Data Engineering for Research Data Pipelines 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 the digital age, the effective collection, transformation, and management of research data is essential for meaningful scientific discovery and innovation. Data Engineering for Research Data Pipelines Training Course equips learners with hands-on, cutting-edge skills in designing scalable, reproducible, and optimized data pipelines using modern data engineering tools and practices. This course is tailored for professionals who need to manage structured and unstructured research data efficiently—from raw ingestion to usable datasets ready for analysis and machine learning.

With data volumes growing exponentially, researchers and engineers need robust knowledge of technologies like Apache Airflow, Apache Spark, SQL, ETL/ELT frameworks, and cloud-based data solutions. This course bridges the gap between research data demands and practical data engineering skills by providing real-world scenarios and case studies in scientific research, healthcare analytics, social science, and environmental data.

Programme Curriculum

Data Engineering for Research Data Pipelines Training Course

Introduction

In the digital age, the effective collection, transformation, and management of research data is essential for meaningful scientific discovery and innovation. Data Engineering for Research Data Pipelines Training Course equips learners with hands-on, cutting-edge skills in designing scalable, reproducible, and optimized data pipelines using modern data engineering tools and practices. This course is tailored for professionals who need to manage structured and unstructured research data efficiently—from raw ingestion to usable datasets ready for analysis and machine learning.

With data volumes growing exponentially, researchers and engineers need robust knowledge of technologies like Apache Airflow, Apache Spark, SQL, ETL/ELT frameworks, and cloud-based data solutions. This course bridges the gap between

research data demands and practical data engineering skills by providing real-world scenarios and case studies in scientific research, healthcare analytics, social science, and environmental data.

Course Objectives

1. Understand the fundamentals of data engineering in research environments.
2. Develop automated data pipelines for various research datasets.
3. Implement ETL and ELT strategies for structured and unstructured data.
4. Use Apache Airflow to schedule and monitor research data workflows.
5. Leverage Apache Spark for big data processing in research contexts.
6. Apply SQL and NoSQL for effective data storage and retrieval.
7. Design data lakes and warehouses tailored for academic and institutional research.
8. Ensure data quality, validation, and lineage in research projects.
9. Integrate cloud platforms like AWS, Azure, or GCP for research data infrastructure.
10. Analyze real-time and batch data for scientific studies.
11. Explore metadata management and documentation best practices.
12. Secure data governance and compliance in research pipelines.
13. Build scalable and reproducible data workflows for collaborative research.

Target Audiences

1. Data engineers in research institutions
2. University researchers and graduate students
3. Health informatics professionals
4. Government and NGO data analysts
5. Data scientists managing research data
6. IT professionals supporting research projects
7. Environmental and social science researchers
8. Research software engineers

Course Duration: 5 days

Course Modules

Module 1: Introduction to Research Data Engineering

- Overview of data engineering in research
- Role of data pipelines in scientific workflows
- Research data lifecycle
- Tools and technologies overview
- Data formats in academia (CSV, JSON, XML, etc.)
- **Case Study: Research pipeline in clinical trial data management**

Module 2: Data Ingestion and Collection

- Batch vs. real-time ingestion
- APIs and web scraping for research data
- IoT and sensor data collection
- Using Python and R for data extraction
- Connecting to research databases
- **Case Study: Environmental data ingestion using IoT sensors**

Module 3: ETL and ELT Design for Research

- ETL vs. ELT in research environments

- Data transformation techniques
- Data validation and cleaning
- Workflow optimization strategies
- Open-source tools: Apache NiFi, Talend
- **Case Study: ETL process for national census datasets**

Module 4: Workflow Orchestration with Apache Airflow

- DAGs and task dependencies
- Scheduling and triggering workflows
- Monitoring and logging
- Versioning research workflows
- Integrating Airflow with cloud platforms
- **Case Study: Airflow-based pipeline for genomics research**

Module 5: Big Data Processing with Apache Spark

- Spark architecture and RDDs
- Spark SQL for research queries
- DataFrames and MLlib
- Spark on Databricks or EMR
- Optimizing research workloads
- **Case Study: Large-scale social media sentiment analysis**

Module 6: Storage Solutions and Data Modeling

- Relational vs. non-relational models
- Schema design for research data
- Cloud storage: S3, Azure Blob
- Data lakes vs. warehouses
- File formats: Parquet, Avro
- **Case Study: Storage design for medical imaging data**

Module 7: Data Governance and Compliance

- Data privacy in academic research
- HIPAA, GDPR, and institutional compliance
- Data lineage and auditing
- Role-based access and IAM
- Metadata and data dictionaries
- **Case Study: Compliance in university-hosted research projects**

Module 8: Final Project and Capstone

- End-to-end pipeline design
- Selecting research data use case
- Implementing ingestion, transformation, storage
- Workflow orchestration and optimization
- Presentation and peer review
- **Case Study: Capstone on public health research pipeline**

Training Methodology

- Instructor-led virtual or onsite sessions
- Hands-on labs with real datasets
- Guided projects with mentor support

- Group discussions and peer review
- Quizzes and knowledge checks
- Final capstone project with feedback

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
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000

Start Date	End Date	Location	Price
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