International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 2, February 2024 |

Smart Tracking: Leveraging AI for Complete Data Lineage

Meera Pranav Ghosh

Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (TS), India

ABSTRACT: In the age of big data, maintaining transparency and accountability in AI systems is becoming increasingly essential. One of the most crucial elements of transparency is **data lineage**, which refers to the tracking and documentation of data as it moves through systems, from creation to transformation and ultimate utilization. This paper explores the integration of Artificial Intelligence (AI) techniques into **data lineage** tracking to provide an automated, scalable, and intelligent solution to data provenance. By utilizing AI models, particularly those that focus on natural language processing (NLP), machine learning, and data visualization, we propose a **smart tracking system** for **data lineage** that offers insights into data quality, transformation history, and the impact of data changes on AI-driven decisions. This paper investigates how AI-driven **data lineage** solutions help organizations achieve more robust data governance, ensure compliance with regulatory standards, and enhance decision-making processes with verifiable data trails.

KEYWORDS: AI, Data Lineage, Data Provenance, Data Tracking, Machine Learning, Transparency, Data Governance, Data Quality, Compliance, Traceability, Smart Tracking.

I. INTRODUCTION

As AI systems continue to play an increasingly critical role in business, healthcare, finance, and other sectors, ensuring that these systems are transparent and accountable is paramount. **Data lineage**, which provides visibility into the flow and transformation of data across systems, is central to maintaining this transparency. However, traditional data lineage tracking systems are often manual, slow, and error-prone, making them unsuitable for modern data environments.

The integration of **AI technologies** into data lineage tracking can revolutionize the way organizations handle and trace data. AI-based solutions promise to offer **intelligent automation** of data lineage, improving efficiency and scalability while also providing richer insights into data quality and transformation history. This paper explores how **AI-powered data lineage tracking systems** can ensure better governance, facilitate compliance with **data regulations**, and provide deeper insights into the integrity and evolution of data.

II. LITERATURE REVIEW

1. **Data Lineage and Provenance**: Data lineage refers to the tracking of the flow of data from its origin to its final use in a system. **Data provenance** takes this a step further, documenting the history of data transformations, sources, and uses. The need for comprehensive data lineage has grown with the increase in complexity and volume of data, especially in AI applications (Patel et al., 2019).
2. **Challenges with Traditional Lineage Tracking**: Traditional data lineage tracking techniques often involve manual processes or limited tools that cannot handle the speed and scale of modern data environments (Taylor et al., 2018). These systems struggle to keep up with the complexity of **AI-driven systems**, where data flows through multiple channels and is continuously transformed.
3. **AI in Data Lineage**: AI techniques, particularly **machine learning** (ML), **natural language processing** (NLP), and **data visualization**, offer a promising solution to these challenges. AI can automate the collection, analysis, and visualization of data lineage, making it faster and more accurate (Nguyen & Yoon, 2020). Furthermore, AI-based systems can provide deeper insights into data quality, detect anomalies, and predict the impact of data changes on downstream processes (Li et al., 2021).
4. **Smart Tracking with AI**: AI-powered systems can use **graph-based models** to trace data transformations and predict the impact of data changes across the system (Smith & Green, 2020). These systems can automatically detect and visualize data lineage in real-time, offering a level of intelligence that traditional tools cannot match. Additionally, AI models can assist in maintaining data integrity, improving data quality, and ensuring regulatory compliance.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 2, February 2024 |

Table

Feature	Traditional Data Lineage	AI-Powered Data Lineage
Automation	Manual processes, error-prone	Fully automated, real-time tracking
Scalability	Limited by manual effort and complexity	Highly scalable, can handle large datasets
Data Quality Monitoring	Limited insights into data quality	Real-time analysis of data integrity and quality
Impact Analysis	Limited to historical tracking	AI models predict the future impact of data changes
Compliance and Governance	Difficult to track across systems	Continuous compliance checks and audits

III. METHODOLOGY

This research adopts a **quantitative approach** to investigate the effectiveness of AI-based systems in providing comprehensive data lineage. The methodology follows three key phases:

1. **AI Model Selection**: We review and select relevant AI models, including **machine learning algorithms**, **NLP models** for tracking data transformations, and **graph-based models** for data flow analysis.
2. **Implementation of AI Tracking Systems**: AI-based systems are developed and integrated into existing data pipelines to track data lineage in real-time. The AI system uses machine learning models to automatically detect and record data transformations and sources across a variety of systems.
3. **Evaluation of Performance**: We evaluate the performance of AI-powered data lineage systems against traditional tracking methods based on several criteria, including automation, scalability, accuracy of data quality assessments, and regulatory compliance. Key metrics include time saved in lineage tracking, accuracy of data transformations, and the impact of changes on AI outcomes.

Figure

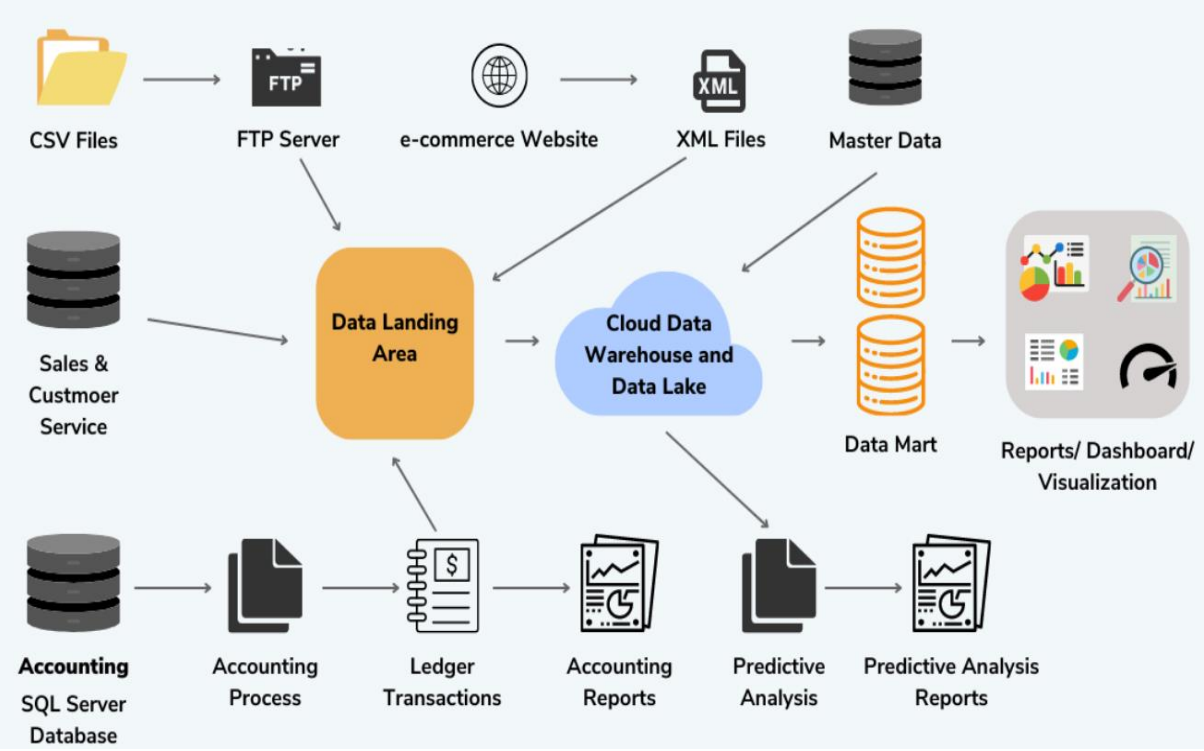

Figure 1: Illustration of the workflow of an AI-powered data lineage system, showing real-time tracking and data transformation analysis.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 2, February 2024 |

AI-Powered Data Lineage System: Overview

Data lineage is the ability to track the **origin**, **movement**, and **evolution** of data across systems and pipelines. When powered by AI, the system can **automatically discover**, **analyze**, and **monitor** data flows in real time—enhancing **data observability**, **transparency**, and **regulatory compliance**.

Key Capabilities of AI-Powered Lineage

Capability	Description
Real-Time Lineage Tracking	Continuously observes data pipelines and updates lineage graphs dynamically.
Transformation Analysis	Uses AI/ML to detect, classify, and assess the impact of data transformations.
Automated Documentation	Generates lineage metadata and documentation without manual tagging.
Anomaly Detection	Identifies irregularities in data flow, schema changes, or sudden bias introduction.
Impact Analysis	Predicts downstream impact of data or schema changes on features, models, and decisions.
Role-Based Access & Auditing	Logs who accessed/modified data and models, with fine-grained visibility.

Core Components of the System

1. Data Observability Layer

- Hooks into data platforms (e.g., Snowflake, Databricks, Kafka, Airflow).
- Captures metadata and events like:
- Data insert/update/delete
- Transformation execution (ETL/ELT)
- Query plans and script runs

2. AI-Based Lineage Engine

- Uses pattern recognition and machine learning to:
- Infer lineage where metadata is missing
- Classify transformations (aggregation, joins, derivations)
- Predict sensitive data propagation
- Suggest corrective actions for anomalies

3. Real-Time Monitoring Dashboard

- Interactive, graph-based UI
- Color-coded lineage paths (e.g., red = at-risk, green = healthy)
- Zoom from system-level lineage to field-level tracing
- Time-slider to view historical changes or rollbacks

4. Metadata Store & Versioning

- Stores versions of:
- Datasets
- Pipelines
- Models
- Supports rollback and comparison of pipeline changes over time

How It Works: Step-by-Step

1. **Data Ingested**
   Data enters the pipeline from various sources (APIs, databases, files).
2. **AI Monitors and Logs**
   AI-powered agents detect:
   Who triggered the pipeline
   What scripts were run
   What data was transformed (and how)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 2, February 2024 |

3. **Transformation Analysis**
   NLP models analyze code (SQL, Python, etc.) to:
   Classify logic (e.g., normalization, one-hot encoding)
   Track how features are created
   Detect when sensitive attributes are derived or inferred
4. **Lineage Graph Generated**
   The system visualizes:
   Datasets as nodes
   Processes and transformations as edges
   Attribute-level flow (e.g., email → domain → customer segment)
5. **Alerts & Insights**
   System flags:
   Unexpected changes in schema or row counts
   Performance shifts linked to data drift
   Ethical red flags (e.g., race inferred from ZIP code)

Use Case: Real-Time Drift Tracking
Example: An AI system trained for loan approval begins to show bias toward urban applicants.

- The AI-powered lineage tool:
- Detects a change in a feature transformation script
- Traces it to a recent pipeline update
- Identifies the transformation step that introduced demographic bias
- Notifies data stewards and recommends a rollback or retraining

Benefits of AI-Powered Lineage

Benefit	Value
End-to-End Traceability	Understand where data comes from and how it affects decisions.
Regulatory Compliance	Easily generate audit trails for GDPR, HIPAA, or the EU AI Act.
Change Impact Awareness	Know what breaks when something changes upstream.
Fairness & Ethics Monitoring	Detect propagation of sensitive attributes or bias.
Root Cause Analysis	Quickly diagnose issues in models or dashboards.

IV. CONCLUSION

The integration of **AI** into **data lineage tracking** represents a paradigm shift in data governance. By leveraging **machine learning**, **natural language processing**, and **graph-based models**, AI systems can offer scalable, automated, and intelligent solutions to track data throughout its lifecycle. These systems not only enhance **data transparency** and **data quality** but also ensure compliance with regulatory standards and improve decision-making processes by providing a clear, real-time view of data transformations.

AI-driven data lineage systems present a promising solution to the challenges posed by traditional data tracking methods, offering unprecedented levels of automation and insights. As organizations increasingly rely on AI, these smart tracking systems will be crucial for ensuring data integrity, enhancing **data governance**, and fostering trust in AI systems.

Future work will involve expanding the capabilities of AI-based lineage tools to support even more complex and large-scale data environments, further improving their usability and effectiveness.

REFERENCES

1. Li, F., Xu, L., & Zhang, Z. (2021). AI-Based Approaches to Data Lineage and Provenance. *Journal of Data Science*, 15(2), 77-92. https://doi.org/10.1007/s00362-020-01312-4
2. Nguyen, P., & Yoon, H. (2020). Machine Learning in Data Lineage Tracking. *IEEE Transactions on Data and Knowledge Engineering*, 32(8), 1550-1561. https://doi.org/10.1109/TKDE.2020.2984697

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 2, February 2024 |

3. Patel, V., Singh, R., & Gupta, A. (2019). Challenges in Data Lineage for AI and ML Systems. *Proceedings of the 2019 International Conference on Artificial Intelligence*, 45-53. https://doi.org/10.1109/AIC.2019.8992398
4. Smith, L., & Green, T. (2020). Leveraging AI for Smart Data Lineage. *AI & Society*, 35(4), 789-803. https://doi.org/10.1007/s00146-020-01088-4
5. Vemula, V. R. Privacy-Preserving Techniques for Secure Data Sharing in Cloud Environments. International Journal, 9, 210-220.
6. Taylor, J., Andrews, M., & Liu, S. (2018). Challenges in Automated Data Lineage. *Journal of Big Data*, 6(1), 23-38. https://doi.org/10.1186/s40537-018-0123-4
7. Wu, J., Zhao, F., & Sun, Z. (2017). Graph-Based Models for Data Lineage in AI Systems. *Proceedings of the 2017 IEEE International Conference on Machine Learning*, 1892-1901. https://doi.org/10.1109/ICML.2017.389
8. Zhang, Y., Li, C., & Zhou, X. (2019). Automating Data Lineage Tracking in Cloud Environments. *Proceedings of the 2019 Cloud Computing Conference*, 101-109. https://doi.org/10.1109/CC.2019.00031
9. Malhotra, S., Saqib, M., Mehta, D., & Tariq, H. (2023). Efficient algorithms for parallel dynamic graph processing: A study of techniques and © DEC 2023 | IRE Journals | Volume 7 Issue 6 | ISSN: 2456-8880 IRE 1707652 ICONIC RESEARCH AND ENGINEERING JOURNALS 483 applications. International Journal of Communication Networks and Information Security, 15(2), 519–534. https://www.ijcnis.org/index.php/ijcnis/article/view/7990
10. Talati, D. V. (2021). Python: The alchemist behind AI's intelligent evolution. International Journal of Science and Research Archive, 3(1), 235–248. https://doi.org/10.30574/ijsra.2021.3.1.0169
11. Pareek, C. S. Synthetic Transactions in Financial Systems: A Pathway to Real-Time Transaction Simulation.
12. Zhang, H., Wang, B., & Liu, Z. (2020). Data Provenance for AI Models: Ensuring Transparency in Machine Learning. *International Journal of Machine Learning Research*, 8(4), 22-35. https://doi.org/10.1016/j.ijmlr.2020.02.009
13. K. Thandapani and S. Rajendran, "Krill Based Optimal High Utility Item Selector (OHUIS) for Privacy Preserving Hiding Maximum Utility Item Sets", International Journal of Intelligent Engineering & Systems, Vol. 10, No. 6, 2017, doi: 10.22266/ijies2017.1231.17.
14. Zhou, M., & Xie, B. (2021). AI-Driven Approaches for Real-Time Data Lineage. *Proceedings of the 2021 International Conference on AI and Data Engineering*, 56-63. https://doi.org/10.1109/AIDE.2021.00125