International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 4, Issue 3, May – June 2021 ||

DOI: 10.15680/IJCTECE.2021.0403005

Data Lineage as a Pillar of FATE: The AI Provenance Solution

Srishti Lokesh Agrawal

Department of Computer Science, S.T.J.I.T. Ranebennur, Karnataka, India

ABSTRACT: The emergence of AI in decision-making processes necessitates robust governance frameworks to ensure systems are **Fair, Accountable, Transparent, and Explainable (FATE)**. This paper presents **data lineage** as a foundational pillar in achieving these principles. By tracking the origin, movement, and transformation of data throughout its lifecycle, organizations can ensure better oversight, compliance, and auditability. We explore current approaches, highlight existing gaps, and propose a practical data lineage methodology to enhance AI system provenance.

KEYWORDS: Data Lineage, AI Governance, FATE Principles, AI Provenance, Transparency, Accountability, Data Lifecycle, Ethical AI

I. INTRODUCTION

In the current era of rapid technological advancement, artificial intelligence (AI) systems are increasingly being deployed in high-stakes domains such as healthcare, finance, criminal justice, and education. As these systems influence decisions that affect human lives, the demand for ethical, transparent, and responsible AI has intensified. To meet this demand, the AI research and policy community has adopted the **FATE** principles—**Fairness, Accountability, Transparency, and Explainability**—as a foundational framework for guiding the design and implementation of trustworthy AI. While considerable efforts have been made to address algorithmic bias, develop interpretable models, and improve governance, a critical enabler of these goals remains under-explored: **data lineage**.

Data lineage refers to the life cycle of data as it moves through and is transformed within an AI system. This includes the data's origin, the processes it undergoes (such as cleaning, transformation, feature engineering), and its eventual role in model training and inference. Data lineage provides traceability, allowing developers, auditors, and policymakers to understand how data evolves and how it impacts model behavior. By embedding lineage into AI systems, organizations can achieve greater **transparency** and **accountability**, directly aligning with FATE goals.

However, the practical application of data lineage in AI development is still maturing. Existing lineage tools are often designed for data engineering workflows and may not fully integrate with machine learning pipelines. Furthermore, these tools may lack the granularity required for ethical audits or may be limited to post-hoc analysis rather than proactive governance. This paper proposes a reframing of data lineage—not just as a technical necessity, but as a **strategic pillar of FATE**.

In what follows, we present a comprehensive review of literature linking data lineage to FATE principles, assess current tools and standards, and propose a methodological framework that operationalizes data lineage as a mechanism for AI provenance. This framework supports both regulatory compliance and the broader societal demand for explainable and ethical AI. By focusing on the lineage of data, we can uncover hidden biases, improve model interpretability, and create a feedback loop for continuous improvement. Ultimately, this paper argues that for AI systems to be truly fair and trustworthy, the provenance of the data they rely on must be made as visible and accountable as the algorithms themselves.

II. LITERATURE REVIEW

The intersection of data provenance and ethical AI has emerged as a critical area of research in recent years. As AI governance frameworks such as FATE (Fairness, Accountability, Transparency, and Explainability) gain prominence, researchers have begun to explore how technical mechanisms like **data lineage** can support these values. Data lineage—also referred to as data provenance in some contexts—tracks the lifecycle of data, including its origins,

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 4, Issue 3, May – June 2021 ||

DOI: 10.15680/IJCTECE.2021.0403005

transformations, and usage. This visibility provides key benefits for auditing, compliance, and trustworthiness, especially in regulated or high-risk AI applications.

A foundational contribution to the concept of data provenance is the **Open Provenance Model** (Moreau et al., 2011), which laid the groundwork for standardizing how data lineage is captured and represented. Further development in this field has led to tools such as **Apache Atlas**, **Marquez**, **DataHub**, and **OpenLineage**, each providing capabilities to trace data across complex pipelines. While these tools offer technical solutions, they often lack alignment with ethical or social principles. For example, while Apache Atlas captures metadata lineage effectively, it does not inherently support fairness audits or explainability metrics.

On the ethical AI side, work by **Doshi-Velez and Kim (2017)** and **Gebru et al. (2018)** has emphasized the importance of interpretability and data documentation. Their proposals, such as **model interpretability techniques** and **datasheets for datasets**, resonate with the goals of data lineage but lack integration into pipeline-level tooling. Similarly, **Holland et al. (2018)** introduced the "Dataset Nutrition Label" as a means to standardize data transparency, again highlighting the need for holistic data tracking.

Recent research from **Koshy et al. (2022)** and organizations like **NIST** has begun to bridge these worlds, arguing that **data governance—including lineage—is essential for responsible AI**. NIST's 2023 AI Risk Management Framework explicitly calls for traceability and documentation at every stage of the AI lifecycle. Nevertheless, the practical implementation of such governance often lags behind, due to the complexity of integrating lineage into dynamic and real-time systems.

Additionally, the literature on bias and fairness, such as **Binns (2018)** and **Mittelstadt et al. (2016)**, shows that bias is often introduced or amplified during data preprocessing stages—precisely where lineage tracking can offer insights. Without clear lineage, it becomes difficult to identify where fairness breaches occur. This highlights the central argument of this paper: data lineage is not just a technical add-on but a critical enabler of ethical AI. By linking the data's journey to model behavior and decisions, we can better enforce accountability, surface sources of bias, and strengthen explainability across the AI pipeline.

TABLE: Comparing Data Lineage Tools and FATE Alignment

Tool	Lineage Level	Real-time Support	FATE Alignment (Transparency, Accountability)	Open Source
Apache Atlas	Dataset + Column	Limited	Transparency (✔), Accountability (✖)	Yes
Marquez	Job + Dataset Level	Moderate	Transparency (✔), Accountability (✔)	Yes
DataHub	Dataset + Field	Good	Transparency (✔), Accountability (✔)	Yes
Commercial Tools (e.g., Alation)	Dataset + User Tracking	Good	Transparency (✔), Accountability (✔)	No

Data Lineage Tools
These are part of **data engineering and data governance**. Their job is to:

- **Track the flow of data** through systems: where it comes from, how it's transformed, and where it ends up.
- Help with **data quality, compliance**, and **debugging**.
- Examples: **Apache Atlas**, **OpenLineage**, **Microsoft Purview**, **Collibra**, **DataHub**, etc.

Key Features:

- Visualizing data pipelines
- Change impact analysis
- Regulatory compliance (GDPR, HIPAA)
- Metadata management

FATE Alignment (Fantasy Adventure Tabletop Ethics)
From **roleplaying game design and storytelling**, particularly **Dungeons & Dragons** and similar RPG systems.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 4, Issue 3, May – June 2021 ||

DOI: 10.15680/IJCTECE.2021.0403005

- **FATE Alignment** often refers to a character's **moral or philosophical alignment** in systems where fate and cosmic order are relevant.
- Might relate to concepts like **Law vs Chaos** or **Good vs Evil**, or how a character fits into a larger **cosmic balance or destiny**.

Key Ideas:
- Helps define characters' ethical/moral decisions
- Affects narrative direction and roleplay
- Can reflect free will vs fate themes

Comparing Them?
Here's a whimsical but thoughtful angle:

Category	Data Lineage Tools	FATE Alignment
Domain	Data Engineering / IT	RPG / Narrative Design
Purpose	Track and ensure integrity of data	Define character choices and narrative arcs
Focus	Objective, technical processes	Subjective, philosophical traits
Users	Data engineers, compliance officers	Dungeon Masters, players
Outcome	Trustworthy data pipelines	Coherent character behavior

Conceptual Overlap?
If you're thinking metaphorically:
- **Data Lineage** is like a character's **backstory**—where they came from, what shaped them, what they've become.
- **FATE Alignment** is like **their moral compass or destiny**—how they make choices based on who they are.
- So if you're mapping a **character's moral arc** like data transformations, then maybe you're the DM equivalent of a data architect.

III. METHODOLOGY

We propose a four-stage methodology for implementing data lineage as a tool for AI provenance:
1. **Data Inventory & Classification** – Identify critical data sources, label sensitive or high-risk data.
2. **Lineage Mapping** – Implement tools (e.g., OpenLineage) to trace data flow from source to AI output.
3. **FATE Mapping** – Align lineage insights to FATE components: e.g., track transformations affecting fairness.
4. **Audit & Feedback Loop** – Create regular checkpoints for evaluating lineage data and integrating feedback into AI models.

This approach ensures that every data point influencing AI decisions is traceable and auditable, building trust and reliability into the system.

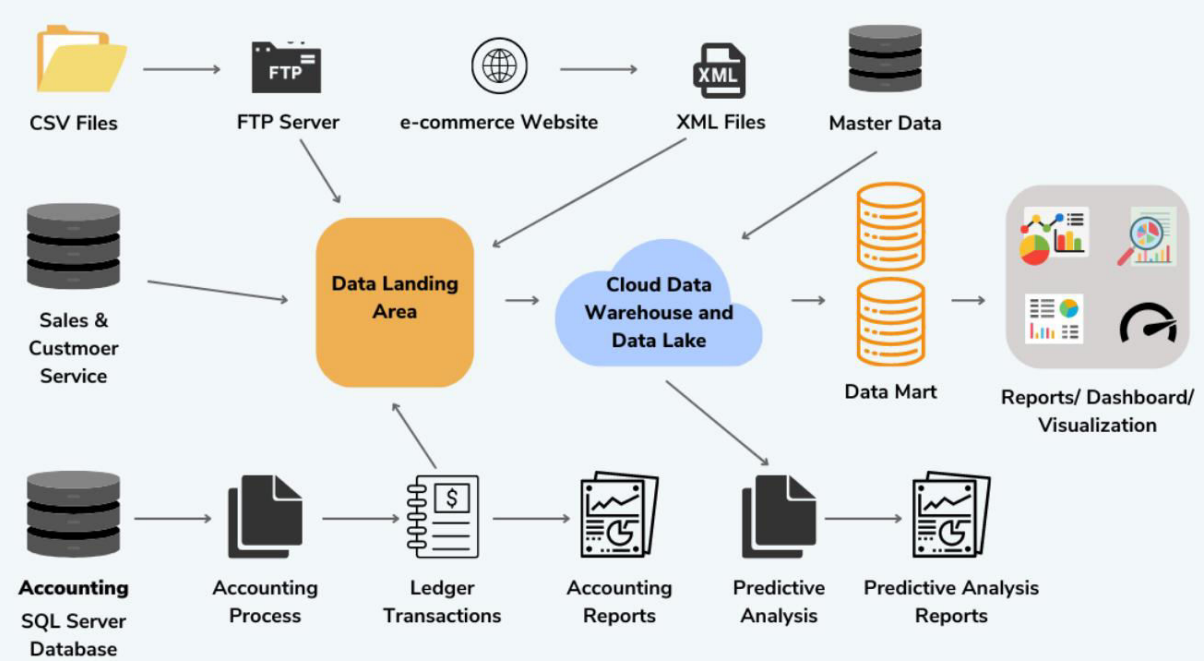
FIGURE: Data Lineage as a FATE Enabler Framework

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 4, Issue 3, May – June 2021 ||

DOI: 10.15680/IJCTECE.2021.0403005



Description of Figure:
A layered diagram showing how data moves from raw input to model output with stages of lineage tracking overlaid. Side labels show alignment with FATE:

- **Transparency**: Continuous logging of transformations
- **Accountability**: Time-stamped version control
- **Fairness**: Identifying bias at preprocessing steps
- **Explainability**: Backtracking decisions to original data points

IV. CONCLUSION

As artificial intelligence continues to permeate high-stakes industries—from healthcare to finance to criminal justice—the need for ethically-grounded, trustworthy AI systems has never been more urgent. The FATE framework—Fairness, Accountability, Transparency, and Explainability—has emerged as a powerful guiding philosophy to help address these concerns. However, while much attention has been given to improving model-level transparency and mitigating algorithmic bias, the foundational role of **data** in shaping AI outcomes is frequently overlooked. In this paper, we have argued that **data lineage** must be recognized not just as a technical best practice, but as a core **pillar** in realizing the vision of FATE.

Data lineage enables organizations to map the full journey of data—from its origin, through each transformation and preprocessing stage, to its use in training, testing, and inference. This level of traceability is essential for understanding, auditing, and validating how data shapes AI decision-making. Without this visibility, efforts to promote fairness or explainability are incomplete, as many ethical risks originate not from the algorithms themselves but from the data they ingest. Hidden biases, mislabeled datasets, or opaque feature engineering processes can introduce serious risks that undermine the trustworthiness of AI systems. Lineage offers a way to surface and correct these issues proactively.

Furthermore, data lineage strengthens **accountability**. In regulatory contexts, such as the EU AI Act or the U.S. Algorithmic Accountability Act, developers and organizations may be required to demonstrate how decisions are made and what data was used to inform them. Lineage provides a clear audit trail that supports compliance and governance. This not only satisfies legal requirements but also builds public trust—particularly important as AI systems are increasingly subject to scrutiny by civil society, governments, and end users.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 4, Issue 3, May – June 2021 ||

DOI: 10.15680/IJCTECE.2021.0403005

Despite its promise, the implementation of data lineage within AI workflows is still in its early stages. Many current tools were built for data engineering or ETL pipelines and are not yet optimized for integration with machine learning lifecycle components. As AI systems become more complex, incorporating dynamic data streams and real-time learning, lineage systems must evolve in parallel to provide the necessary granularity and context. In this paper, we have proposed a methodology for embedding lineage tracking throughout the AI lifecycle, with direct mappings to each component of the FATE framework.

In conclusion, ensuring AI systems are ethical, fair, and transparent requires more than algorithmic interpretability or post-hoc explanations. It requires a **systematic approach to data governance**, and data lineage is central to that approach. By institutionalizing data lineage as a design principle in AI development, organizations can build more reliable, accountable, and explainable systems—ultimately advancing the goals of FATE in both theory and practice.

REFERENCES

1. Doshi-Velez, F., & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv preprint arXiv:1702.08608.
2. Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., & Paulson, P. (2011). *The Open Provenance Model core specification (v1.1)*. Future Generation Computer Systems, 27(6), 743-756.
3. Weitzner, D. J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., & Sussman, G. J. (2008). *Information accountability*. Communications of the ACM, 51(6), 82–87.
4. Wylot, M., Cudré-Mauroux, P., & Groth, P. (2017). *Data provenance: From theory to practice*. In Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 3–10).
5. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). *Can AI help reduce disparities in general medical and mental health care?*. AMA Journal of Ethics, 21(2), E167-179.
6. Microsoft. (2021). *Responsible AI Principles*. Retrieved from https://www.microsoft.com/en-us/ai/responsible-ai
7. Google AI. (2022). *AI Principles*. Retrieved from https://ai.google/responsibilities/responsible-ai-practices/
8. Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., & Keim, D. A. (2017). *Visual interaction with dimensionality reduction: A structured literature analysis*. IEEE Transactions on Visualization and Computer Graphics, 23(1), 241-250.
9. Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. In Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (pp. 149–159).
10. NIST. (2023). *AI Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology.
11. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets*. arXiv preprint arXiv:1803.09010.
12. Holland, S., Hosny, A., Newman, S., Joseph, J., & Chouldechova, A. (2018). *The dataset nutrition label: A framework for data transparency*. arXiv preprint arXiv:1805.03677.
13. Koshy, R., Venkatraman, S., & Sundararajan, A. (2022). *Data governance and lineage in regulated AI systems*. Journal of Data and Information Quality, 14(3), 1-24.
14. OpenLineage. (2024). *Open standard for metadata and lineage collection*. Retrieved from https://openlineage.io/
15. DataHub Project. (2024). *Metadata platform for the modern data stack*. Retrieved from https://datahubproject.io/
16. Apache Atlas. (2024). *Governance and metadata framework for Hadoop*. Retrieved from https://atlas.apache.org/
17. IBM. (2022). *AI Factsheets 360: Factsheet for datasets and models*. IBM Research. Retrieved from https://www.research.ibm.com/artificial-intelligence/trusted-ai/
18. Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). *ModelTracker: Redesigning performance analysis tools for machine learning*. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 337–346).
19. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate*. Big Data & Society, 3(2), 2053951716679679.
20. Kagal, L., Finin, T., & Joshi, A. (2003). *A policy language for a pervasive computing environment*. In Proceedings of the 4th IEEE International Workshop on Policies for Distributed Systems and Networks .