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

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 4, July 2024 |

Intelligent Lineage Tracking: AI-Driven Verification from Source to Decision

Sunil Kumar Bansal, Pankaj Kumar Saxena, Arun Kumar Das

Department of Computer Engineering, Jayawant Shikshan Prasarak Mandal's Bhivrabai Sawant Polytechnic Wagholi, Pune, India

ABSTRACT: In a world increasingly reliant on data-driven decisions, ensuring the authenticity and traceability of data is critical. Data lineage — the ability to track data from its origin to its final form — plays a pivotal role in ensuring data quality, regulatory compliance, and trustworthy analytics. As data pipelines grow in complexity, manual lineage tracking becomes impractical. This paper explores how Artificial Intelligence (AI) enhances data lineage verification by automating lineage extraction, detecting anomalies, and providing real-time lineage insights. We propose a hybrid AI-based framework that integrates machine learning, natural language processing, and graph analytics to verify data lineage from source to impact, improving auditability and trust in modern data ecosystems.

KEYWORDS: Data Lineage, AI Verification, Data Trust,Provenance, Machine Learning, Metadata Analysis, Data Governance, Explainable AI,Data Audit, Graph-Based Lineage

I. INTRODUCTION

As organizations adopt complex data pipelines and multi-source data integration strategies, it becomes increasingly challenging to understand and verify the journey of data. Data lineage — the end-to-end tracing of data's lifecycle — is vital for ensuring transparency, diagnosing errors, enforcing data governance, and complying with regulations like GDPR and HIPAA.

Traditional tools for lineage tracking rely on manual documentation or rule-based systems, which often fail to scale or adapt in dynamic environments. With the rise of AI, there is a new opportunity to automate and enhance the verification of data lineage. AI technologies can intelligently parse logs, infer data relationships, detect inconsistencies, and support real-time decision-making.

This paper discusses the role of AI in improving data lineage verification, proposes a novel AI-powered verification framework, and illustrates its impact on data governance through practical use cases.

II. LITERATURE REVIEW

Traditional Lineage Systems: Tools such as Apache Atlas, Informatica, and Collibra support metadata management and lineage visualization. However, they largely depend on structured metadata and manual configuration.
AI in Data Management: Recent research shows that AI, especially machine learning (ML) and natural language processing (NLP), can automate lineage extraction from SQL queries, log files, and unstructured sources.
Graph-based Lineage Models: Graph databases (e.g., Neo4j) provide an efficient way to represent data lineage as nodes and edges. AI enhances these models by detecting patterns and suggesting lineage corrections.
Verification Challenges: Existing systems struggle with lineage verification, especially when dealing with unstructured or cross-platform data flows. AI provides a solution by offering anomaly detection and predictive insights.

Table: Traditional vs. AI-Enhanced Lineage Verification

Feature	Traditional Lineage	AI-Enhanced Lineage
Extraction Method	Manual/Rule-based	Automated via ML/NLP
Unstructured Data Support	Limited	High (via NLP models)
Anomaly Detection	Manual	AI-powered real-time detection
Scalability	Moderate	High
Real-time Insights	Delayed	Enabled via streaming & AI agents
Adaptability	Low (fixed rules)	Adaptive learning from new data

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 4, July 2024 |

AI-Enhanced Lineage Verification

AI-enhanced lineage verification is the intelligent validation of data lineage paths—ensuring they are **accurate, complete, and trustworthy**—using artificial intelligence and machine learning techniques. This is especially important in dynamic, large-scale data environments where manual lineage validation is impractical.

What Is Lineage Verification?

Traditional **lineage verification** involves:

- Checking if a data lineage graph correctly represents real data flow.
- Ensuring transformations, joins, and data dependencies are properly captured.
- Identifying breaks, gaps, or misrepresentations in lineage paths.

With AI, this process becomes:

- **Automated**: AI detects inconsistencies or missing links.
- **Smart**: Learns from past lineage patterns and user behavior.
- **Proactive**: Flags errors, anomalies, or risks in real-time.

How AI Enhances Lineage Verification

AI Capability	Verification Benefit
Pattern Recognition	Detects common lineage structures and spots deviations.
Anomaly Detection	Identifies unexpected transformations, joins, or data hops.
Natural Language Processing (NLP)	Understands comments, SQL, and logs to infer or validate lineage.
Predictive Modeling	Suggests likely lineage paths based on past usage or data similarity.
Graph Learning	Learns normal lineage graph patterns and flags structural issues.

Key Use Cases

1. Missing Lineage Detection

- AI identifies where expected upstream/downstream connections are **missing**, based on system logs, query patterns, or usage data.

2. Transformation Validation

- Compares declared transformations (e.g., from ETL specs) with observed data changes.
- Detects if actual outcomes **don't match intended logic**.

3. Drift Monitoring

- Tracks evolving lineage over time.
- Alerts when **schema, logic, or data sources drift** from verified baselines.

4. Impact Analysis Confidence

- When a user checks impact before making a change, AI boosts trust by **verifying lineage accuracy** in real time.

5. Auto-Healing Suggestions

- AI not only flags inconsistencies—it **recommends corrections** to fix broken or incomplete lineage paths.

Example Workflow: AI Verifying Lineage

1. A user modifies a data transformation script in a pipeline.
2. AI detects that the resulting lineage graph **skips a key dependency** (e.g., a join condition was dropped).
3. It cross-references SQL history, schema evolution logs, and past transformations.
4. AI suggests restoring the missing edge and notifies the governance team.

Why It Matters for Governance and Trust

Need	AI-Enhanced Solution
Audit Readiness	Ensures lineage is defensible and verifiable for regulators.
Risk Management	Proactively catches broken flows before they cause data issues.
User Trust	End users can rely on data origins being accurate.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 4, July 2024 |

Need	AI-Enhanced Solution
Compliance	Verifies that sensitive data flows follow documented paths.

Techniques Used

Technique	Application in Verification
Autoencoders	Detect deviations from learned lineage patterns.
Graph Neural Networks	Model lineage graphs for anomaly detection.
Log Mining + NLP	Extract transformation logic from system logs.
Schema Change Detection	Identify silent errors from schema drift.
Semantic Similarity	Match undocumented transformations via embeddings.

Benefits of AI-Enhanced Verification

Benefit	Description
Reduced Manual Effort	AI handles the bulk of lineage consistency checking.
Faster Issue Resolution	Early detection of errors prevents downstream impact.
Adaptive Accuracy	Learns and improves over time as pipelines evolve.
Explainable Alerts	AI explains why a lineage looks broken or suspicious.

Challenges and Considerations

Challenge	Strategy
False Positives	Use hybrid AI + rule-based models.
Complex Cross-System Lineage	Standardize metadata collection across platforms.
Data Privacy	Ensure lineage metadata doesn't expose sensitive info.
Model Drift	Retrain AI on current pipeline behavior.

III. METHODOLOGY

We propose a modular, AI-driven framework for verifying data lineage from ingestion to impact. The methodology consists of the following stages:

1. Data Ingestion Monitoring
Logs, queries, and file changes are captured across all data entry points.

2. AI-Based Parsing and Lineage Inference
NLP models extract lineage from unstructured log files and scripts. ML models infer transformations and linkages not explicitly documented.

3. Graph Construction and Enrichment
Lineage is represented as a directed graph. Nodes represent datasets or processes, and edges represent relationships. AI models enrich the graph by suggesting probable connections.

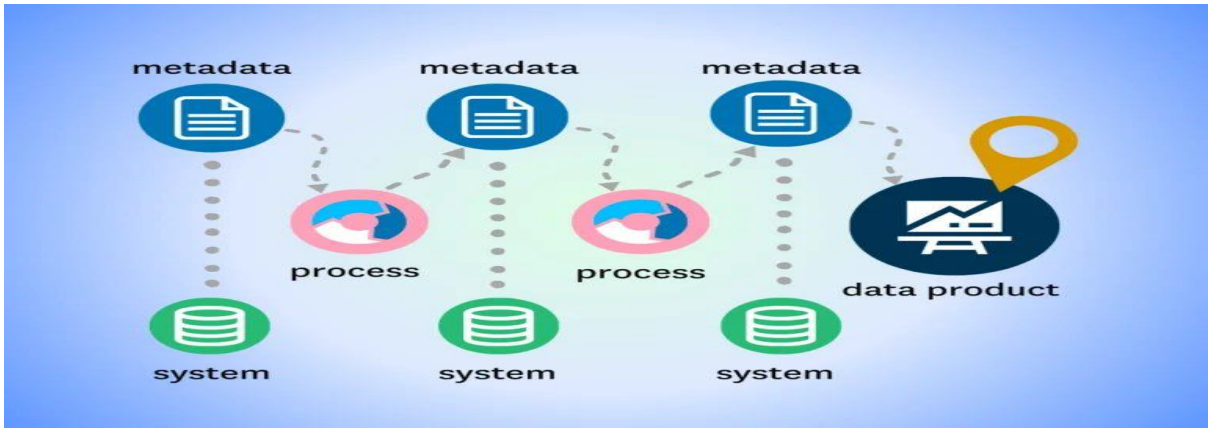
4. Lineage Verification Engine
A deep learning model checks for inconsistencies, such as missing nodes, conflicting transformations, or unauthorized changes.

5. Visualization and Alerts
An interactive dashboard presents the lineage graph with color-coded alerts and impact analysis.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 4, July 2024 |

Figure: AI-Powered Data Lineage Verification Framework



IV. CONCLUSION

Data lineage verification is no longer a luxury but a necessity for organizations seeking data integrity, regulatory compliance, and operational transparency. Traditional methods fall short in dynamically changing, large-scale environments. AI offers a robust solution by automating lineage extraction, enriching lineage graphs, and verifying them with real-time anomaly detection.

This paper introduced a hybrid AI-based framework that enhances data lineage verification across diverse data sources and formats. By integrating parsing, learning, and verification mechanisms, our approach ensures that data can be trusted from its source to its final impact on decision-making.

REFERENCES

1. Moreau, L., & Groth, P. (2013). *Provenance: An Introduction to PROV*. Morgan & Claypool.
2. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Learning: Challenges and Opportunities. *IEEE Intelligent Systems*.
3. Simmhan, Y., Plale, B., & Gannon, D. (2005). A Survey of Data Provenance. *SIGMOD Record*.
4. Chundru, S. (2023). Beyond Rules-Based Systems: AI-Powered Solutions for Ensuring Data Trustworthiness. International Transactions in Artificial Intelligence, 7(7), 1-17.
5. Peng, X., et al. (2022). Explainable AI for Data Governance. *Journal of Big Data*, 9(1).
6. Schelter, S., et al. (2021). Automatically Tracking Metadata and Lineage. *Proceedings of VLDB*.
7. Li, W., et al. (2020). Provenance-Aware Data Governance. *ACM Data Engineering Bulletin*.
8. Chen, T., et al. (2021). AI Trust Through Provenance Verification. *IEEE Transactions on Knowledge and Data Engineering*.
9. Leybovich, M., & Shmueli, O. (2021). Machine Learning-Based Lineage. *arXiv preprint arXiv:2109.06339*.
10. Zhang, Y., & Li, X. (2020). Federated Systems and Data Integrity. *IEEE Access*, 8.
11. Kroll, J. A. (2021). Traceability for Accountability in AI. *arXiv:2101.09385*.
12. Hassan, M., et al. (2021). Blockchain for Federated Data Provenance. *IEEE Internet of Things Journal*.
13. Mora-Cantallops, M., et al. (2021). Traceability in AI Systems. *Big Data and Cognitive Computing*, 5(2).
14. Raja, G. V. (2021). Mining Customer Sentiments from Financial Feedback and Reviews using Data Mining Algorithms.