International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 6.421|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 2, March – April 2024 ||

DOI:10.15680/IJARETY.2024.1102041

Harnessing AI for Transparent and Trustworthy Data Lineage

Gaurav Kumar Deshmukh, Abhishek Kumar Kulkarni

Department of Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India

ABSTRACT: In the age of big data and artificial intelligence (AI), ensuring transparency and traceability of data across its lifecycle has become crucial for ethical, legal, and operational reasons. Data lineage — the ability to track data's origin, movement, transformation, and use — is essential for ensuring data integrity and regulatory compliance. This paper investigates the transformative role of AI in enhancing data lineage systems. It examines how AI technologies, including machine learning, natural language processing (NLP), and graph-based analytics, automate and strengthen the data lineage process. The proposed AI-driven approach enhances visibility, accountability, and reliability in complex, distributed data ecosystems. This research emphasizes AI's potential to not only document data flows but also predict anomalies, enforce compliance, and support decision-making with trusted data.

KEYWORDS: Data Lineage, Artificial Intelligence, Data Provenance, Data Transparency, Explainable AI, Machine Learning, Data Traceability, Metadata Management, Compliance

I. INTRODUCTION

In today's data-intensive world, organizations rely on vast amounts of data to drive decisions, train machine learning models, and deliver services. As data flows through multiple systems, it undergoes transformations that must be transparent and traceable to ensure integrity and trust. Traditional data management tools often fall short when it comes to offering real-time, automated tracking of data movement and changes.

Data lineage provides visibility into the lifecycle of data, from its origin to its consumption. With increasing regulatory requirements (e.g., GDPR, HIPAA), it is no longer sufficient to know where data is — organizations must know how it got there and what happened to it along the way. AI offers promising solutions to enhance lineage by identifying, mapping, and even predicting data flows automatically.

This paper explores how AI technologies can be leveraged to improve the traceability and transparency of data lineage. We argue that AI, with its ability to process complex patterns and automate insights, is a natural fit for the increasingly dynamic and distributed data environments we operate in today.

II. LITERATURE REVIEW

Traditional data lineage systems have been largely rule-based and manually maintained, which limits their scalability and responsiveness. Tools such as Informatica, Apache Atlas, and Collibra provide metadata-based lineage tracking but often require manual integration and constant updates. Moreover, they struggle with unstructured or semi-structured data.

Recent academic research has explored the intersection of AI and metadata management. AI techniques, such as deep learning and knowledge graphs, are being tested to automate lineage detection in data lakes. Natural language processing (NLP) is being used to infer lineage from unstructured log files and SQL queries. Meanwhile, reinforcement learning and anomaly detection algorithms are showing promise in identifying data drift or unauthorized transformations.

A major challenge lies in the opacity of black-box AI models themselves. Researchers in the field of explainable AI (XAI) advocate for data traceability not just in input data but also in model logic and outputs, calling for lineage systems that trace through both data and algorithms.

Table: Comparison of Traditional vs. AI-Driven Data Lineage

Feature	Traditional Lineage Tools	AI-Driven Lineage Systems
Scalability	Limited by manual processes	High — automated tracking
Unstructured Data Support	Minimal or manual	NLP and AI-based extraction
Real-Time Monitoring	Delayed or batch mode	Near real-time processing
Anomaly Detection	Not built-in	Built-in ML anomaly detection
Predictive Capabilities	None	Predictive lineage via ML
Adaptability	Rigid, rule-based	Adaptive to schema changes

AI-Driven Data Lineage

AI-driven data lineage is a next-gen approach to tracing how data **flows, transforms, and evolves** across systems, pipelines, and applications—powered by **machine learning (ML)** and **artificial intelligence (AI)** to enhance automation, accuracy, and usability.

What Is Data Lineage?

Data lineage is the process of tracking the **origins, movement, transformation, and destination** of data across its lifecycle. It answers:

- Where did the data come from? (Source)
- How was it changed? (Transformation)
- Where is it used? (Destination)
- Who touched it and when? (Audit trail)

What Makes It "AI-Driven"?

Traditional lineage relies on static metadata, manual annotations, or rigid rule-based systems. **AI-driven lineage systems go further** by:

AI-Driven Feature	Description
Automated Discovery	AI scans systems to detect undocumented data flows, joins, or transformations.
Lineage Inference	ML infers lineage from logs, SQL queries, data patterns, or even API usage.
Anomaly Detection	AI flags unexpected data changes or suspicious lineage paths (e.g., schema drift).
Semantic Understanding	NLP models extract meaning from labels, fields, and logs to enrich lineage metadata.
Intelligent Visualization	Graph models and AI-driven layouts highlight the most important nodes, impacts, and dependencies.

Core Components of an AI-Driven Data Lineage System

1. Lineage Extractors

- AI agents that **parse code, SQL, data pipelines**, ETL scripts, and APIs.
- Use pattern recognition and language models to extract transformation logic (even in unstructured or semi-structured scripts).

2. Lineage Graph Builder

- Builds a **graph-based view** of datasets, transformations, tables, and dashboards.
- Nodes = data entities; Edges = operations (joins, filters, mappings).

3. ML-Based Inference Engine

- Uses historical patterns to **guess missing lineage connections**, especially in complex or loosely documented systems.
- Can infer dependencies from **query patterns**, **data similarity**, or **schema propagation**.

4. Monitoring & Anomaly Detection

- Continuously monitors lineage paths for unusual patterns (e.g., data appearing in new locations without pipeline changes).
- AI flags anomalies such as:
- Unexpected joins
- Schema changes

- Sudden spikes in downstream data usage

5. Search & Explanation Interface

- Allows users to ask:
- "Where does this number come from?"
- "Who transformed this field and when?"
- AI assists by **ranking the most relevant lineage paths** and **explaining transformations in plain language**.

How AI Enhances Each Phase

Phase	**AI Contribution**
Capture	Automatically detects undocumented flows, ETL patterns, schema changes
Construct	Infers relationships and joins from observed usage patterns
Visualize	Builds dynamic, interactive lineage graphs with impact analysis
Monitor	Detects anomalies or drift in lineage paths or transformations
Explain	Uses NLP to describe transformations in business-friendly terms

Use Cases

1. Regulatory Compliance (e.g., GDPR, HIPAA)

- Trace personal data from input to output
- Validate data masking and access controls
- Audit historical changes to sensitive fields

2. Root Cause Analysis

- Quickly pinpoint where data went wrong in a complex pipeline
- AI can highlight recent transformation steps affecting a specific field or report

3. Impact Analysis

- Before changing a schema or data source, AI helps you see **what downstream systems** will be affected

4. Model Auditing in ML Pipelines

- Trace which version of training data, features, and transformations led to a specific model outcome

Example Workflow: AI-Driven Lineage in Action

1. Data analyst runs an ad-hoc SQL transformation not documented in ETL.
2. AI detects the new transformation by analyzing query logs and compares input-output relationships.
3. System infers that a new data field has entered a downstream dashboard.
4. AI updates the lineage graph and alerts the governance team.
5. If the field contains PII, the AI recommends tagging it and applying policy-based controls.

Popular Tools Using AI for Data Lineage

Tool/Platform	**AI Capabilities in Lineage**
Alation	Behavioral AI for usage-based lineage inference
Collibra	Automated lineage extraction and anomaly detection
Microsoft Purview	Intelligent classification + lineage across Azure services
Amundsen (Lyft)	Open-source with ML integration for metadata and lineage
DataHub (LinkedIn)	AI-driven metadata propagation and lineage graph

Benefits of AI-Driven Data Lineage

Benefit	Description
Automation at Scale	Reduces need for manual lineage documentation
Real-Time Visibility	Continuously updates lineage with live data
Improved Trust	Increases confidence in data by showing full origin story
Faster Debugging	Speeds up resolution of data issues
Compliance Support	Streamlines audits and data governance checks

Challenges to Consider

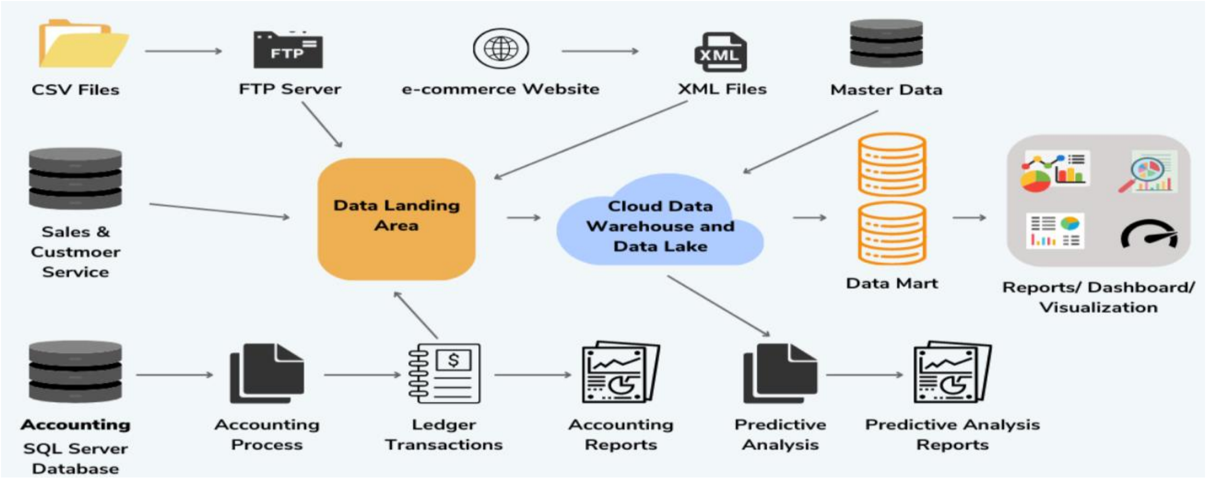
Challenge	Mitigation Strategy
Incomplete Metadata	Use ML inference + active feedback loops
Cross-system Integration	APIs, connectors, and common lineage formats (e.g., OpenLineage)
Privacy and Access	Role-based access and lineage anonymization
Model Drift	Monitor lineage accuracy over time

III. METHODOLOGY

The proposed methodology outlines the integration of AI into a comprehensive data lineage system that emphasizes automation, accuracy, and compliance:

1. **Metadata Collection**: Data is first collected from various sources (databases, APIs, data lakes) along with metadata, logs, and transformation scripts.
2. **AI-Based Parsing**: NLP algorithms analyze scripts, SQL queries, and log files to infer data sources, transformation logic, and destinations.
3. **Graph-Based Lineage Mapping**: AI constructs a knowledge graph where each node represents a data entity or process, and edges represent lineage connections. This graph continuously updates with system changes.
4. **Anomaly Detection Module**: Machine learning models are trained to detect abnormal data flow behaviors, such as unexpected schema changes, unauthorized access, or data drift.
5. **Visualization and Reporting**: A user interface displays the lineage graph, highlights risks, and provides detailed data provenance paths for compliance auditing.
6. **Feedback Loop for Continuous Learning**: Human validation is incorporated where AI uncertainly classifies lineage, improving the model over time through feedback.

Figure: AI-Driven Data Lineage System Architecture



IV. CONCLUSION

AI is reshaping the way organizations manage and understand data lineage. By leveraging AI tools for parsing, modeling, and anomaly detection, data lineage systems can move from passive documentation tools to intelligent agents capable of providing actionable insights and ensuring data trust. AI-driven data lineage not only improves operational efficiency but also supports transparency, compliance, and governance in increasingly complex data environments.

Future work includes integrating AI lineage systems with blockchain for immutable recordkeeping and extending the methodology to include model lineage, thereby offering end-to-end visibility across data and algorithms.

REFERENCES

1. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M.-A. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. Big Data and Cognitive Computing, 5(2), 20. https://doi.org/10.3390/bdcc5020020
2. P Pulivarthy, S Semiconductor, IT Infrastructure.(2023), " ML-driven automation optimizes routine tasks like backup and recovery, capacity planning and database provisioning ", Excel International Journal of Technology, Engineering and Management", 10, 1_Page_22-31
3. Goriparthi, S. (2023). Tracing Data Lineage with Generative AI: Improving Data Transparency and Compliance. International Journal of Artificial Intelligence & Machine Learning, 2(1), 155–165. https://iaeme.com/Home/article_id/IJAIML_02_01_015
4. Leybovich, M., & Shmueli, O. (2021). ML Based Lineage in Databases. arXiv preprint arXiv:2109.06339. https://arxiv.org/abs/2109.06339
5. Kroll, J. A. (2021). Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. arXiv preprint arXiv:2101.09385. https://arxiv.org/abs/2101.09385
6. Patel, S., Rahevar, M., & Parmar, M. (2020). Data Provenance and Data Lineage in the Cloud: A Survey. International Journal of Advanced Science and Technology, 29(05), 4883–4900. https://sersc.org/journals/index.php/IJAST/article/view/13882
7. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M.-A. (2021). Traceability for Trustworthy AI: A Review of Models and Tools. Big Data and Cognitive Computing, 5(2), 20. https://doi.org/10.3390/bdcc5020020