removal of cross-team dependencies [5].

This approach aligns with modern microservices architectures and agile methodologies, enabling faster innovation cycles. As demonstrated in recent industrial case studies, companies implementing Data Mesh architectures alongside microservices have achieved a 28% improvement in time-to-market for new data-driven features [6]. The architectural alignment is particularly beneficial for complex industrial platforms, where the traditional separation between operational and analytical systems has historically created significant data integration challenges.

By decentralizing data ownership, organizations can scale their data capabilities horizontally across the enterprise while maintaining cohesion through standardized protocols and governance frameworks. Recent research examining industrial implementations found that Data Mesh adopters processed an average of 3.5 times more data sources in their analytical ecosystems without proportional increases in complexity or governance costs [6]. This scalability advantage becomes particularly evident in manufacturing environments, where a single automotive manufacturer successfully integrated 237 distinct sensor data streams from 42 production systems using a domain-oriented data platform—a task that had previously failed under centralized data lake architectures.

The economic benefits are equally compelling, with organizations reporting 45-55% reductions in overall time spent on data integration activities and a 60% decrease in the number of data pipeline failures following Data Mesh adoption. Furthermore, with self-service capabilities properly implemented, 83% of business users reported increased data accessibility, while data engineering teams reclaimed an average of 37% of their capacity for innovation rather than maintenance tasks [5].

4. Implementing an AI-Driven Data Mesh with Databricks

A truly effective Data Mesh implementation requires sophisticated tools that balance domain autonomy with enterprise interoperability. Modern data architectures increasingly adopt hybrid approaches, with 76% of enterprises now employing some combination of cloud, on-premises, and edge computing resources to address their diverse analytical needs [7]. Databricks provides an ideal foundation through its Unity Catalog and Delta Sharing capabilities, offering the flexibility required for these multi-environment deployments.

The Unity Catalog enables federated governance with fine-grained access controls that respect domain boundaries while ensuring compliance with organizational policies. This capability addresses a critical challenge in modern architectures, where data frequently exists in multiple locations—with the typical enterprise utilizing three to five different storage platforms across cloud and on-premises environments [7]. By centralizing metadata management while allowing data to remain distributed, organizations can achieve what industry experts call "governance without borders," essential for maintaining compliance across complex multi-cloud and hybrid architectures.

Delta Sharing facilitates secure cross-domain data exchange using an open protocol, allowing teams to publish and consume data products seamlessly. This protocol enables a streamlined workflow where providers create shares with specific tables or views, generate activation links containing secure credentials, and allow recipients to access data directly through familiar tools and frameworks [8]. The innovation of Delta Sharing lies in its ability to maintain a clear separation between storage and compute, with recipients accessing only the data they're authorized to view without requiring specialized infrastructure or complex extract-transform-load pipelines.

The implementation architecture includes domain-specific Databricks workspaces for isolated development, where each domain can operate autonomously while adhering to enterprise standards. Unity Catalog provides unified metadata management across domains, creating a consistent governance framework despite the decentralized operational model. Delta Lake tables serve as the foundation for high-quality, versioned data products, offering ACID transaction guarantees and time travel capabilities essential for maintaining data lineage. Databricks SQL enables analytics workloads with built-in query federation across distributed data assets. Finally, Delta Sharing establishes secure internal and external data exchange, supporting open-standard formats like Apache Parquet that ensure compatibility with virtually any modern analytics tool or platform [8].

This architectural approach aligns perfectly with the requirements of data mesh implementations, where domain autonomy must be balanced with enterprise interoperability to deliver meaningful business value while maintaining governance and security standards.

5. AutoML-Powered Data Quality Engine

At the heart of the AI-driven Data Mesh is an AutoML-powered Data Quality Engine that ensures high data integrity across all domain products. Recent research demonstrates that automated data quality management systems can reduce error rates by up to 87% compared to manual processes, while simultaneously cutting assessment time by approximately 76% [9]. This efficiency is critical as organizations struggle with growing data volumes—the average enterprise now manages over 10 terabytes of data per domain, making manual quality assurance increasingly impractical.

Built on Spark ML and MLflow, this engine employs machine learning to automatically detect anomalies and outliers using unsupervised learning algorithms. These techniques have proven highly effective, with isolation forest algorithms demonstrating precision rates of 92.3% in identifying anomalous patterns across diverse datasets [9]. The engine intelligently imputes missing values based on data patterns and relationships, addressing a pervasive challenge in enterprise data landscapes where missing data rates typically range from 15-30% in operational systems.

The engine generates data drift reports to identify changing data characteristics over time, leveraging statistical methods that can detect distribution shifts with 94% accuracy. This capability is particularly valuable as research shows that approximately 52% of production machine learning models experience significant performance degradation within six months due to undetected data drift [9]. By proactively identifying these changes, organizations can maintain analytical integrity even as underlying data patterns evolve.

Furthermore, the system creates explainable quality metrics accessible to business users. Research indicates that contextual, interpretable quality indicators increase stakeholder trust by 63% compared to technical metrics, leading to significantly higher adoption rates for data-driven decision making [10]. This approach transforms quality management from a technical exercise to a business enabler that directly supports organizational objectives.

The quality engine runs as scheduled jobs within each domain's Databricks environment, with results published to a central quality catalog that provides transparency across the organization. This distributed yet centrally visible approach has been shown to reduce data quality management overhead by approximately 42% while increasing cross-domain data reuse by 57% [10]. Additionally, automation in data quality measurement enables organizations to efficiently monitor 4-5 times more data assets without increasing staff, supporting critical use cases like customer sentiment analysis and personalized experiences.

This architectural approach enables domain teams to maintain quality standards autonomously while providing enterprise-wide visibility into data health, creating what researchers describe as "democratized quality management"—a critical enabler for successful Data Mesh implementations at scale.

Metric	Traditional/Manual Approach	AutoML-Powered Approach
Error Rate Reduction	Baseline	87% reduction
Assessment Time		76% reduction
Anomaly Detection Precision	Comparison baseline	92.3%
Missing Data in Operational Systems	15-30%	Addressed via imputation
Data Drift Detection Accuracy	Comparison baseline	94%
ML Models with Performance Degradation (within 6 months)	52%	Reduced via proactive detection
Stakeholder Trust (with interpretable metrics)	Baseline	63% increase
Data Quality Management Overhead		42% reduction
Cross-Domain Data Reuse		57% increase
Data Assets Monitored (with same staff)		4-5x more

Table 2: Impact Metrics of AI-Driven Data Quality Automation in Enterprise Environments [9, 10]

6. Event-Driven AI Analytics with Kafka and Delta Live Tables

To enable real-time insights, the architecture incorporates event-driven AI analytics pipelines using Apache Kafka and Databricks Delta Live Tables (DLT). A comprehensive analysis of event-driven business process management reveals that organizations implementing event-driven architectures achieve 60-70% faster response times to critical business events compared to traditional approaches [11]. This improvement enables what researchers refer to as "real-time enterprise capabilities," where operational decisions are made within the same timeframe as the events themselves.

Apache Kafka serves as the enterprise event backbone, capturing real-time signals from operational systems. Research indicates that event-driven architectures built on robust event brokers like Kafka have achieved 85% higher business process monitoring effectiveness and 63% improved exception handling compared to traditional pull-based integration approaches [11]. The architectural advantages are particularly evident in complex enterprise environments, where systems must coordinate across dozens or hundreds of applications operating at different tempos and with varying availability patterns.

Databricks Delta Live Tables provide declarative pipeline definitions with built-in quality controls, dramatically simplifying the development and maintenance of data pipelines. DLT enables developers to focus on transformations rather than framework code, reducing implementation time by up to 80% compared to traditional pipeline development [12]. The platform's integrated data quality capabilities, including schema enforcement, data validation, and constraint checking, ensure that faulty data is identified and handled before it impacts downstream analytics, addressing what practitioners identify as the most significant challenge in streaming data processing.

Streaming analytics jobs process events in real-time, updating analytical models and dashboards. According to industry analysis, approximately 82% of organizations consider real-time monitoring and analytics to be critical or very important to their business operations, yet only 31% have successfully implemented such capabilities due to technical challenges [11]. The combination of Kafka's reliable event delivery with DLT's declarative processing model bridges this implementation gap, enabling organizations to achieve what was previously beyond technical reach for many teams.

AI-based forecasting models deployed through Databricks Model Serving make predictions on streaming data. This architecture creates a continuous intelligence loop where data flows directly from operational systems through analytical pipelines to predictive models and back to operational systems. The continuous flow enables organizations to reduce data-to-decision latency by approximately 76%, creating what researchers describe as a "sense-and-respond" capability essential for digital business operations [11].

This event-driven approach enables organizations to monitor customer behavior, track inventory levels, and optimize operations with minimal latency, shifting from reactive to proactive operational models that deliver measurable competitive advantages in today's fast-moving markets.

Metric	Event-Driven Approach	Improvement
Response Time to Critical Business Events	60-70% faster	60-70%
Business Process Monitoring Effectiveness	85% higher	85%
Exception Handling Efficiency	63% improved	63%
Pipeline Implementation Time	80% reduction	80%
Data-to-Decision Latency	76% reduction	76%

Table 3: Comparative Metrics of Traditional vs. Event-Driven Analytics Architectures [11, 12]

7. Conclusion

The AI-augmented Data Mesh represents a fundamental shift in how enterprises manage, process, and derive value from their data assets. By combining decentralized domain ownership with sophisticated AI capabilities, organizations overcome the limitations of traditional architectures while embracing future-oriented analytics practices. The integration of AutoML for data quality management ensures the integrity of domain-specific data products, while event-driven processing enables real-time decision intelligence across the enterprise. As data volumes continue growing and business requirements evolve, this architecture provides a sustainable framework for turning data into actionable insights. Through technologies like Databricks, Apache Spark, MLflow, and cloud-native services, the AI-driven Data Mesh creates a resilient ecosystem that scales with organizational needs while continuously delivering value to stakeholders throughout the enterprise.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Amazon Web Services, "What is Data Mesh?," AWS, 2025. [Online]. Available: https://aws.amazon.com/what-is/data-mesh/
[2] Andre Ripla PgCert, "The Risks of Centralized IT Control in Large Enterprises," LinkedIn, 2025. [Online]. Available: https://www.linkedin.com/pulse/risks-centralized-control-large-enterprises-andre-ripla-pgcert-pgdip-wmp6e
[3] Ben Aurel Schill, "Exploring Data Mesh Architecture: A Comparative Study of Implementation Archetypes Across Different Sectors and Industries," Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2025, 2023. [Online]. Available: https://dl.gi.de/server/api/core/bitstreams/6c6a48a2-b7b9-46cb-88b7-847d778844fb/content
[4] Data Ideology, "Enterprise Data Architecture Overview," DataIdeology, 2025. [Online]. Available: https://www.dataideology.com/enterprise-data-architecture-overview/
[5] Databricks, "Share data using the Delta Sharing open sharing protocol (for providers)," Databricks, 2025. [Online]. Available: https://docs.databricks.com/aws/en/delta-sharing/share-data-open
[6] Geeksforgeeks, "Latest Trends in Distributed Systems," GeeksforGeeks, 2024. [Online]. Available: https://www.geeksforgeeks.org/latest-trends-in-distributed-systems/
[7] Julian Krumeich et al., "Event-Driven Business Process Management: where are we now? A comprehensive synthesis and analysis of literature," ResearchGate, 2014. [Online]. Available: https://www.researchgate.net/publication/265856985_Event-Driven_Business_Process_Management_where_are_we_now_A_comprehensive_synthesis_and_analysis_of_literature
[8] Mikolaj Klepacz, "Create declarative ETL pipelines in Databricks with Delta Live Tables," DSStream, 2025. [Online]. Available: https://www.dsstream.com/post/create-declarative-etl-pipelines-in-databricks-with-delta-live-tables
[9] Philipp Kernstock et al., "Data Mesh -A Case Study Perspective On Building Industrial Data Platforms," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/379839899_DATA_MESH_-A_CASE_STUDY_PERSPECTIVE_ON_BUILDING_INDUSTRIAL_DATA_PLATFORMS
[10] Praneeth Reddy Amudala Puchakayala, "Data Quality Management for Effective Machine Learning and AI Modelling, Best Practices and Emerging Trends," International Research Journal of Innovations in Engineering and Technology, 2022. [Online]. Available: https://www.researchgate.net/publication/386230230_Data_Quality_Management_for_Effective_Machine_Learning_and_AI_Modelling_Best_Practices_and_Emerging_Trends
[11] Sidd TUMKUR, "Modern Data Architecture: Embracing the Cloud, Edge, and Hybrid Models," LinkedIn, 2024. [Online]. Available: https://www.linkedin.com/pulse/modern-data-architecture-embracing-cloud-edge-hybrid-models-tumkur-fe4ne
[12] Thu Nguyen, Hong-Tri Nguyen, Tu-Anh Nguyen-Hoang, "Scaling data quality monitoring for distributed intelligent systems: Challenges and solutions," Journal of Parallel and Distributed Computing, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731525000346