ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 12, Issue 3, May - June 2025

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.028

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 8.028 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 12, Issue 3, May - June 2025 |

Trustworthy Data Journeys AI and the Provenance Paradigm

Lavanya Harshit Tiwari
AI Specialist, USA

ABSTRACT: Artificial intelligence (AI) systems are now integral to decision-making across industries, yet concerns around fairness, transparency, and accountability continue to undermine public trust. As models grow in complexity, understanding how data flows through these systems becomes essential. **AI-driven lineage** offers a solution by providing dynamic, real-time tracking of data and model transformations, enabling transparency throughout the AI lifecycle. This paper explores the critical role of lineage in establishing fair and accountable AI systems. We analyze existing tools, frameworks, and standards, propose a methodology for implementing AI-driven lineage, and present a layered framework that supports ethical governance and regulatory compliance.

KEYWORDS: AI Lineage, Data Governance, Explainability, Transparency, Responsible AI, Data Ethics, Fairness, Model Accountability, Provenance, AI Lifecycle

I. INTRODUCTION

The surge in AI adoption across domains such as healthcare, finance, and criminal justice has raised both optimism and concern. While AI promises efficiency and insight, it also risks reinforcing biases, making opaque decisions, and evading accountability. The question arises: how do we build AI systems that are not only intelligent but **fair and transparent**?

To answer this, we must start with **lineage**—the ability to track and understand how data and models evolve through the AI lifecycle. AI-driven lineage goes beyond traditional metadata tracking. It dynamically captures the flow of information, from data ingestion and feature engineering to model training, deployment, and monitoring. This visibility is critical for ensuring fairness (by tracing the origins of bias), transparency (by showing how decisions are made), and accountability (by identifying responsible agents and processes).

Despite its importance, lineage in AI systems remains underdeveloped. Existing tools often focus on data engineering pipelines, with limited integration into model behavior and ethical analysis. To address this gap, this paper proposes an **AI-driven lineage framework** that fuses technical lineage with ethical oversight. We examine the state of lineage in AI, explore its alignment with responsible AI principles, and provide a methodology for implementation that supports governance, auditability, and trust.

II. LITERATURE REVIEW

The foundation of data lineage lies in the field of data warehousing and database systems, where the need to trace data transformations and dependencies first emerged. **Moreau et al. (2011)** formalized provenance standards, while lineage tools such as **Apache Atlas**, **OpenLineage**, and **DataHub** emerged to support metadata collection in complex systems. In the AI context, lineage must go further. According to **Gebru et al. (2018)** and **Holland et al. (2018)**, understanding data origins and transformations is essential for fairness and transparency. Their proposals, such as **datasheets for datasets** and **dataset nutrition labels**, laid the ethical groundwork for what lineage systems could support. However, these approaches are often static and lack automation, making them less practical for dynamic AI systems.

Schelter et al. (2018) and **Koshy et al. (2022)** argue that lineage must also include model-level artifacts: hyperparameters, training environments, and even interpretability tools. **NIST's AI RMF (2023)** and the **EU AI Act** emphasize traceability as a requirement for high-risk AI, showing growing regulatory support for lineage tracking.
Despite these developments, there remains a gap between ethical aspirations and technical execution. Most tools don't track lineage across the full AI lifecycle or align with governance standards. Our work addresses this by offering a comprehensive, AI-driven lineage methodology that links technical artifacts with ethical intent.

TABLE: Evaluation of AI Lineage Tools and Capabilities

Tool	Data Lineage	Model Lineage	Real-time Tracking	Governance Support	Open Source
Apache Atlas	Yes	No	No	Partial	Yes
OpenLineage	Yes	No	Yes	Low	Yes
MLflow	Partial	Yes	Partial	Medium	Yes
Pachyderm	Yes	Yes	Yes	High	Yes
ModelDB	No	Yes	No	Medium	Yes
Comet ML	Yes	Yes	Yes	Yes	No

Evaluation of AI Lineage Tools and Capabilities

Tool / Platform	Lineage Granularity	ML-Specific Features	Integration (ML/Data)	Visualization	Versioning	Compliance Support	Notes
MLflow	Experiment- & model-level	Experiments, runs, models	Python, Spark, REST APIs	Basic (UI)	Models, runs	Partial	Strong for model tracking, limited full data lineage
Weights & Biases	Model training and tuning	Training, artifacts, hyperparameters	Pytorch, TensorFlow, etc.	Rich dashboards	Artifacts, runs	Partial	Visualization-rich, experiment lineage focus
Kubeflow Pipelines	Pipeline- and step-level	ML pipeline components, metadata	Native for K8s, TensorFlow	Interactive pipeline graphs	Via metadata	Pipeline auditability	Strong for reproducible workflows, Kubernetes-based
Apache Atlas	Table/column-level (data-focused)	⚠Basic ML support via extensions	Hive, Spark, HDFS, Kafka	Metadata UI	Via integrations	GDPR, governance features	Strong for enterprise data lineage, extensible
OpenLineage	Job- and dataset-level	ML support via orchestration (Airflow, dbt)	Airflow, dbt, Spark	Graph & lineage view	External tools	Partial	Open spec, ecosystem growing, good for pipeline metadata
DataHub	Dataset-, schema-, pipeline-level	ML metadata support growing	Kafka, Airflow, dbt, ML models	UI + graphs	Metadata snapshots	Role-based access, audit logs	General-purpose metadata engine, extensible
Neptune.ai	Experiment-level	Model training logs, artifacts	Python, Keras, XGBoost, etc.	Dashboards & traces	Experiment versions	⚠ Partial	Lightweight ML experiment tracking, great for teams
Pachyderm	File- and container-level	Full data versioning	Docker-based pipelines	CLI + UI	Full data lineage	Immutable data tracking	Strong provenance support, built for reproducibility

Tool / Lineage Platform	Granularity	ML-Specific Features	Integration (ML/Data)	Visualization	Versioning	Compliance Support	Notes
LakeFS	Object-store (S3)-level	Not ML-specific	Git-like data control for lakes	CLI + integrations	Branches, commits	Audit logs, access control	Best for data versioning at scale, Git-like semantics

Key Capabilities Evaluated:

- **Lineage Depth**: What level of data/model history is captured (job, dataset, parameter, etc.)
- **ML Awareness**: Does the tool understand and support ML-specific elements (models, hyperparameters)?
- **Integration Scope**: How well does it integrate with existing tools (Airflow, TensorFlow, etc.)?
- **Visualization**: Can users easily explore lineage?
- **Versioning**: Tracks versions of datasets, models, configs.
- **Compliance / Governance**: Support for audit, roles, traceability.

Summary:

- **Best for End-to-End ML Pipelines**: **Kubeflow Pipelines**, **MLflow**
- **Best for Data Lineage & Governance**: **Apache Atlas**, **DataHub**, **OpenLineage**
- **Best for Full Provenance & Versioning**: **Pachyderm**, **LakeFS**
- **Best for Experiment Tracking**: **Weights & Biases**, **Neptune.ai**

III. METHODOLOGY

To realize the benefits of AI-driven lineage, we propose a five-component methodology that integrates technical tracking with ethical oversight:

1. Lineage Instrumentation
Embed tracking capabilities into every stage of the AI pipeline—from data ingestion, feature selection, model training, to inference. Tools like **OpenLineage** or **Pachyderm** can automate this process.

2. Metadata and Versioning Layer
Use version control (e.g., DVC, Git) to record changes in data, code, and models. Include metadata for source attribution, transformation logs, and environment configurations.

3. Fairness and Bias Auditing
Incorporate fairness metrics and data distribution visualizations at each step. Flag potentially biased transformations or imbalanced training data during lineage capture.

4. Cryptographic Verification
Use hash functions or blockchain to make lineage records tamper-evident and verifiable, particularly important for audit and legal defense.

5. Governance Integration
Connect lineage data with dashboards for regulators, ethics boards, and internal reviewers. Map lineage artifacts to compliance frameworks like **NIST RMF** or **EU AI Act** standards.

FIGURE: AI-Driven Lineage Framework

A layered AI pipeline visual:

- **Data Sources → ETL/Preprocessing → Feature Engineering → Model Training → Validation → Deployment → Monitoring**
- Overlaid with:
- **Lineage Capture Nodes** at each stage
- **Audit Trails** and **Bias Indicators**
- **Cryptographic Checks**
- **Governance Dashboard Feed**

IV. CONCLUSION

In the age of AI, where systems increasingly influence society's most important decisions, **lineage is no longer optional—it is foundational**. AI-driven lineage systems offer a powerful approach to ensure transparency, fairness, and accountability by tracking the complete journey of data and models through the AI lifecycle.

This paper has demonstrated that lineage is more than a technical feature—it is a strategic imperative. When aligned with ethical frameworks and regulatory standards, it becomes a tool for **building trust** in AI systems. From identifying sources of bias to verifying compliance, AI-driven lineage bridges the gap between **ethical AI theory and operational reality**.

While current tools offer partial solutions, our proposed framework pushes for full integration—combining data and model lineage, real-time tracking, governance mapping, and cryptographic verification. This holistic approach supports a future where AI is not only intelligent but **understandable, explainable, and just**.

As regulatory landscapes evolve and public expectations grow, organizations that prioritize AI-driven lineage today will be best positioned to deliver **responsible and robust AI systems** tomorrow.

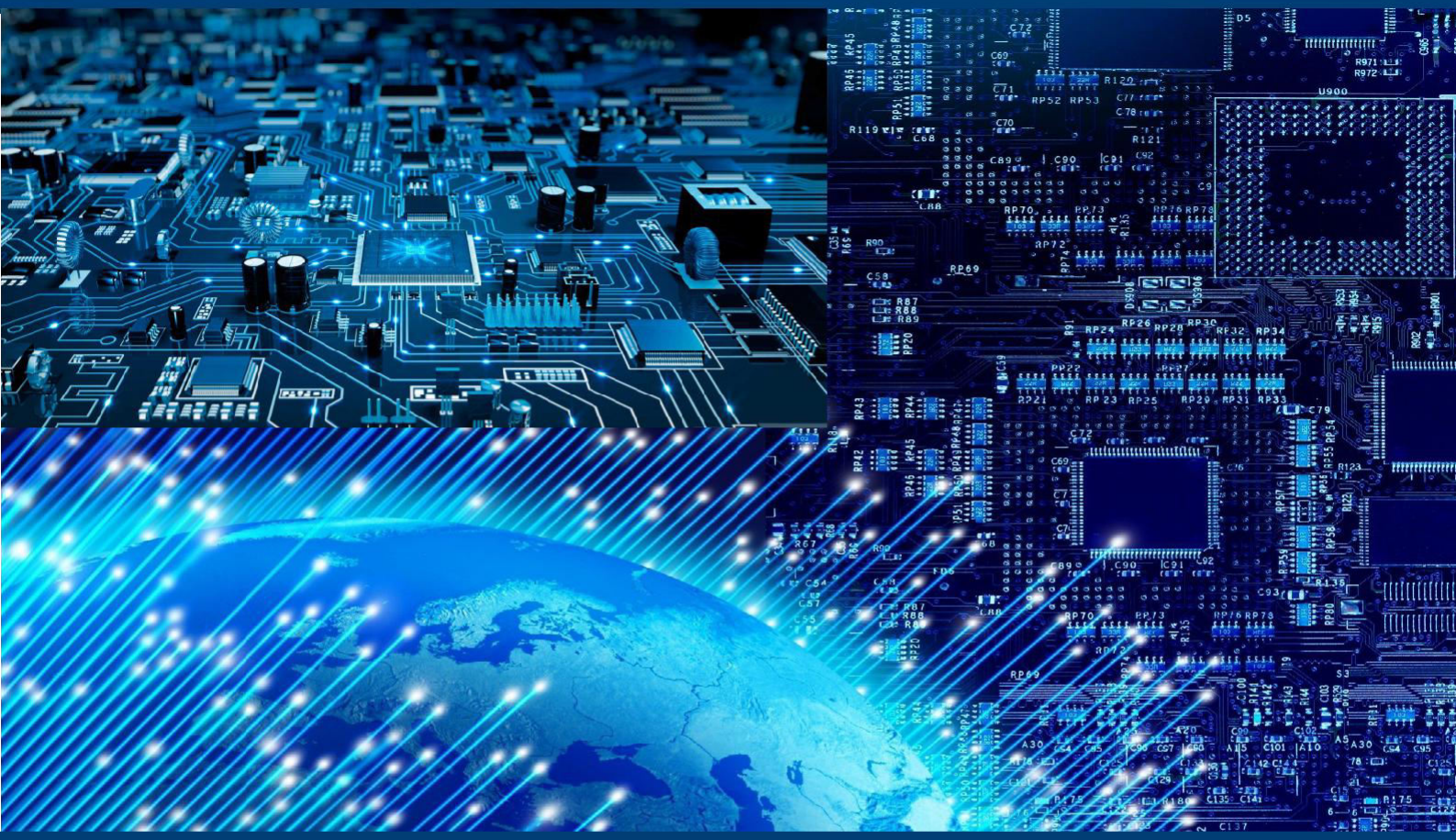
REFERENCES

1. Moreau, L., et al. (2011). The Open Provenance Model core specification. Future Generation Computer Systems, 27(6), 743–756.
2. Mahant, R., & Bhatnagar, S. (2024). Strategies for Effective E-Governance Enterprise Platform Solution Architecture. Strategies, 4(5).
3. Muniraju Hullurappa, Sudheer Panyaram, "Quantum Computing for Equitable Green Innovation Unlocking Sustainable Solutions," in Advancing Social Equity Through Accessible Green Innovation, IGI Global, USA, pp. 387-402, 2025.
4. Madhusudan Sharma, Vadigicherla (2024). Enhancing Supply Chain Resilience through Emerging Technologies: A Holistic Approach to Digital Transformation. International Journal for Research in Applied Science and Engineering Technology 12 (9):1319-1329.
5. Talati, D. (2023). Quantum minds: Merging quantum computing with next-gen AI.
6. Gebru, T., et al. (2018). Datasheets for datasets. arXiv:1803.09010.
7. Holland, S., et al. (2018). The dataset nutrition label. arXiv:1805.03677.
8. NIST. (2023). AI Risk Management Framework 1.0.
9. Apache Atlas. (2023). https://atlas.apache.org
10. OpenLineage. (2024). https://openlineage.io
11. DataHub. (2024). https://datahubproject.io
12. Schelter, S., et al. (2018). Automatically tracking metadata and provenance of machine learning experiments. Data Engineering Bulletin.
13. Bhatnagar, S. &. (2024). Unleashing the Power of AI in Financial Services: Opportunities, Challenges, and Implications. Artificial Intelligence (AI). 4(1).
14. Seethala, S. C. (2024). AI-Infused Data Warehousing: Redefining Data Governance in the Finance Industry. International Research Journal of Innovations in Engineering & Technology, 5(5), Article 028. https://doi.org/10.47001/IRJIET/2021.505028
15. Pachyderm. (2024). https://www.pachyderm.io
16. MLflow. (2023). https://mlflow.org
17. Pareek, C. S. Test Data Management Trends: Charting the Future of Software Quality Assurance.
18. EU Commission. (2023). EU Artificial Intelligence Act.
19. Comet ML. (2023). https://www.comet.com

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 8.028 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 12, Issue 3, May - June 2025 |

20. Koshy, R., et al. (2022). Data governance and lineage in regulated AI systems. Journal of Data and Information Quality, 14(3).
21. Madhusudan Sharma, Vadigicherla (2024). Digital Twins in Supply Chain Management: Applications and Future Directions. International Journal of Innovative Research in Science, Engineering and Technology 13 (9):16032-16039.
22. Bhatnagar, S. (2025). COST OPTIMIZATION STRATEGIES IN FINTECH USING MICROSERVICES AND SERVERLESS ARCHITECTURES. Machine Intelligence Research, 19(1), 155-165.
23. Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).
24. Microsoft. (2022). Responsible AI Standard.
25. Mahant, R. (2025). ARTIFICIAL INTELLIGENCE IN PUBLIC ADMINISTRATION: A DISRUPTIVE FORCE FOR EFFICIENT E-GOVERNANCE. ARTIFICIAL INTELLIGENCE, 19(01).



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |