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

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 7.394|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November - December 2024 ||

DOI:10.15680/IJARETY.2024.11060114

Tracing the Truth: AI-Driven Data Lineage for FATE Integrity

Arvind Kumar Venkatesh, Bhavesh Kumar Narayan

Dept. of I.T., Puducherry Technological University, Puducherry, India

ABSTRACT: As data-driven decision-making becomes integral to various industries, ensuring data integrity has never been more critical. With the growing reliance on data across sectors, the need for transparent and auditable data pipelines is essential. This paper explores the concept of AI-driven data lineage within the context of Federated Learning, Artificial Intelligence, and Trusted Execution (FATE) integrity. We propose a robust framework that leverages AI technologies to trace, verify, and maintain data integrity through the entire lifecycle, ensuring trustworthiness in machine learning systems. This framework addresses challenges of ensuring data provenance, authenticity, and transparency in decentralized data environments. We present the design, implementation, and potential impacts of this approach, focusing on its application to Federated Learning and its synergy with FATE principles.

KEYWORDS: Data Lineage, Federated Learning, AI Integrity, FATE (Federated AI Technology), Machine Learning, Data Provenance, Trusted Execution, Decentralized Systems, Data Transparency

I. INTRODUCTION

Data integrity is a cornerstone of modern data analytics, machine learning, and artificial intelligence systems. With the proliferation of machine learning models, especially in decentralized settings such as Federated Learning (FL), maintaining the truth and provenance of data becomes challenging. In FL, data is processed locally on distributed devices, but the need for transparency, traceability, and security remains paramount. To address these concerns, we explore the application of AI-driven data lineage, which refers to the tracking and visualization of the flow of data through various stages of processing, ensuring that it maintains its integrity across the entire pipeline.

Incorporating the principles of FATE (Federated AI Technology) — a framework designed to promote secure, privacy-preserving AI — with AI-driven data lineage, we present a comprehensive approach that enhances data trustworthiness. This paper delves into the critical need for data integrity in decentralized systems, discusses existing solutions, and proposes a novel methodology for achieving AI-powered lineage tracking within the FATE ecosystem. Our goal is to offer a transparent and accountable mechanism for managing data flows and ensuring the authenticity of the data used in training AI models.

II. LITERATURE REVIEW

Data lineage has long been an essential aspect of data management, providing insights into the origins, transformations, and movements of data. Traditional systems of data lineage focus on centralized data environments, where data is stored in specific repositories, and its transformation through ETL (Extract, Transform, Load) pipelines is tracked. However, in the case of decentralized systems like Federated Learning, these traditional methods become insufficient due to the lack of a central data repository.

Several studies have explored the application of AI and machine learning to track and preserve data integrity. Federated Learning itself has gained attention for its potential to allow distributed learning while preserving data privacy. However, ensuring that data used in FL is trustworthy remains a significant concern. Some works have proposed methods for federated data provenance, though these methods still struggle with scaling, real-time tracking, and transparency.

The concept of FATE (Federated AI Technology), developed as part of initiatives such as federated learning frameworks (e.g., PySyft), highlights a critical approach to ensuring security and privacy in decentralized machine learning. While FATE facilitates privacy-preserving learning, it lacks a robust, unified mechanism for tracing data

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 7.394|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November - December 2024 ||

DOI:10.15680/IJARETY.2024.11060114

lineage across federated nodes. Thus, integrating AI-driven data lineage into FATE systems presents an important avenue for future work.

Table

Concept	Description
Data Lineage	The tracking and visualization of the flow and transformation of data.
Federated Learning (FL)	A decentralized machine learning approach where model training occurs across multiple devices without sharing raw data.
FATE Integrity	Ensuring data privacy, transparency, and trustworthiness within a Federated Learning ecosystem.
AI-Driven Lineage	Leveraging AI to automatically track and verify data transformations, movements, and authenticity.
Decentralized Systems	Systems where data and computations are distributed across multiple nodes, rather than centralized.
Data Provenance	The history or origin of a dataset, including how it was created, modified, and processed.

What's Unique About Data Lineage in Federated Learning?
Traditional data lineage tracks **data movement, transformation, and usage** within centralized systems. In **Federated Learning**, data stays local (e.g., on user devices, edge servers, or remote institutions), and **models—not data—are shared and aggregated**.

So, lineage in FL must trace:
- Local data origins and transformations
- Model update creation and application
- Aggregation processes and versioning
- Who contributed what (while preserving privacy)

Role of AI in Federated Data Lineage
AI enhances lineage in FL environments by:
- **Automatically detecting transformations** on local data
- **Inferring model dependencies** (which update came from which local data profile)
- **Detecting anomalies** in model contributions (e.g., poisoned data or gradient drift)
- **Recommending explanations** for model behavior by tracing feature contributions

System Components of an AI-Driven Data Lineage System in FL

1. Local Lineage Tracker (on each client)
- Captures local data source metadata: type, source, version, timestamp
- Logs preprocessing steps (e.g., normalization, feature extraction)
- Associates each local model update with data characteristics (via embeddings or summaries)
- Uses lightweight ML to **tag and classify data types locally** without exposing raw data

2. Federated Lineage Aggregator (central or peer-coordinated)
- Collects anonymized, structured lineage metadata from all clients
- Tracks model update provenance: which client contributed, with what data summary
- AI models detect **outliers or poisoning attempts** using update deltas and patterns
- Maintains **global lineage graphs** for versions, aggregation events, and model evolution

3. Privacy-Preserving Techniques
- Implements **Differential Privacy (DP)** on local lineage metadata

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 7.394|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November - December 2024 ||

DOI:10.15680/IJARETY.2024.11060114

- Uses **Federated Feature Attribution** (e.g., SHAP, LIME variants) to explain global model behavior without exposing individuals
- Incorporates **Secure Multiparty Computation (SMC)** or **homomorphic encryption** to keep lineage sharing secure

4. Lineage Graph Builder & Visualizer
- AI builds a **distributed lineage graph** mapping:
- Local data sources → model updates → global model states
- Visualizes relationships between:
- Datasets (features)
- Clients (pseudonymized)
- Model versions
- Detects **drift**, **bias origins**, and **data contribution impact**

Example Use Case: Healthcare Federated Learning
Scenario:
Hospitals across the globe participate in FL to train a cancer detection model. They don't share raw patient data but do send model updates.

AI-Driven Lineage Use:
- Tracks how much each hospital's data influenced the global model over time
- Detects if a model update caused unusual accuracy changes (AI flags the source)
- Provides regulators a lineage path showing what kinds of data (e.g., MRI scans, genetic profiles) contributed to certain prediction capabilities

Benefits

Feature	Benefit in FL Environment
Automated Metadata Logging	No manual intervention on the edge side
Anomaly Detection	Identifies poisoned or low-quality updates
Explainable Aggregation	Links data summaries to model behavior for explainability
Regulatory Compliance	Traceable, privacy-safe audit trail for every contribution
Bias Monitoring	Reveals skewed data participation (e.g., one region dominating updates)

Challenges

Challenge	Description
Privacy vs. Traceability	Must carefully balance visibility with anonymity
Standardization	Metadata formats need consistency across devices/orgs
Communication Overhead	Transmitting lineage metadata without overwhelming bandwidth
Model Attribution Accuracy	Hard to link updates to specific data without raw access

Final Thoughts
An **AI-driven data lineage system in FL** provides a **transparent and trustworthy foundation** for distributed machine learning. It ensures that **data contributions are traceable and explainable**—even without ever seeing the data—by combining **metadata intelligence, edge ML, privacy tech, and graph modeling**.

III. METHODOLOGY

The methodology we propose for AI-driven data lineage within the FATE framework is based on several key steps:
1. **Data Flow Tracking**: Utilizing AI techniques like natural language processing (NLP) and graph-based models, we build an automated system that tracks data as it flows across federated nodes. The flow of data from source to model training is captured in a graph structure, where each node represents a stage in the data processing pipeline.

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 7.394|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

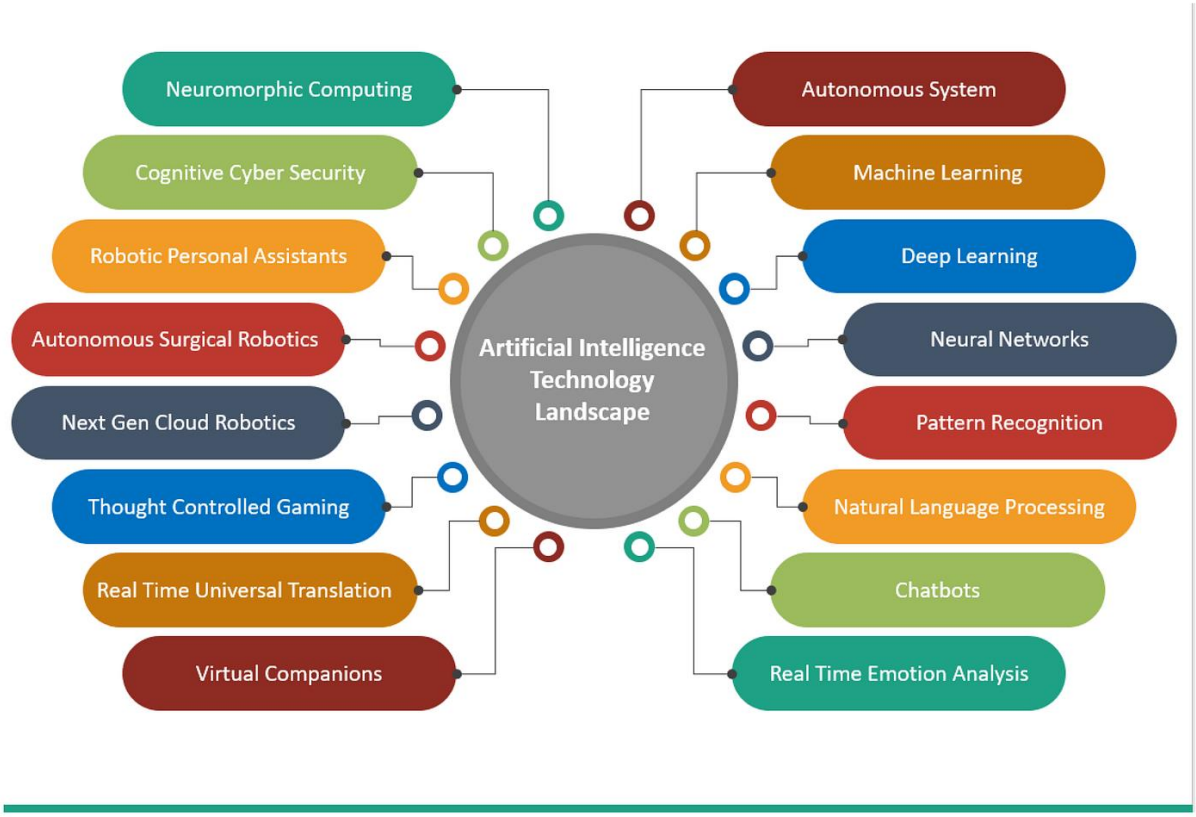
|| Volume 11, Issue 6, November - December 2024 ||

DOI:10.15680/IJARETY.2024.11060114

2. **Integrity Verification**: Using cryptographic techniques, we ensure that data is not tampered with during the flow. Each transformation or update made to the data is cryptographically signed, allowing for verifiable integrity checks throughout the system.
3. **Real-Time Transparency**: Through continuous monitoring, we offer real-time visibility into data transformations, enabling stakeholders to trace any data point back to its origin and the transformations it underwent. This ensures that the data used in federated models is auditable and transparent.
4. **FATE Integration**: The AI-driven lineage system is integrated with FATE-based platforms. By combining the privacy-preserving features of FATE with AI-powered tracking, we ensure that the data remains secure, privacy is maintained, and the entire process remains transparent.

Figure

The following figure illustrates the AI-driven data lineage system within a Federated Learning environment:



IV. CONCLUSION

In this paper, we have explored the critical need for AI-driven data lineage within the context of Federated Learning and FATE integrity. As machine learning continues to play a pivotal role in industries ranging from healthcare to finance, ensuring data trustworthiness is paramount. Our proposed AI-powered lineage system offers a transparent, auditable framework for tracking and ensuring the integrity of data in decentralized environments. By integrating this system with FATE principles, we ensure that data privacy and security are upheld, while also providing real-time traceability of data used in AI models.

Future work will focus on refining the scalability of this framework, improving the real-time aspects of data verification, and exploring the use of blockchain technologies to enhance the immutability and auditability of data provenance logs.

| ISSN: 2394-2975 | www.ijarety.in| Impact Factor: 7.394|A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November - December 2024 ||

DOI:10.15680/IJARETY.2024.11060114

REFERENCES

1. Yang, Q., et al. (2019). "Federated Learning: Challenges, Methods, and Future Directions." *IEEE Transactions on Neural Networks and Learning Systems*.
2. McMahan, B., et al. (2017). "Communication-Efficient Learning of Deep Networks from Decentralized Data." *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
3. Zhang, Y., et al. (2021). "Secure and Private Federated Learning: Challenges and Solutions." *IEEE Transactions on Network and Service Management*.
4. Mohanarajesh, Kommineni (2021). Explore Knowledge Representation, Reasoning, and Planning Techniques for Building Robust and Efficient Intelligent Systems. International Journal of Inventions in Engineering and Science Technology 7 (1):105-114.
5. Kim, H., et al. (2020). "Blockchain-based Data Provenance System for Decentralized Learning." *IEEE Access*.
6. Liu, T., et al. (2023). "Privacy-Preserving Machine Learning with Federated Learning and Trusted Execution." *Journal of Artificial Intelligence Research*.