FROM PIPELINES TO PLATFORMS: THE IMPACT OF PLATFORM-CENTRIC DATA ARCHITECTURE ON ENTERPRISE AI IN REGULATED INDUSTRIES

Mohammed Arbaaz Shareef

Lead Data Engineer at Anblicks

ABSTRACT

The article discusses the transition from pipeline-oriented data engineering to platform-centric data management as enterprises scale AI operations under regulatory constraints. The practical motivation stems from the rising complexity of heterogeneous sources, the growing demand for traceability, and operational friction caused by isolated ETL pipelines that fragment quality controls and accountability. Scientific novelty lies in aligning platform concepts (lakehouse unification, data products, mesh/fabric logics, and governed feature stores) with governance requirements for auditability and trustworthy AI lifecycle oversight. The article aims to analytically substantiate how platform-centric data design affects delivery speed, reuse economics, and compliance posture compared to pipeline-first operating models. The study relies on a comparative synthesis of recent peer-reviewed literature on open data platforms, data products, data fabric paradigms, and feature store architectures, complemented by governance research on data quality management and responsible AI governance. The conclusion formulates structured implications for regulated enterprises designing AI-ready data foundations.

KEYWORDS

platform-centric data architecture, data platform, lakehouse, data products, data mesh

1. INTRODUCTION

Regulated enterprises increasingly face a dual pressure: accelerated adoption of AI for decision automation and heightened scrutiny over the provenance, privacy, and reproducibility of data transformations. Pipeline-oriented delivery, built as sequences of ETL jobs and point solutions, often scales in volume but not in accountability: lineage becomes fragmented, quality gates become inconsistent across teams, and operational ownership blurs when multiple pipelines replicate similar transformations. Platform-centric data management proposes a different operating logic: shared data services, standardized interfaces, and productized data assets that remain traceable across analytical and AI workloads.

This article aims to conceptually examine how the transition from pipeline-centric to platform-centric data architecture reshapes the execution, governance, and oversight of enterprise AI in regulated environments. To achieve this aim, the study systematizes architectural drivers behind platform adoption, identifies platform capabilities that enable reuse and time-consistent AI data, and derives governance implications for auditability, data quality discipline, and responsible AI oversight.

To make the analytical trajectory explicit, the study addresses three research questions.

1) Which architectural limitations of pipeline-centric delivery become most consequential for enterprise AI in regulated industries?
2) Which platform-centric capabilities most directly improve traceability, reuse, and reproducibility across analytical and ML workloads?
3) How does relocating control mechanisms from isolated pipelines to shared platform services affect the compliance posture and responsible AI oversight?

In line with these questions, the objectives of the article are to systematize the drivers of platform adoption, compare pipeline-centric and platform-centric operating logics, and derive governance implications for regulated enterprise environments.

The scientific novelty of this work lies in the integrated alignment of platform-centric data mechanisms with governance objectives specific to regulated industries. Unlike prior studies that examine data platforms or AI governance in isolation, this article synthesizes architectural and governance perspectives to position the data platform as an operational substrate for trustworthy and auditable enterprise AI.

2. MATERIALS AND METHODS

2.1. Materials

The material base consists of recent peer-reviewed studies selected for their direct relevance to platform architecture, governance instrumentation, and ML data lifecycle control. The source set was constructed to cover four analytical clusters: unifying storage and processing architectures, platform coordination models, governed ML data services, and governance frameworks for data and AI oversight. This composition enabled comparison not of isolated tools but of alternative architectural logics for enterprise AI delivery in regulated settings.

The empirical base comprises peer-reviewed sources that describe platform and governance patterns. M. Armbrust [1] proposes the lakehouse paradigm to unify analytics and ML across open formats. B. M. V. Bernardo [2] synthesizes data governance and data quality management frameworks relevant for enterprise control systems. I. Blohm [3] conceptualizes the relationships among data products, the data mesh, and the data fabric as competing and complementary management logics. P. K. Donta [4] frames data fabric as an adaptive data layer across distributed environments. A. Gieß [5] offers a structured delimitation between data platforms, data spaces, data meshes, and data fabrics, clarifying governance modes. R. Liu [6] analyzes feature stores as “DBMS-for-ML” and studies optimization and point-in-time correctness in training pipelines. J. de la Rúa Martínez [7] presents feature store architecture as a shared data layer linking feature, training, and inference pipelines. A. Nambiar [8] reviews differences between data warehouses and data lakes, clarifying the limitations of earlier centralized paradigms. A. Nizamis [9] operationalizes “data-as-a-product” in industrial value networks, emphasizing quality, security, trust, and traceability services. E. Papagiannidis [10] develops a research framework for responsible AI governance that connects structural and procedural practices with organizational outcomes. R. Eichler [11] develops the concept of an enterprise data marketplace as a metadata-driven platform that extends data discovery with access management and self-service provisioning. X. Ye [12] examines privacy and personal data risk governance in generative AI, highlighting the growing significance of rights-sensitive control mechanisms in data-intensive AI environments. A. Meroño-Peñuela [13] proposes a knowledge-graph-based governance framework for AI workflows, showing how semantic structures can support data governance beyond traditional cataloging. A. Ahmed [14] provides empirical evidence from healthcare that general governance frameworks require sector-specific adaptation to ensure

security and privacy in sensitive data environments. E. Permin [15] analyzes use-case-driven data platform architectures in manufacturing, clarifying how platform design is shaped by operational scenarios rather than by tooling choices alone.

2.2. Methods

The study uses a structured conceptual review design combined with comparative synthesis and interpretive modeling. The review protocol was organized in four stages. First, the literature pool was delimited to peer-reviewed journal articles and conference papers that address one or more of the following themes: lakehouse architecture, data platforms, data products, data mesh or data fabric, feature stores, data governance, data quality management, and responsible AI governance. Second, the source set was screened for conceptual relevance to enterprise AI delivery under compliance pressure. Publications were retained when they discussed architectural coordination, control mechanisms, lifecycle traceability, or governance practices applicable to regulated organizations. Third, the selected studies were coded along a common comparison grid comprising architectural unit of delivery, governance distribution, metadata and lineage mechanisms, reuse logic, ML lifecycle support, and compliance-related evidence generation. Fourth, the coded material was synthesized into an integrative model explaining how platform-centric architecture relocates controls from isolated transformation chains to shared data services.

The synthesis followed a cross-source triangulation logic. The analysis compared how different strands of the literature converged or diverged on five analytical dimensions: traceability, reproducibility, access control, data quality discipline, and coupling between analytical and ML workloads. This procedure increased the replicability of the conceptual analysis by making the comparison criteria explicit.

The scope of the review is intentionally limited. The article focuses on enterprise environments where AI delivery is constrained by auditability, privacy, and reproducibility requirements. The analysis is most applicable to industries such as financial services, healthcare, insurance, and other sectors with high documentation and control obligations. The conclusions are less directly transferable to low-regulation digital environments where platform adoption is driven mainly by scale or developer productivity.

3. RESULTS

The results reveal a consistent pattern across the analyzed literature: while pipeline-centric architectures optimize localized data transformation efficiency, they systematically underperform in providing unified governance, reuse, and lifecycle traceability for enterprise AI.

Pipeline-first data engineering has historically optimized throughput of ingestion and transformation; however, its scaling pattern frequently leads to parallel pipelines that replicate joins, feature logic, and validation routines. The lakehouse paradigm reframes this by consolidating storage and analytical processing on open, direct-access formats while serving both BI and ML workloads within a single architectural pattern, reducing architectural bifurcation between “warehouse for reporting” and “lake for science” [1]. In parallel, comparative reviews of warehouse and lake architectures describe how schema-on-write rigidity and raw-data sprawl form opposing failure modes that often trigger migration toward more unified platform constructs [8].

Platform-centric design redefines data assets as managed products with explicit lifecycle, ownership, and consumption contracts. Conceptual work on data products, data mesh, and data

fabric positions them as self-contained units that combine data, metadata, policies, and operational dependencies [11]. In contrast, mesh and fabric emphasize different coordination mechanisms for access, interoperability, and metadata-driven automation [3]. A delimitation study further clarifies that centralized data platforms and data fabrics tend to preserve stronger centralized control. At the same time, decentralized meshes prioritize flexibility through domain ownership, altering governance distribution [5]. Within regulated environments, this distinction is crucial because accountability for data quality, lineage, and access requires explicit assignment, regardless of the level of centralization.

A practical instantiation of productized data is evident in "data-as-a-product" implementations for industrial networks, where the architecture is organized around services for abstraction, interoperability, quality, security, and traceability, complemented by mechanisms for sovereignty and discovery [9]. This service orientation aligns with governance research that frames data governance not as a documentation activity, but as a structured discipline that links quality management, stewardship, and maturity models with operational controls [2]. Taken jointly, these sources indicate that platform-centric data is not a tooling upgrade; it is a redesign of how control points are embedded into the data layer, shifting assurance from manual approvals to repeatable platform services [11,13].

Feature stores provide a concrete platform mechanism that bridges data engineering and model operations through a shared, governed feature layer. Feature store literature describes architectures separating offline stores for high-throughput training from online stores for low-latency inference, supported by catalogs for discovery and reuse [6]. The same work highlights point-in-time joins as a correctness primitive that prevents training data leakage by reconstructing the historical state at a given timestamp [6]. In a complementary platform view, feature stores are treated as a shared data layer that connects feature, training, and inference pipelines through well-defined read/write contracts, enabling independent development and operation of each pipeline while keeping data consistency anchored in the shared layer [7]. This pattern directly supports regulated requirements for reproducibility and audit trails: training datasets can be regenerated from governed feature definitions.

Figure 1 synthesizes the pipeline-to-platform transition around the control surfaces most relevant to regulated AI delivery. The figure is an adapted conceptual consolidation derived from lakehouse unification [1], product/mesh/fabric integration logic [3], governance mode distinctions [5], and feature store lifecycle coupling [6,7].

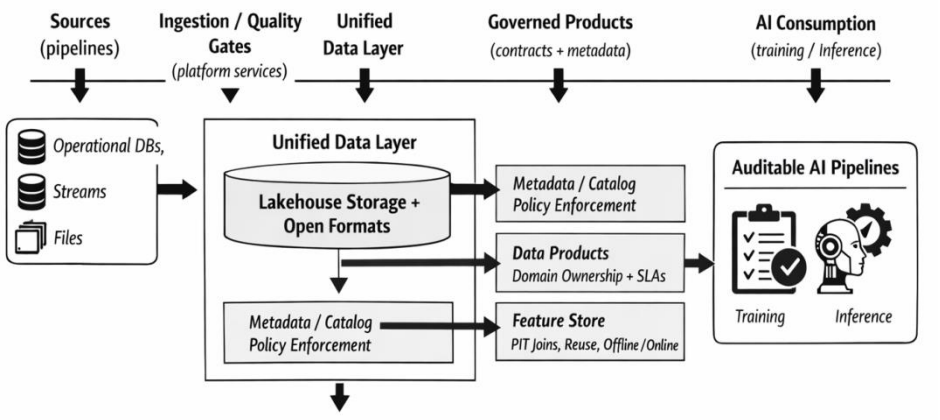

Figure 1. Platform-centric data layer: transition from pipeline orchestration to governed data products and ML feature services (adapted from [1,3,5–7]).

Figure 1 can be read as a layered architectural transition model. The bottom layer provides data persistence and unified storage, anchoring heterogeneous analytical and ML workloads in a shared data foundation. The next layer concentrates metadata, lineage, and policy services that standardize visibility and control across domains. Above it, productized data contracts organize reusable datasets as governed units with explicit ownership, quality expectations, and access rules. The upper ML-oriented layer contains feature generation and feature-serving mechanisms that preserve time-consistent joins and training-serving alignment. The top governance layer connects these technical services to auditability, reproducibility, and oversight requirements. In this reading, the figure does not merely illustrate technology components; it shows how control surfaces migrate from fragmented pipelines into persistent platform services.

More specifically, the left-hand side of the figure corresponds to pipeline-era orchestration, where delivery logic is distributed across multiple ETL chains. The right-hand side condenses the platform logic, where governance, reuse, and ML consistency are anchored in shared infrastructure. The analytical value of the figure lies in showing where evidence, validation, and accountability are produced.

From a governance standpoint, platform-centric data strengthens the link between data controls and responsible AI oversight. Responsible AI governance research emphasizes the separation of governance across data, models, and AI systems, advocating for structured practices that operationalize oversight [10]. When combined with data governance and quality management synthesis, the results indicate a convergence: platform services (catalogs, lineage, policy enforcement, quality monitoring) become the technical substrate through which governance practices gain operational expression [12,13]. In regulated industries, this convergence reduces dependence on post-hoc evidence collection during audits by embedding evidence generation (such as lineage, access logs, feature definitions, and dataset snapshots) into the platform layer.

4. DISCUSSION

The results support a conceptual reinterpretation of “pipelines versus platforms” as a transition from transformation sequences to managed data services with explicit accountability surfaces. Lakehouse unification reduces architectural fragmentation across analytics and ML workloads; yet, governance discipline remains the differentiator. Without product contracts, rigorous metadata management, and quality control, a unified storage layer can still devolve into uncontrolled accumulation [1,2,8]. Data products and mesh/fabric research indicate that decentralization shifts where governance decisions are executed; it does not eliminate governance workload, and it can amplify governance complexity unless computational policies and shared platform services are institutionalized [3,5]. Feature stores exemplify a mature platform pattern for ML data, characterized by point-in-time correctness, reuse economics, and shared lifecycle coupling across feature engineering, training, and inference, which are platform properties [6,7].

Table 1. Comparative characteristics of pipeline-centric and platform-centric data operating models (sources: [1–3,5–7,9,10]).

Dimension	**Pipeline-centric model**	**Platform-centric model**
Unit of delivery	ETL job/workflow chain	Data product + governed services
Control surface	Job-level checks are uneven across teams	Platform policies, catalogs, lineage, and standardized validation

Reuse economics	Repeated transformations across pipelines	Shared features and products are reused across consumers
ML correctness	Leakage risk when historical joins are ad hoc	Point-in-time joins and governed feature definitions
Governance distribution	Central governance detached from execution	Governance embedded in platform services and product lifecycle
Audit evidence	Reconstructed after the incidents	Generated continuously via metadata, access controls, and lineage

A regulated-industry framing highlights that platform-centric architecture gains value when mapped to explicit control objectives: traceability, reproducibility, controlled access, and lifecycle oversight. Data-as-a-product implementations emphasize traceability and security services as architectural building blocks rather than add-ons, aligning with governance synthesis that treats data quality management as an operational system with maturity pathways [2,9]. Feature stores extend this mapping into ML execution by anchoring training/serving consistency in shared definitions, and time-consistent joins [6,7]. Responsible AI governance research contributes a higher-level governance structure, connecting data governance to broader governance of models and AI systems, which becomes actionable when the data layer provides durable evidence artifacts [10].

The practical relevance of this architectural transition becomes clearer when mapped to sector-specific compliance pressures. In financial services, platform-centric architecture supports retention of transformation lineage for risk models, model input reconstruction, and controlled access to sensitive customer data under supervisory review and internal audit procedures. In healthcare, the same architectural logic strengthens traceability of clinical or claims-related data transformations, improves reproducibility of feature extraction used in predictive models, and reduces the governance burden associated with fragmented data preparation across departments [14]. In insurance and adjacent risk-intensive sectors, governed data products and shared feature services help standardize underwriting inputs, fraud-detection features, and model evidence trails across multiple analytical teams. Across these sectors, the architectural gain does not reduce to higher throughput; the gain lies in making compliance-relevant evidence persistently available at the data layer [12].

At the same time, the argument developed in this article has defined boundaries of applicability. The conclusions are strongest for enterprises subject to formal obligations regarding auditability, privacy, model documentation, and reproducibility. They are less persuasive for small organizations with limited governance exposure, low model criticality, or short-lived analytical use cases, where the coordination overhead of a platform may exceed the control benefits. The analysis speaks most directly to medium and large enterprises whose AI operations require durable evidence, cross-team reuse, and policy-enforced data access.

Table 2. Regulated control objectives and platform mechanisms that operationalize them [1–3,5–7,9,10]

Control objective	**Platform mechanism**
Traceability of data transformations	Lineage + metadata services, product documentation, traceability services

Reproducible model training datasets	Unified storage layer + governed feature definitions + dataset regeneration
Prevention of training data leakage	Point-in-time joins for historical reconstruction
Controlled access and policy enforcement	Centralized policy enforcement with catalog-driven access patterns
Continuous evidence for oversight	Operational governance practices linked to data, model, and system governance

A balanced interpretation of platform adoption must account for organizational and technical barriers. Legacy estates often contain duplicated schemas, undocumented dependencies, and fragmented ownership, which complicate migration from pipeline-first delivery to shared platform services. Organizationally, the transition frequently requires redistributing responsibilities between central data teams and domain teams, creating friction over stewardship, funding, and control authority. Culturally, platform adoption presupposes a shift from project-specific delivery toward reusable internal products, which may conflict with local optimization habits and short-term delivery incentives [15]. For regulated enterprises, these frictions are amplified by the need to preserve operational continuity during migration. As a result, platform transition is best understood as a staged reorganization of data ownership, control points, and evidence production.

Taken together, the discussion suggests a design principle for regulated enterprises: platform-centric data architecture gains operational legitimacy when governance practices are expressed as platform services (catalogs, lineage, policy engines, standardized validation) and when ML-specific data primitives (point-in-time correctness, offline/online feature separation, shared feature lifecycle) are treated as first-class data layer functions. This orientation places auditability and explain ability prerequisites at the data layer, where provenance, consistency, and access decisions originate. In regulated enterprise AI, that relocation strengthens both operational trust and oversight readiness.

5. CONCLUSIONS

The analysis shows that the transition from pipeline-centric delivery to platform-centric data architecture changes the operational foundation of enterprise AI in regulated industries. The architectural shift is expressed through a move from isolated transformation chains toward governed data products, shared metadata and lineage services, and feature layers that preserve reproducibility across training and inference. Under such conditions, auditability is produced continuously within the data layer.

For practitioners, five implications follow from the synthesis. First, platform transition should begin with standardizing metadata, capturing lineage, and enforcing policies, because these services establish the control baseline for later architectural scaling. Second, reusable data products require explicit ownership, service-level expectations, and documented consumption contracts; without them, platform language remains nominal. Third, ML-intensive organizations

benefit from introducing governed feature definitions and point-in-time reconstruction rules early, since these mechanisms directly improve reproducibility and reduce leakage risk. Fourth, migration should proceed in phases, prioritizing high-risk and high-reuse data domains. Fifth, governance practices need to be embedded into the delivery infrastructure so that evidence for audits, internal controls, and model reviews is generated during operation.

The study remains limited by its literature-based design and by the absence of longitudinal organizational cases. Further research may proceed in several directions: comparative case studies of platform adoption across regulated industries; empirical assessment of how lineage, feature stores, and productized data contracts affect audit preparation effort; maturity models for staged migration from warehouse- and pipeline-first estates to platform operating models; and measurement frameworks linking platform controls to responsible AI outcomes such as reproducibility, policy compliance, and incident response quality.

REFERENCES

[1] M. A. Zaharia, A. Ghodsi, R. Xin and M. Armbrust, "Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics," in Proceedings of the Conference on Innovative Data Systems Research (CIDR), 2021.

[2] B. M. V. Bernardo, H. S. Mamede, J. M. P. Barroso and V. M. P. D. dos Santos, "Data Governance & Quality Management—Innovation and Breakthroughs across Different Fields," Journal of Innovation & Knowledge, vol. 9, no. 4, Art. no. 100598, 2024. doi: 10.1016/j.jik.2024.100598.

[3] I. Blohm, F. Wortmann, C. Legner et al., "Data Products, Data Mesh, and Data Fabric," Business & Information Systems Engineering, vol. 66, pp. 643–652, 2024. doi: 10.1007/s12599-024-00876-5.

[4] P. K. Donta, C. K. Dehury and Y.-C. Hu, "Learning-Driven Data Fabric Trends and Challenges for Cloud-to-Thing Continuum," Journal of King Saud University – Computer and Information Sciences, vol. 36, no. 7, Art. no. 102145, 2024. doi: 10.1016/j.jksuci.2024.102145.

[5] A. Gieß and A. Hutterer, "The Future of Data Management: A Delimitation of Data Platforms, Data Spaces, Data Meshes, and Data Fabrics," Information Systems and E-Business Management, vol. 23, pp. 971–997, 2025. doi: 10.1007/s10257-025-00707-4.

[6] R. Liu, K. Park, F. Psallidas, X. Zhu, J. Mo, R. Sen, M. Interlandi, K. Karanasos, Y. Tian and J. Camacho-Rodríguez, "Optimizing Data Pipelines for Machine Learning in Feature Stores," Proceedings of the VLDB Endowment, vol. 16, no. 13, pp. 4230–4239, 2023. doi: 10.14778/3625054.3625060.

[7] J. de la Rúa Martínez, F. Buso, A. Kouzoupis, A. A. Ormenisan, S. Niazi, D. Bzhalava, K. Mak, V. Jouffrey, M. Ronström, R. Cunningham, R. Zangis, D. Mukhedkar, A. Khazanchi, V. Vlassov and J. Dowling, "The Hopsworks Feature Store for Machine Learning," in Companion of the 2024 International Conference on Management of Data (SIGMOD '24), pp. 135–147, 2024. doi: 10.1145/3626246.3653389.

[8] A. Nambiar and D. Mundra, "An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management," Big Data and Cognitive Computing, vol. 6, no. 4, Art. no. 132, 2022. doi: 10.3390/bdcc6040132.

[9] A. Nizamis, M. Julian, C. I. Valero, M. Foti, I. Drigkopoulou, M. Á. Esbrí, R. Costa, A. Yortholt, D. Ioannidis, P. Gkonis, P. Trakadas, D. Tzovaras and C. E. Palau, "Data-as-a-Product to Enable Data-Driven Value Networks in Industries 4.0 & 5.0: The Swiss Smart Factory Experiment," Procedia Computer Science, vol. 257, pp. 793–800, 2025. doi: 10.1016/j.procs.2025.03.102.

[10] E. Papagiannidis, P. Mikalef and K. Conboy, "Responsible Artificial Intelligence Governance: A Review and Research Framework," Journal of Strategic Information Systems, vol. 34, no. 2, Art. no. 101885, 2025. doi: 10.1016/j.jsis.2024.101885.

[11] R. Eichler, A. Kaltenbrunner, P. Drews and I. Schirmer, "Introducing the Enterprise Data Marketplace: a Platform for Democratizing Company Data," Journal of Big Data, vol. 10, Art. no. 165, 2023. doi: 10.1186/s40537-023-00843-z.

[12] X. Ye, Y. Yan, J. Li and B. Jiang, "Privacy and Personal Data Risk Governance for Generative Artificial Intelligence: A Chinese Perspective," Telecommunications Policy, vol. 48, no. 10, Art. no. 102851, 2024. doi: 10.1016/j.telpol.2024.102851.

[13] A. Meroño-Peñuela, E. Simperl, A. Kurteva and I. Reklos, "KG.GOV: Knowledge Graphs as the Backbone of Data Governance in AI," Web Semantics, vol. 83, Art. no. 100847, 2025. doi: 10.1016/j.websem.2024.100847.

[14] A. Ahmed, A. Shahzad, A. Naseem, S. Ali and I. Ahmad, "Evaluating the Effectiveness of Data Governance Frameworks in Ensuring Security and Privacy of Healthcare Data: A Quantitative Analysis of ISO Standards, GDPR, and HIPAA in Blockchain Technology," PLOS ONE, vol. 20, no. 5, Art. no. e0324285, 2025. doi: 10.1371/journal.pone.0324285.

[15] E. Permin, C. Wohlgemuth and T. Keller, "Use-Case-Driven Architectures for Data Platforms in Manufacturing," Platforms, vol. 3, no. 3, Art. no. 15, 2025. doi: 10.3390/platforms3030015.