Training Course on Data Lineage and Provenance for Trustworthy AI

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

In the rapidly evolving landscape of artificial intelligence **(AI)** and machine learning **(ML)**, the demand for **trustworthy AI systems** has become paramount. Organisations leveraging AI face increasing scrutiny regarding model fairness, accountability, and regulatory compliance. The training course on data lineage & provenance for trustworthy AI delves into the critical disciplines of **data lineage** and **data provenance**, equipping professionals with the essential knowledge and practical skills to meticulously track data origins, transformations, and usage within complex AI pipelines. By mastering these concepts, participants will be empowered to build and manage AI systems that are not only powerful but also transparent, explainable, and ethically sound.

The ability to demonstrate the **auditability** and **reproducibility** of AI model outcomes is no longer optional but a fundamental requirement for **responsible AI development**. This course provides a deep dive into the technical and governance frameworks necessary to establish robust data trails, ensuring data quality, mitigating bias, and adhering to evolving **data privacy regulations** and **ethical AI guidelines**. Participants will gain hands-on experience with tools and methodologies to implement end-to-end data traceability, fostering greater confidence and trust in AI-driven decisions across diverse industries.

Programme Curriculum

Training Course on Data Lineage & Provenance for Trustworthy AI

Introduction

In the rapidly evolving landscape of **Artificial Intelligence (AI)** and **Machine Learning (ML)**, the demand for **trustworthy AI systems** has become paramount. Organizations leveraging AI face increasing scrutiny regarding model fairness, accountability, and regulatory compliance. Training Course on Data Lineage & Provenance for Trustworthy AI delves into the critical disciplines of **Data Lineage** and **Data Provenance**,

equipping professionals with the essential knowledge and practical skills to meticulously track data origins, transformations, and usage within complex AI pipelines. By mastering these concepts, participants will be empowered to build and manage AI systems that are not only powerful but also transparent, explainable, and ethically sound.

The ability to demonstrate the **auditability** and **reproducibility** of AI model outcomes is no longer optional but a fundamental requirement for **responsible AI development**. This course provides a deep dive into the technical and governance frameworks necessary to establish robust data trails, ensuring data quality, mitigating bias, and adhering to evolving **data privacy regulations** and **ethical AI guidelines**. Participants will gain hands-on experience with tools and methodologies to implement end-to-end data traceability, fostering greater confidence and trust in AI-driven decisions across diverse industries.

Course Duration

10 days

Course Objectives

1. Comprehend the fundamental differences and critical interdependencies between data lineage and data provenance in the context of **AI lifecycles**.
2. Implement principles of **Responsible AI** and **AI Governance** through robust data tracking mechanisms.
3. Identify and mitigate data quality issues at their source by tracing data transformations and ensuring **data accuracy**.
4. Navigate complex data regulations (e.g., **GDPR, CCPA, EU AI Act**) by building auditable data trails for **AI accountability**.
5. Analyze data origins and transformations to identify and address **algorithmic bias** and promote **AI fairness**.
6. Develop strategies to provide clear and transparent explanations of AI model decisions by understanding underlying data flows.
7. Design and manage efficient **AI data pipelines** with integrated lineage and provenance capabilities.
8. Integrate data lineage and provenance into broader **data governance strategies** for holistic data management.
9. Utilize **metadata management** and **data cataloging** tools to automate and streamline lineage and provenance tracking.
10. Implement practices for data versioning and ensure the **reproducibility of AI experiments** and results.
11. Conduct effective **AI data audits** to verify data integrity, compliance, and ethical usage.
12. Understand and apply best practices for **ethical data sourcing**, usage, and destruction within AI systems.
13. Empower organizations to make informed and reliable decisions based on traceable and trustworthy AI outputs.

Organizational Benefits

- Build greater confidence in AI systems among stakeholders, customers, and regulators.
- Ensure compliance with evolving data privacy and AI regulations, minimizing legal and financial penalties.
- Identify and resolve data issues proactively, leading to more accurate and dependable AI models.
- Streamline data pipeline development and debugging through clear data traceability.
- Prevent incidents of unfair or biased AI decisions by addressing data-related issues.
- Gain insights into data usage patterns to improve data storage, processing, and governance efficiency.

- Establish a robust framework for managing data throughout its lifecycle, supporting broader organizational goals.
- Provide clear, demonstrable evidence of data handling practices for internal and external audits.

Target Audience

1. Data Scientists & Machine Learning Engineers.
2. AI/ML Architects.
3. Data Governance & Compliance Officers
4. Data Engineers & ETL Developers
5. Risk Management Professionals
6. Chief Data Officers (CDOs) & Chief AI Officers (CAIOs).
7. Auditors & Legal Professionals
8. Business Analysts & Stakeholders.

Course Outline

Module 1: Introduction to Trustworthy AI & Its Data Imperatives

- Defining Trustworthy AI: Fairness, Accountability, Transparency, Explainability.
- The Critical Role of Data in AI Trustworthiness.
- Understanding the AI Lifecycle and Data Touchpoints.
- Emerging Regulations: EU AI Act, Data Governance Act, and their impact.
- Challenges in achieving AI transparency and auditability.
- **Case Study:** Analysis of a high-profile AI bias incident (e.g., facial recognition misidentification) and the role of untraceable data.

Module 2: Fundamentals of Data Lineage

- What is Data Lineage? Definition, Scope, and Importance.
- Types of Data Lineage: Business, Technical, Operational.
- Key Components of Data Lineage: Sources, Transformations, Destinations.
- Data Flow Mapping and Visualization Techniques.
- Benefits of comprehensive data lineage for AI development.
- **Case Study:** Tracing data from raw customer input to a credit scoring AI model's output in a financial institution.

Module 3: Deep Dive into Data Provenance

- What is Data Provenance? Understanding Data Origin and History.
- Distinction between Data Lineage and Data Provenance.
- Capturing Provenance Metadata: Who, What, When, Where, Why.
- Importance of Immutable Data Records for AI.
- Tools and Technologies for Provenance Tracking.
- **Case Study:** Investigating the provenance of a medical dataset used to train a diagnostic AI, including collection methods and any pre-processing steps.

Module 4: Data Quality and Integrity for AI Models

- Impact of Data Quality on AI Performance and Bias.
- Common Data Quality Issues: Inaccuracies, Inconsistencies, Missing Data.
- Strategies for Data Profiling and Validation.

- Implementing Data Cleansing and Transformation Rules.
- Leveraging Lineage to Pinpoint Data Quality Problems.
- **Case Study:** Anomaly detection in a manufacturing process AI where data quality issues in sensor readings led to false positives.

Module 5: Regulatory Compliance & AI Accountability

- Overview of Global Data Protection and AI Regulations.
- Data Lineage as a Compliance Enabler (GDPR, HIPAA, CCPA, etc.).
- Establishing AI Accountability Frameworks.
- Demonstrating Due Diligence in AI Development.
- Audit Trails and Reporting for Regulatory Scrutiny.
- **Case Study:** A company demonstrating GDPR compliance for its AI-driven customer service chatbot by proving data handling and deletion processes.

Module 6: Mitigating Bias and Ensuring Fairness in AI

- Understanding Sources of Bias in AI Data (Selection, Collection, Annotation).
- The Role of Lineage and Provenance in Identifying Bias.
- Techniques for Bias Detection and Mitigation.
- Fairness Metrics and Their Application.
- Ethical Considerations in AI Data Collection.
- **Case Study:** Analyzing a recruiting AI that showed gender bias, tracing the bias back to historical training data and its collection methods.

Module 7: Explainable AI (XAI) and Data Transparency

- The Need for Explainability in Trustworthy AI.
- Connecting Data Lineage to AI Model Interpretability.
- Techniques for Explaining AI Decisions through Data Context.
- Visualizing Data Flows for Enhanced Understanding.
- Communicating AI Transparency to Non-Technical Stakeholders.
- **Case Study:** How a financial institution uses data lineage to explain a loan approval AI's decision to a customer.

Module 8: Designing and Managing AI Data Pipelines

- Architecting Robust Data Pipelines for AI.
- Integrating Lineage and Provenance into ETL/ELT Processes.
- Automating Data Flow Capture and Documentation.
- Version Control for Data and AI Models.
- Best Practices for Scalable Data Pipeline Management.
- **Case Study:** A tech company implementing an MLOps pipeline with automated data lineage tracking for continuous model retraining.

Module 9: Metadata Management and Data Cataloging

- The Power of Metadata in Data Lineage & Provenance.
- Types of Metadata: Technical, Business, Operational.
- Implementing Data Catalogs for Discoverability and Governance.
- Automated Metadata Capture and Enrichment.

- Leveraging Data Catalogs for AI Data Discovery and Trust.
- **Case Study:** A large enterprise using a data catalog to manage thousands of datasets and enable self-service data discovery for AI teams.

Module 10: Data Versioning and Reproducibility in AI

- The Importance of Data Versioning for AI Experimentation.
- Strategies for Managing Data Changes Over Time.
- Ensuring Reproducibility of AI Model Training and Results.
- Tools for Data Version Control (e.g., DVC).
- Impact of Reproducibility on AI Reliability and Auditability.
- **Case Study:** A research lab using data versioning to reproduce scientific findings from an AI model and validate its results.

Module 11: Data Auditability for AI Systems

- Principles of Auditing AI Systems.
- Designing an AI Data Audit Program.
- Tools and Techniques for Data Audit Trails.
- Responding to Audit Inquiries and Compliance Checks.
- Continuous Monitoring for AI Data Integrity.
- **Case Study:** An external auditor assessing an insurance company's AI-powered claims processing system for fairness and data compliance.

Module 12: Ethical AI Data Practices

- Ethical Principles for AI Data Collection and Usage.
- Data Minimization and Purpose Limitation.
- Consent Management and Data Subject Rights.
- Addressing Data Security and Privacy Concerns.
- Developing and Implementing Ethical AI Guidelines.
- **Case Study:** A social media company redesigning its data collection practices for AI to align with ethical user data principles and privacy.

Module 13: Advanced Topics in Data Lineage & Provenance for AI

- Graph Databases for Complex Data Lineage.
- Blockchain for Immutable Data Provenance.
- AI-Powered Data Lineage Automation.
- Federated Learning and Data Lineage Challenges.
- Future Trends in Trustworthy AI and Data Governance.
- **Case Study:** Exploring how blockchain could be used to create an indisputable record of data transformations in a supply chain AI.

Module 14: Practical Implementation & Tooling

- Overview of Commercial and Open-Source Lineage Tools.
- Hands-on Exercises with Selected Data Lineage & Provenance Tools.
- Integrating Lineage Tools with Existing Data Ecosystems.
- Building a Business Case for Data Lineage & Provenance Investment.
- Developing an Implementation Roadmap.

- **Case Study:** A startup integrating a new data lineage tool into their existing cloud data warehouse and MLOps platform.

Module 15: Capstone Project & Future Outlook

- Applying Learned Concepts to a Real-World AI Scenario.
- Designing a Data Lineage & Provenance Strategy for a Fictional AI Product.
- Presenting Solutions and Justifications.
- The Evolving Landscape of AI Trustworthiness.
- Continuous Learning and Professional Development.
- **Case Study:** Participants work in groups to develop a data governance plan for a new generative AI application, focusing on data lineage, provenance, and ethical use.

Training Methodology

This course employs a **blended learning approach** designed for maximum engagement and practical application.

- **Interactive Lectures:** Core concepts will be introduced through dynamic presentations with real-world examples.
- **Hands-on Labs & Workshops:** Participants will gain practical experience using industry-relevant tools for data lineage, provenance, and metadata management.
- **Case Study Analysis:** In-depth examination of real-world scenarios to understand challenges and best practices.
- **Group Discussions & Collaborative Exercises:** Fostering peer learning and diverse perspectives on complex topics.
- **Practical Demonstrations:** Live walkthroughs of data pipeline and governance tool functionalities.
- **Q&A Sessions:** Dedicated time for addressing participant questions and clarifying concepts.
- **Capstone Project:** A culminating project to apply all learned skills in a practical, integrated scenario.
- **Expert Guest Speakers:** Industry leaders and practitioners sharing insights and experiences.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

-

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com