Data Lakehouse Architecture for Research Data Training Course

Professional Development Training Course Outline

Category	Research and Data Analysis	**Duration**	5 Days
Base Fee	$1,500 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

In today’s rapidly evolving data-driven research landscape, traditional data management systems are no longer sufficient to meet the needs of modern research workflows. The Data Lakehouse Architecture is revolutionizing how research institutions, universities, and organizations handle vast and diverse datasets by combining the best features of data lakes and data warehouses. Data Lakehouse Architecture for Research Data Training Course is designed to equip researchers, data architects, and data analysts with the critical knowledge and technical expertise needed to implement and manage a high-performance data lakehouse system that supports real-time analytics, scalable storage, and structured/unstructured data processing.

Through hands-on labs, real-world case studies, and expert-led modules, participants will explore the integration of big data frameworks (like Apache Spark and Delta Lake) with cloud platforms, ensuring governance, data quality, and enhanced collaboration in research environments. Whether you are dealing with genomic data, climate simulations, or behavioral research, this course offers a comprehensive guide to unlocking the full potential of data lakehouse architecture for scientific research.

Programme Curriculum

Data Lakehouse Architecture for Research Data Training Course

Introduction

In today’s rapidly evolving data-driven research landscape, traditional data management systems are no longer sufficient to meet the needs of modern research workflows. The Data Lakehouse Architecture is revolutionizing how research institutions, universities, and organizations handle vast and diverse datasets by combining the best features of data lakes and data warehouses. Data Lakehouse Architecture for Research Data Training Course is designed to equip researchers, data architects, and data analysts with the critical knowledge and technical expertise needed to implement and manage a

high-performance data lakehouse system that supports real-time analytics, scalable storage, and structured/unstructured data processing.

Through hands-on labs, real-world case studies, and expert-led modules, participants will explore the integration of big data frameworks (like Apache Spark and Delta Lake) with cloud platforms, ensuring governance, data quality, and enhanced collaboration in research environments. Whether you are dealing with genomic data, climate simulations, or behavioral research, this course offers a comprehensive guide to unlocking the full potential of data lakehouse architecture for scientific research.

Course Objectives

1. Understand the core concepts of data lakehouse architecture and its relevance in research.
2. Compare and contrast data lakes, warehouses, and lakehouses.
3. Learn to design scalable data lakehouse infrastructures using cloud-native tools.
4. Implement Delta Lake and Apache Iceberg for optimized research data storage.
5. Enable real-time research analytics using Apache Spark in a lakehouse.
6. Apply data governance and compliance best practices in research data management.
7. Leverage machine learning workflows within the lakehouse framework.
8. Integrate structured and unstructured research data seamlessly.
9. Perform ETL/ELT operations and automation in a lakehouse environment.
10. Secure multi-tenant access and ensure role-based data security in academic institutions.
11. Manage metadata and schema evolution effectively for evolving research needs.
12. Explore interoperability with scientific tools and languages (Python, R, SQL).
13. Develop real-world research solutions through guided case studies and team projects.

Target Audiences

1. Data Scientists involved in scientific or academic research.
2. Academic Researchers working with large-scale datasets.
3. University IT Administrators handling research infrastructure.
4. Data Engineers building big data solutions in research.
5. Bioinformatics Analysts managing genomic data pipelines.
6. Climate and Environmental Scientists needing scalable storage systems.
7. Government Research Institutions adopting cloud-based analytics.
8. PhD Students and Postdoctoral Researchers exploring advanced data architectures.

Course Duration: 5 days

Course Modules

Module 1: Introduction to Data Lakehouse Architecture

- Define data lakehouse and its evolution
- Advantages over traditional data lakes/warehouses
- Key components: storage, processing, governance
- Technologies powering lakehouses (Delta Lake, Apache Spark)
- Challenges in research data storage and how lakehouses help
- **Case Study**: Transitioning from a university data warehouse to lakehouse

Module 2: Building a Scalable Lakehouse Infrastructure

- Cloud vs on-premise lakehouse platforms
- Storage formats and optimization (Parquet, ORC)
- Lakehouse with AWS, Azure, Google Cloud
- Data ingestion and streaming techniques

- Cost management strategies
- **Case Study**: Lakehouse setup for a national weather research center

Module 3: Research Data Processing with Apache Spark

- Introduction to Apache Spark for research
- Batch vs. stream processing
- Research query optimization
- Spark SQL for scientific data analytics
- Integrating Spark MLlib for predictive modeling
- **Case Study**: Real-time analysis of COVID-19 datasets

Module 4: Delta Lake and Apache Iceberg in Research

- ACID transactions for research data
- Time travel and version control
- Schema evolution in scientific datasets
- Partitioning and performance tuning
- Iceberg vs. Delta Lake for reproducibility
- **Case Study**: Longitudinal health data storage using Delta Lake

Module 5: Data Governance and Security

- Data privacy regulations (HIPAA, GDPR)
- Encryption and secure data access
- Audit trails and role-based permissions
- Metadata cataloging and lineage tracking
- Governance tools: Unity Catalog, Apache Ranger
- **Case Study**: Secure access in clinical trial data lakehouse

Module 6: ETL, Automation & Workflow Orchestration

- Data wrangling and ETL pipelines with Databricks
- Airflow and dbt for workflow orchestration
- Automating data quality checks
- Data validation and testing frameworks
- Continuous integration/continuous deployment (CI/CD)
- **Case Study**: Automating lab sensor data ETL in academic labs

Module 7: Machine Learning and Advanced Analytics

- ML lifecycle in a lakehouse environment
- Feature store for reproducible research
- Model versioning and deployment
- Integration with Jupyter, MLflow, TensorFlow
- Use of AI for scientific discovery
- **Case Study**: Predicting disease spread using lakehouse ML workflows

Module 8: Future Trends & Real-World Applications

- Data mesh and decentralized research data management
- Lakehouse for open science and FAIR principles
- Multi-language support (SQL, Python, R)
- Collaboration tools in research ecosystems
- Evaluating ROI and sustainability of lakehouses
- **Case Study**: Implementing a cross-university data mesh lakehouse

Training Methodology

- Hands-on Lab Sessions using cloud-based environments
- Live Expert Lectures with industry case examples
- Group Activities and Peer Learning for collaboration
- Downloadable Course Materials & Resources
- Final Capstone Project to evaluate application of knowledge
- Q&A and Feedback Sessions for clarity and improvement

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	27 Nov 2026	Online	$1,000
23 Nov 2026	27 Nov 2026	Physical	$1,500
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

Start Date	End Date	Location	Price
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

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

Register Online Now

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