Training Course on Data Orchestration with Airflow/Dagster

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 today's data-driven landscape, efficient and reliable **data workflow automation** is paramount for organizations striving for **data-driven decision-making** and **operational efficiency**. Complex data pipelines, ranging from **ETL/ELT processes** to **machine learning workflows**, demand robust orchestration tools to manage dependencies, ensure data quality, and enable **real-time data processing**. This comprehensive training course delves into the intricacies of **data orchestration**, focusing on two industry-leading platforms: **Apache Airflow** and **Dagster**. Participants will gain hands-on expertise in designing, deploying, monitoring, and scaling **modern data platforms**, mastering the art of **data pipeline management** for enhanced productivity and reduced data bottlenecks.

Training Course on Data Orchestration with Airflow/Dagster provides a practical, project-based learning experience, equipping data professionals with the essential skills to build **resilient data architectures** and drive significant **business value** through automated data initiatives. By exploring advanced features, best practices, and real-world case studies, attendees will be empowered to optimize their organization's data infrastructure, fostering greater **data governance**, **data democratization**, and ultimately, a more agile and responsive data ecosystem.

Programme Curriculum

Training Course on Data Orchestration with Airflow/Dagster

Introduction

In today's data-driven landscape, efficient and reliable **data workflow automation** is paramount for organizations striving for **data-driven decision-making** and **operational efficiency**. Complex data pipelines, ranging from **ETL/ELT processes** to **machine learning workflows**, demand robust orchestration tools to manage dependencies, ensure data quality, and enable **real-time data processing**. This comprehensive

training course delves into the intricacies of **data orchestration**, focusing on two industry-leading platforms: **Apache Airflow** and **Dagster**. Participants will gain hands-on expertise in designing, deploying, monitoring, and scaling **modern data platforms**, mastering the art of **data pipeline management** for enhanced productivity and reduced data bottlenecks.

Training Course on Data Orchestration with Airflow/Dagster provides a practical, project-based learning experience, equipping data professionals with the essential skills to build **resilient data architectures** and drive significant **business value** through automated data initiatives. By exploring advanced features, best practices, and real-world case studies, attendees will be empowered to optimize their organization's data infrastructure, fostering greater **data governance**, **data democratization**, and ultimately, a more agile and responsive data ecosystem.

Course Duration

10 days

Cours Objectives

1. Develop proficiency in designing, building, and managing complex data pipelines using Airflow and Dagster.
2. Create robust and fault-tolerant Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) processes.
3. Orchestrate end-to-end machine learning workflows for seamless model training, deployment, and monitoring.
4. Understand and apply Airflow (e.g., Google Cloud Composer, Amazon MWAA) and Dagster within cloud environments.
5. Implement strategies for data validation, error handling, monitoring, and gaining deep insights into data flow health.
6. Establish clear data governance policies and track data lineage for compliance and auditing.
7. Design scalable data architectures capable of handling increasing data volumes and complexity.
8. Connect and orchestrate data flows from various databases, APIs, data lakes, and streaming platforms.
9. Utilize container technologies for deploying and managing Airflow and Dagster environments.
10. Explore techniques for near real-time data ingestion and processing with orchestration tools.
11. Apply Continuous Integration/Continuous Delivery principles to data workflow development and deployment.
12. Extend Airflow and Dagster functionalities with custom components for specific use cases.
13. Enable broader access and usability of organizational data through well-orchestrated pipelines.

Organizational Benefits

- Automate repetitive data tasks, reducing manual effort and human error.
- Ensure consistent, accurate, and validated data across all systems.
- Accelerate data processing and delivery, enabling quicker business intelligence and decision-making.
- Build data infrastructure that can easily adapt to growing data volumes and evolving business needs.
- Optimize resource utilization and minimize the need for extensive manual data management.
- Centralize data management, simplifying policy enforcement and regulatory adherence.
- Efficiently allocate computing power, storage, and network bandwidth for data workflows.
- Provide a unified framework for data engineers, data scientists, and analysts to work collaboratively.
- Leverage real-time data capabilities for proactive market response and informed strategic planning.

Target Audience

1. Data Engineers.
2. ETL Developers.
3. Data Scientists.
4. Cloud Architects.
5. DevOps Engineers.
6. Analytics Engineers.
7. Data Architects.
8. Software Engineers with Data Focus

Course Outline

Module 1: Introduction to Data Orchestration and Modern Data Stack

- Define data orchestration and its significance in modern data platforms.
- Explore the evolution of data architectures: from data warehouses to data lakes and lakehouses.
- Understand the key components of a modern data stack and where orchestration fits.
- Discuss the challenges of manual data workflow management and the need for automation.
- **Case Study:** A global e-commerce company struggling with fragmented data sources and manual ETL processes, leading to delayed reporting and inconsistent analytics.

Module 2: Apache Airflow Fundamentals

- Introduction to Airflow architecture: scheduler, web server, worker, and database.
- Understanding Directed Acyclic Graphs (DAGs), tasks, and operators.
- Setting up an Airflow development environment (local, Docker Compose).
- Writing your first DAG: basic task dependencies and execution.
- **Case Study:** A marketing analytics firm uses Airflow to automate daily report generation from various ad platforms, ensuring timely insights.

Module 3: Advanced Airflow Concepts

- Scheduling and triggering DAGs: cron expressions, dataset-driven scheduling.
- Managing connections, variables, and XComs for data sharing.
- Implementing branching, sub-DAGs, and task groups for complex workflows.
- Error handling, retries, and SLAs in Airflow.
- **Case Study:** A financial institution uses Airflow's branching capabilities to handle different data validation paths based on transaction types.

Module 4: Deploying Airflow in Production

- Containerizing Airflow with Docker.
- Deploying Airflow on Kubernetes using Helm charts.
- Scaling Airflow components (scheduler, workers, webserver).
- Monitoring Airflow deployments with Prometheus and Grafana.
- **Case Study:** A SaaS company migrating its on-premise data pipelines to a cloud-based Airflow deployment on Kubernetes for scalability and reliability.

Module 5: Introduction to Dagster: Asset-Oriented Orchestration

- Understanding Dagster's asset-based approach vs. Airflow's task-based.
- Defining assets, ops, and jobs in Dagster.
- Setting up a Dagster development environment.

- Exploring the Dagit UI for observability and development.
- **Case Study:** A bioinformatics research lab adopts Dagster to manage complex genomic data processing workflows, focusing on data lineage and reproducibility.

Module 6: Building Data Assets with Dagster

- Designing and implementing data assets with dependencies.
- Working with I/O Managers for diverse data storage.
- Partitioning assets for efficient data processing.
- Materializing assets and understanding data versioning.
- **Case Study:** A media company uses Dagster to build a data catalog of its content assets, tracking transformations from raw video to optimized streaming formats.

Module 7: Advanced Dagster Features

- Dynamic partitioning and auto-materialize policies.
- Configuring sensors and schedules for event-driven orchestration.
- Testing Dagster assets and jobs for reliability.
- Using Dagster's software-defined assets for data quality checks.
- **Case Study:** A logistics company implements real-time shipment tracking using Dagster sensors triggered by IoT device data, ensuring immediate updates.

Module 8: Integrating Airflow and Dagster with Data Ecosystems

- Connecting with databases (PostgreSQL, MySQL, Snowflake).
- Integrating with cloud storage (S3, GCS, Azure Blob Storage).
- Working with data processing frameworks (Spark, dbt).
- Leveraging external APIs for data ingestion.
- **Case Study:** A retail chain integrates Airflow with its CRM and ERP systems, orchestrating data flows into a centralized data lake for customer analytics.

Module 9: Data Quality, Monitoring, and Alerting

- Implementing data validation checks within pipelines (e.g., Great Expectations).
- Setting up comprehensive monitoring dashboards for workflow health.
- Configuring alerts for failures, anomalies, and performance issues.
- Best practices for logging and debugging data pipelines.
- **Case Study:** A healthcare provider uses Airflow and Dagster with integrated data quality checks to ensure the accuracy and integrity of patient records for compliance.

Module 10: ETL/ELT Pipeline Automation

- Designing efficient data ingestion strategies.
- Implementing data transformations (SQL, Python, Spark).
- Loading data into data warehouses or data lakes.
- Building incremental load patterns for large datasets.
- **Case Study:** An online gaming company automates its daily ETL process using Airflow to consolidate player activity data for behavioral analysis.

Module 11: Machine Learning Workflow Orchestration

- Orchestrating data preparation for ML models.

- Automating model training and evaluation pipelines.
- Deployment strategies for machine learning models (e.g., MLOps).
- Monitoring model performance and retraining workflows.
- **Case Study:** A ride-sharing company uses Dagster to orchestrate its fraud detection ML pipeline, from data collection to model deployment and continuous monitoring.

Module 12: Data Governance and Security in Orchestration

- Implementing role-based access control (RBAC) in Airflow and Dagster.
- Managing sensitive data and credentials securely.
- Data lineage tracking for auditing and compliance.
- Best practices for data security in orchestrated environments.
- **Case Study:** A government agency utilizes Airflow with strict access controls and data lineage tracking to manage sensitive public sector data.

Module 13: Performance Optimization and Cost Management

- Optimizing DAG and job performance in Airflow and Dagster.
- Resource allocation and management for efficient execution.
- Identifying and resolving performance bottlenecks.
- Strategies for cost optimization in cloud-based orchestration.
- **Case Study:** A cloud-native startup reduced its data processing costs by 20% through optimizing Airflow DAGs and leveraging spot instances.

Module 14: Advanced Topics and Future Trends

- Exploring dynamic DAG generation in Airflow.
- Advanced Dagster concepts: graphs, subgraphs, and op factories.
- Integration with event-driven architectures (Kafka, RabbitMQ).
- Emerging trends in data orchestration: real-time data, AI-driven automation.
- **Case Study:** A telecommunications company experiments with real-time data ingestion and processing using Airflow for network anomaly detection.

Module 15: Capstone Project: End-to-End Data Pipeline

- Design and implement a complete data orchestration solution for a real-world scenario.
- Apply learned concepts to build a complex data pipeline using either Airflow or Dagster.
- Present and justify architectural decisions, demonstrating problem-solving skills.
- Troubleshoot and optimize the developed pipeline.
- **Case Study:** Participants work on a project involving ingesting data from a simulated IoT device, transforming it, and loading it into a data warehouse, with full orchestration.

Training Methodology

This training course employs a highly interactive and hands-on methodology to ensure deep understanding and practical skill development. The approach includes:

- **Interactive Lectures**: Clear explanations of concepts with practical examples.
- **Hands-on Labs**: Extensive coding exercises and guided projects using real-world datasets.
- **Live Coding Demonstrations**: Expert instructors showcase best practices and problem-solving techniques.
- **Case Studies**: In-depth analysis of industry examples to illustrate concepts and their application.

- **Group Discussions**: Collaborative learning and problem-solving sessions.
- **Q&A Sessions**: Opportunities for participants to clarify doubts and engage with instructors.
- **Project-Based Learning**: A capstone project to solidify learning and apply acquired skills.
- **Best Practices and Troubleshooting**: Focus on real-world scenarios, common pitfalls, and effective debugging strategies.

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

Start Date	End Date	Location	Price
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