Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 06| 2025

Research Article **Received:** 17-05-2025 | **Accepted:** 10-06-2025 | **Published:** 29-06-2025

Data-Centric AI: Engineering Platforms for Pre-Model Intelligence

Rahul Joshi

IIT Kharagpur, India

Abstract: This article explores the paradigm shift from model-centric to data-centric approaches in artificial intelligence, emphasizing how the quality, structure, and governance of data have become primary determinants of AI system success. The article examines the theoretical foundations of data-centric AI and demonstrates its particular significance in regulated industries where explainability, reproducibility, and auditability are non-negotiable requirements. The article outlines the essential architectural components, including feature stores, schema-aware pipelines, and metadata management systems, that enable robust AI implementations. The article further investigates how comprehensive data lineage, sophisticated labeling infrastructure, and drift detection mechanisms create resilient foundations for enterprise AI systems. The article on responsible AI implementation through data engineering highlights how fairness assessment, bias mitigation, and ethical considerations can be systematically addressed at the data layer. The article illustrates successful patterns for data-centric platform engineering in practice. The article concludes by identifying emerging standards, open challenges, and promising research directions that will shape the future of pre-model intelligence, establishing data platform engineering as a critical discipline for organizations seeking to build AI systems that deliver consistent value while maintaining the highest standards of reliability, fairness, and transparency.

Keywords: Data-centric AI, Feature store architecture, Lineage tracking, Drift detection, Responsible AI implementation.

INTRODUCTION

The artificial intelligence landscape has undergone a profound transformation in recent years, shifting away from a predominantly model-centric approach toward one that recognizes the paramount importance of data quality and governance. While architectural innovations in neural networks have historically dominated AI research discourse, practitioners across industries are increasingly discovering that sophisticated models built on flawed, inconsistent, or poorly managed data inevitably yield disappointing results. This realization has catalyzed the emergence of data-centric AI—a paradigm that positions data engineering as the foundation upon which successful machine learning systems must be built.

Data-centric AI represents a fundamental reorientation of priorities, where engineering effort focuses first on creating robust, consistent, and well-labeled datasets before addressing model architecture. As Andrew Ng, a prominent advocate of this approach, observed, "The model and the code are commoditizing... It's the data that's key" (Patel, H. 2025). This perspective has gained particular traction in enterprise environments, where the complexity of legacy systems, regulatory requirements, and organizational data silos create significant challenges for AI deployment.

For organizations in regulated industries such as financial services, healthcare, and insurance, the stakes are especially high. These sectors face stringent requirements for explainability, reproducibility, and auditability—qualities that cannot be retrofitted onto systems lacking proper data infrastructure. The consequences of poor data governance extend beyond model performance to include regulatory penalties, reputational damage, and erosion of stakeholder trust. Accordingly, data platform engineering has emerged as a critical discipline that enables not only technical excellence but also responsible and compliant AI systems.

This article examines the architectural patterns, engineering practices, and organizational approaches that enable effective data-centric AI. The article explores how feature stores, schema-aware pipelines, lineage tracking, and drift detection systems create the necessary foundation for pre-model intelligence. Through examination of implementation patterns across industries, we demonstrate that investing in data platform engineering yields dividends not only in model performance but also in accelerated development cycles, reduced operational risk, and enhanced regulatory compliance. As AI systems become increasingly embedded in critical business functions, the quality of the underlying data infrastructure will continue to determine the boundary between success and failure. By focusing engineering effort on pre-model intelligence, organizations can build AI systems that deliver consistent value while maintaining the highest standards of reliability, fairness, and transparency.

***Corresponding Author:** Rahul Joshi

DOI-https://doi.org/10.5281/zenodo.15768724

Theoretical Foundations of Data-Centric AI

The data-centric AI paradigm rests on the fundamental premise that improving data quality yields better results than iterative model refinement. This approach reframes machine learning development by emphasizing systematic data engineering over algorithmic optimization. As defined by Roh *et al.*, data-centric AI represents "a systematic engineering discipline focused on data quality management for achieving reliable and trustworthy AI systems" (Roh, Y., Heo. *et al.*, 2019). Unlike model-centric approaches that maintain static datasets while experimenting with model architectures, data-centric AI keeps the model relatively fixed while systematically improving data quality. This inversion of priorities recognizes that even simple models can achieve remarkable performance when trained on high-quality, well-structured data. Key metrics for measuring data quality in AI systems include completeness, consistency, accuracy, timeliness, and representativeness. These dimensions must be systematically measured and improved throughout the ML lifecycle. Particularly important is the concept of "fitness for purpose," which evaluates whether a dataset appropriately represents the specific problem domain and target distribution.

The relationship between data governance and model performance operates through multiple mechanisms. Effective governance ensures proper versioning, accessibility, and documentation of datasets, enabling reproducible experiments and reliable deployments. Governance frameworks also enforce validation rules that prevent data degradation over time, maintaining model performance in production environments.

Table 1: Comparison of Model-Centric vs. Data-Centric AI Approaches (Patel, H. 2025; Roh, Y., Heo. *et al.*, 1918).

Dimension	Model-Centric Approach	Data-Centric Approach
Primary Focus	Algorithm selection and hyperparameter tuning	Data quality, consistency, and representativeness
Development Cycle	Fixed dataset, iterative model refinement	Fixed model architecture, iterative data improvement
Success Metrics	Model performance on the test set	Data quality metrics and model generalization capability
Engineering Emphasis	Neural architecture search, optimization methods	Data validation, lineage tracking, feature engineering
Regulatory Alignment	Post-hoc explanations and documentation	Built-in transparency through comprehensive data governance

Enterprise Requirements in Regulated Industries

Regulated industries face stringent requirements that directly impact AI system design and implementation. In financial services, regulations such as GDPR, CCPA, and industry-specific frameworks like SR 11-7 (for model risk management) impose strict controls on data usage, model validation, and ongoing monitoring (Federal Reserve. 2011). Healthcare organizations must navigate HIPAA compliance while ensuring patient data remains secure throughout the ML lifecycle.

Compliance requirements typically mandate documentation of training data sources, preprocessing steps, validation procedures, and model performance metrics. Organizations must demonstrate that their AI systems produce consistent, explainable results and that appropriate controls exist to prevent discriminatory outcomes or security vulnerabilities.

Auditability and reproducibility pose significant challenges in enterprise environments. Systems must maintain comprehensive audit trails documenting data transformations, feature engineering steps, and model training procedures. This lineage information must be preserved to enable regulatory examinations and internal governance reviews.

Notable regulatory failures highlight the consequences of inadequate data governance. Examples include algorithmic discrimination in credit scoring systems and medical diagnosis tools that perform poorly on underrepresented patient populations. These cases demonstrate that insufficient attention to data quality, representativeness, and governance creates substantial regulatory and reputational risks, even when sophisticated models are employed.

Publisher: SARC Publisher

Core Architectural Components for Data-Centric AI

Feature stores have emerged as critical infrastructure for data-centric AI, serving as centralized repositories that standardize feature engineering, ensure consistency between training and serving, and enable feature reuse across models. Modern feature store architectures typically implement dual storage systems: an offline store for training data (often using columnar formats) and an online store for low-latency inference (typically using key-value databases) (Feast.). Implementation considerations include time-travel capabilities, feature versioning, and transformation consistency guarantees.

Schema-aware pipeline architectures enforce structural constraints on data flowing through ML systems, preventing silent failures caused by unexpected format changes. These pipelines utilize explicit schema definitions to validate inputs and outputs at each transformation stage. Tools like TensorFlow Data Validation have popularized this approach by enabling automated schema inference and drift detection.

Data validation frameworks establish systematic quality checks throughout the ML lifecycle. These frameworks implement rule-based validation (range checks, uniqueness constraints), statistical validation (distribution tests, anomaly detection), and semantic validation (domain-specific rules). The most effective implementations integrate validation directly into data ingestion workflows, preventing contamination of feature stores with low-quality data. Metadata management systems provide searchable catalogs of datasets, features, and models along with their relationships and properties. These systems capture technical metadata (schema, volume, freshness) and business metadata (ownership, sensitivity classification, usage policies). As noted by Sculley *et al.*, comprehensive metadata management is essential for avoiding hidden technical debt in ML systems (Sculley, D. *et al.*, 2015).

Table 2: Core Components of Data-Centric AI Platforms (Feast. ; Ratner, A *et al.*, 2017).

Component	Primary Function	Implementation Considerations
Feature Store	Centralized repository for feature management	Dual storage (offline/online), versioning, transformation consistency
Schema Registry	Definition and enforcement of data contracts	Evolution management, validation rules, compatibility checking
Data Validation Framework	Systematic quality assessment	Rule-based, statistical, and semantic validation methods
Lineage System	Tracking data transformations	Granularity level, query capabilities, integration with CI/CD
Drift Detection	Monitoring for distribution shifts	Statistical methods, alerting thresholds, and automated responses

Data Lineage and Provenance Systems

Tracking data transformations across the ML lifecycle requires capturing detailed provenance information at each processing step. Effective lineage systems record input sources, transformation operations, parameter configurations, and output destinations. This granular tracking enables root cause analysis when models underperform and provides evidence for regulatory compliance.

Implementation patterns for comprehensive lineage capture include embedded instrumentation, where processing frameworks automatically log transformation details, and declarative pipelines that generate lineage as a byproduct of execution. Leading organizations implement graph-based data models that represent the entire ML workflow as interconnected nodes, enabling powerful query capabilities for impact analysis and troubleshooting.

Integration with CI/CD workflows extends lineage tracking beyond data transformations to include code changes, configuration updates, and deployment events. This integration creates end-to-end traceability from source data through model deployment. Modern implementations use orchestration platforms that combine workflow automation with comprehensive provenance tracking. Technical approaches to immutable audit trails often leverage append-only logs or blockchain-inspired structures to ensure lineage records cannot be altered. These systems implement cryptographic verification mechanisms to detect tampering and establish trusted provenance chains. Enterprise implementations frequently combine these technical safeguards with

Publisher: SARC Publisher

governance controls that restrict access to lineage systems based on role-based permissions.

Labeling and Annotation Infrastructure

Scalable data labeling infrastructure forms a critical component of data-centric AI systems, particularly for supervised learning applications. Modern approaches combine automated pre-labeling with human verification to maximize efficiency while maintaining quality. Organizations increasingly adopt programmatic labeling techniques, where subject matter experts develop labeling functions that can be applied at scale, rather than manually annotating individual examples (Ratner, A *et al.,* 2017). This approach has proven particularly effective for specialized domains where expert knowledge is required. Human-in-the-loop system design addresses the challenges of coordinating annotation work across distributed teams while ensuring consistency and quality. Effective implementations employ specialized UIs that reduce cognitive load, provide contextual information, and incorporate built-in quality checks. These systems typically implement workflow management capabilities that route difficult cases to experienced annotators while distributing routine tasks more broadly. Quality control mechanisms for labeled datasets operate at multiple levels, including redundant annotation (having multiple annotators label the same examples), gold standard comparisons (measuring against expert-verified references), and statistical outlier detection. The most sophisticated implementations calculate inter-annotator agreement metrics and identify systematic biases in labeling patterns. Regular calibration sessions among annotators help maintain consistency for subjective labeling tasks.

Active learning implementation patterns enable more efficient labeling by prioritizing the most informative examples for human review. These systems iteratively identify examples with high uncertainty or that lie near decision boundaries, focusing human effort where it provides the greatest model improvement. Practical implementations combine multiple sampling strategies (uncertainty sampling, diversity sampling, and expected model change) to achieve balanced training datasets.

Drift Detection and Monitoring

Statistical methods for identifying data and concept drift include distribution-based approaches (using KL divergence, JS distance, or population stability index), prediction-based methods (monitoring changes in model outputs), and performance-based techniques (tracking accuracy metrics over time). As outlined by Gama *et al.*, comprehensive drift detection requires monitoring multiple metrics to capture different types of distribution shifts (Gama, J. *et al.,* 2014). Modern systems implement both univariate tests for individual features and multivariate methods for capturing complex interactions. Real-time monitoring architectures typically implement a pipeline of specialized detectors that process incoming data streams, comparing current distributions against reference distributions established during training. These architectures maintain statistical profiles of both input features and model predictions, generating alerts when significant deviations occur. The most effective implementations separate detection (identifying potential issues) from diagnosis (determining root causes) to enable appropriate responses.

Feedback loops and model retraining triggers operate through automated decision rules that initiate retraining workflows when specific conditions are met. These conditions may include performance degradation below threshold levels, data drift exceeding acceptable bounds, or scheduled intervals based on domain knowledge. Leading organizations implement graduated response systems that can apply lightweight corrections (such as calibration adjustments) for minor drift while initiating full retraining for substantial shifts. Performance degradation prevention strategies combine proactive and reactive approaches. Proactive techniques include ensemble methods that incorporate models trained on different periods, adversarial training to build robustness against distribution shifts, and continual learning architectures that incrementally update models with new data. Reactive strategies focus on rapid detection and adaptation, with automated fallback mechanisms that can revert to previous model versions when degradation is detected.

Responsible AI Implementation through Data Engineering

Fairness assessment via data analysis forms the foundation of responsible AI implementation. Data-centric approaches to fairness begin with comprehensive demographic analysis of training datasets, examining representation across protected attributes and identifying potential sources of bias. Techniques include group fairness metrics (demographic parity, equal opportunity, equalized odds) and individual fairness measures that compare similar individuals across groups (Mehrabi, N. *et al.,* 2021). Leading organizations

Publisher: SARC Publisher

implement fairness assessments as automated components of their data validation pipelines, preventing biased datasets from entering production systems.

Bias detection and mitigation at the data layer addresses issues before they become embedded in models. Effective approaches include balanced sampling techniques, fairness-aware data augmentation, and counterfactual data generation to address underrepresentation. As documented by Holstein *et al.*, practitioners increasingly recognize that bias mitigation is most effective when applied during data preparation rather than as a post-processing step (Holstein, K. *et al.*, 2019). Modern platforms implement bias dashboards that visualize fairness metrics across dataset versions, enabling data scientists to quantify the impact of mitigation techniques.

Transparency and explainability through data lineage provide critical foundations for trustworthy AI. Comprehensive lineage systems enable tracing model predictions back to their source data, documenting all transformations and enrichments. This capability allows organizations to answer regulatory inquiries, investigate unexpected behaviors, and verify compliance with data usage policies. The most effective implementations integrate lineage tracking with feature importance measures, connecting model explanations to specific data sources and transformations.

Ethical considerations in data platform design include privacy preservation mechanisms, consent management, and data minimization principles. Leading organizations implement tiered access controls based on data sensitivity, automated PII detection and redaction, and privacy-preserving computation techniques such as differential privacy (Belhassen, 2022). These safeguards are increasingly embedded directly into data platform architecture rather than implemented as afterthoughts, reflecting the growing recognition that ethical AI requires ethical data infrastructure.

Table 3: Data Quality Metrics for AI Systems(Gama, J. *et al.*, 2014 ;Holstein, K. *et al.*, 2019)

Metric Category	Examples	Relevance to AI Performance
Completeness	Missing value rates, coverage of key subgroups	Prevents model bias and improves generalization
Consistency	Schema conformance, logical relationship validation	Ensures reliable feature extraction and transformation
Accuracy	Error rates against ground truth, annotation quality	Directly impacts supervised learning performance
Timeliness	Data freshness, temporal consistency	Critical for real-time systems and concept drift management
Representativeness	Distribution alignment with the production environment	Fundamental for model generalization to real-world scenarios

Case Studies: Data-Centric Platform Engineering in Practice

Financial services organizations have pioneered data-centric AI platforms to address their strict regulatory requirements. JPMorgan Chase's Omni AI platform exemplifies this approach, implementing centralized feature stores, comprehensive lineage tracking, and automated bias detection for models across trading, risk management, and customer service domains (John China. 2024). Their platform emphasizes governance controls that enforce model validation protocols and maintain audit trails for regulatory examinations, demonstrating how data-centric architecture enables both innovation and compliance in highly regulated environments. Healthcare data platform architectures must address unique challenges related to data heterogeneity, privacy concerns, and interoperability requirements. Mayo Clinic's Clinical Data Analytics Platform illustrates effective patterns, including de-identification pipelines that preserve analytical utility, federated learning approaches that maintain data locality, and annotation workflows tailored to clinical specialists Mayo Clinic). These implementations emphasize data quality validation specific to medical contexts, such as temporal consistency checks and physiological plausibility rules. Manufacturing quality control systems leverage data-centric approaches to improve defect detection and process optimization. Siemens' Industrial Edge platform demonstrates how time-series data management, sensor validation frameworks, and domain-specific feature engineering create robust foundations for AI-powered quality inspection (Siemens. 2024). Their implementation emphasizes real-time drift detection to account for changing environmental conditions and wear patterns that affect sensor

Publisher: SARC Publisher

readings, highlighting how data-centric engineering addresses the dynamic nature of industrial environments.

Retail recommendation engines benefit significantly from data-centric approaches that integrate diverse customer signals while maintaining privacy safeguards. Walmart's Customer 360 platform exemplifies effective patterns, including feature stores that unify online and in-store behaviors, contextual enrichment services that incorporate environmental factors, and consent-aware processing pipelines (8th & Walton. 2022). Their implementation emphasizes an A/B testing infrastructure that enables rapid iteration on recommendation strategies while maintaining consistent measurement frameworks.

Table 4: Responsible AI Implementation Patterns at the Data Layer (Mehrabi, N. *et al.*, 2021; John China. 2024)

Objective	Data Engineering Approach	Implementation Example
Fairness Assessment	Demographic analysis of training data	Automated fairness metrics in validation pipelines
Bias Mitigation	Balanced sampling, counterfactual generation	Fairness-aware data augmentation techniques
Transparency	Comprehensive data and model lineage	End-to-end traceability from source data to predictions
Privacy Protection	Data minimization, anonymization	PII detection and differential privacy implementation
Ethical Governance	Tiered access controls, usage policies	Consent management infrastructure and audit mechanisms

FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

Emerging standards for data-centric AI represent a critical area of development as organizations seek to establish common frameworks for data quality assessment, lineage tracking, and responsible AI implementation. Industry consortia and academic collaborations are working to standardize metadata schemas, interchange formats for data documentation, and fairness metrics. The MLOps community is converging around standardized interfaces for feature stores, data validation systems, and lineage tracking, though significant work remains to achieve interoperability across vendor ecosystems.

Open challenges in platform engineering for AI include scaling data validation to high-dimensional, unstructured data types; implementing efficient lineage tracking for complex transformation graphs; and developing automated remediation techniques for detected quality issues. Particularly challenging is the creation of unified governance frameworks that can span hybrid cloud environments while maintaining consistent controls. As AI systems become more pervasive, platform engineers must also address growing complexity in managing cross-team dependencies and resource contention.

Integration with emerging AI technologies presents both opportunities and challenges for data-centric platforms. Foundation models require novel approaches to data curation, prompt engineering governance, and fine-tuning dataset management. Techniques like retrieval-augmented generation necessitate reliable knowledge bases with rigorous fact verification. As noted by (Bommasani, R. *et al., 2021)* these models introduce new considerations for data platform design, including prompt versioning, retrieval corpus maintenance, and alignment dataset curation (Bommasani, R. *et al.*, 2021).

Research directions for pre-model intelligence include automated dataset repair techniques, causal structure discovery for improved feature engineering, and transfer learning approaches for data enhancement. Promising work explores self-supervised methods for data quality assessment that can identify problematic examples without human intervention. Additional research opportunities include developing interpretable data representations that naturally support downstream explainability, privacy-preserving data enrichment techniques, and automated dataset generation for underrepresented scenarios. These advancements will further strengthen the foundations upon which reliable, responsible AI systems are built.

CONCLUSION

The evolution toward data-centric AI represents a fundamental shift in how organizations conceptualize, build, and maintain intelligent systems. As this article has demonstrated, the quality, governance, and engineering of data

Publisher: SARC Publisher

infrastructure now serve as the primary determinants of AI success in enterprise environments. By investing in robust feature stores, comprehensive lineage tracking, rigorous validation frameworks, and responsible AI practices at the data layer, organizations establish the necessary foundation for building reliable, explainable, and compliant AI systems. The case studies across financial services, healthcare, manufacturing, and retail illustrate that data platform engineering is not merely a technical consideration but a strategic imperative that enables both innovation and risk management. As AI continues to transform business operations, the organizations that thrive will recognize data as their most valuable asset and implement the architectural patterns, governance frameworks, and engineering practices needed to harness its full potential. The future of enterprise AI belongs not to those with marginally superior model architectures but to those who excel at engineering the pre-model intelligence that transforms raw data into trusted, actionable insights that drive business value while upholding the highest standards of responsibility and ethics.

REFERENCES

1. Patel, H. “Data-Centric Approach vs Model-Centric Approach in Machine Learning”. *Neptune.ai, 25th April*, (2025). https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning
2. Roh, Y., Heo, G., & Whang, S. E. "A survey on data collection for machine learning: a big data-ai integration perspective." *IEEE Transactions on Knowledge and Data Engineering* 33.4 (2019): 1328-1347.
3. Federal Reserve. "Supervisory Guidance on Model Risk Management." *Board of Governors of the Federal Reserve System*. April 4, 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
4. Feast. "Introduction" Feast Feature Store. https://docs.feast.dev/
5. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. "Hidden technical debt in machine learning systems." *Advances in neural information processing systems* 28 (2015).
6. Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. "Snorkel: Rapid training data creation with weak supervision." *Proceedings of the VLDB endowment. International conference on very large data bases*. 11.3 (2017).
7. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. "A survey on concept drift adaptation." *ACM computing surveys (CSUR)* 46.4 (2014): 1-37.
8. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35.
9. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. "Improving fairness in machine learning systems: What do industry practitioners need?." *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019.
10. John China, "How JPMorgan Chase is preparing the workforce for the future of AI”. *August 13,* (2024). https://www.jpmorganchase.com/newsroom/stories/how-jpmc-is-preparing-workforce-for-ai
11. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).
12. Mayo Clinic, “Mayo Clinic Research Core Facilities: Office of Core Shared Services”. https://www.mayo.edu/research/core-facilities/services/data-analytics
13. Siemens, “Siemens drives AI adoption with Industrial Operations X and NVIDIA-accelerated Industrial PCs”., 11 November (2024). https://press.siemens.com/global/en/pressrelease/siemens-drives-ai-adoption-industrial-operations-x-and-nvidia-accelerated-industrial
14. 8th & Walton, “What Is Item 360? A Guide for Walmart Suppliers”. July 13, 2022, https://www.8thandwalton.com/blog/walmart-item-360/
15. Belhassen, A. “Machine learning for predictive maintenance: Fusing vibration sensor data and thermal imaging to forecast bearing failure.” *Sarcouncil Journal of Engineering and Computer Sciences*(2022).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Joshi, R. " Data-Centric AI: Engineering Platforms for Pre-Model Intelligence." *Sarcouncil Journal of Multidisciplinary 5.6* (2025): pp 48-54

Publisher: SARC Publisher