Training Course on ETL/ELT Pipelines with Modern Data Stacks

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, organizations are flooded with vast amounts of information from diverse sources. The ability to effectively extract, transform, and load (ETL) or extract, load, and transform (ELT) this data into actionable insights is paramount for **business intelligence**, **data analytics**, and **machine learning**. Training Course on ETL/ELT Pipelines with Modern Data Stacks provides participants with the **hands-on skills** and **strategic understanding** necessary to design, build, and manage robust **data pipelines** using **modern data stacks**, ensuring data quality, scalability, and efficiency.

This course delves into the core concepts of **data ingestion**, various **transformation strategies**, and optimal **data loading techniques** for both **on-premise** and **cloud-native environments**. Participants will explore leading **data engineering tools** and frameworks, including **Apache Airflow**, **Snowflake**, **dbt**, and **Apache Kafka**, to orchestrate complex **data workflows** and establish a **single source of truth** for informed decision-making. Through practical **case studies** and **real-world projects**, attendees will gain the expertise to build resilient, high-performance data pipelines that power advanced analytics and drive significant **organizational value**.

Programme Curriculum

Training Course on ETL/ELT Pipelines with Modern Data Stacks

Introduction

In today's data-driven landscape, organizations are flooded with vast amounts of information from diverse sources. The ability to effectively extract, transform, and load (ETL) or extract, load, and transform (ELT) this data into actionable insights is paramount for **business intelligence**, **data analytics**, and **machine learning**. Training Course on ETL/ELT Pipelines with Modern Data Stacks provides participants with the **hands-on skills** and **strategic understanding** necessary to design, build, and manage robust **data pipelines** using **modern**

data stacks, ensuring data quality, scalability, and efficiency.

This course delves into the core concepts of **data ingestion**, various **transformation strategies**, and optimal **data loading techniques** for both **on-premise** and **cloud-native environments**. Participants will explore leading **data engineering tools** and frameworks, including **Apache Airflow**, **Snowflake**, **dbt**, and **Apache Kafka**, to orchestrate complex **data workflows** and establish a **single source of truth** for informed decision-making. Through practical **case studies** and **real-world projects**, attendees will gain the expertise to build resilient, high-performance data pipelines that power advanced analytics and drive significant **organizational value**.

Course Duration

10 days

Course Objectives

1. Master the fundamental concepts of ETL and ELT methodologies in the context of modern data warehousing and data lakes.
2. Design and implement scalable data ingestion strategies from diverse sources, including relational databases, APIs, streaming data (Kafka), and file systems.
3. Apply advanced data transformation techniques for data cleansing, data validation, data enrichment, and data normalization using Python and SQL.
4. Optimize data loading strategies for various target systems like Snowflake, Amazon Redshift, and Databricks, considering batch and incremental loading.
5. Utilize Apache Airflow for robust data pipeline orchestration and workflow automation, including DAG creation and monitoring.
6. Implement data quality checks and error handling mechanisms within ETL/ELT processes to ensure data integrity.
7. Explore the capabilities of cloud data platforms (AWS, Azure, GCP) for building and managing serverless data pipelines.
8. Leverage data transformation tools like dbt (data build tool) for efficient and version-controlled data modeling in the data warehouse.
9. Understand data governance and security best practices within ETL/ELT workflows to ensure compliance and data privacy.
10. Analyze and troubleshoot common performance bottlenecks in data pipelines and implement optimization techniques.
11. Evaluate and select appropriate ETL/ELT tools and frameworks based on specific business requirements and data volume.
12. Integrate ETL/ELT pipelines with BI and analytics tools (e.g., Tableau, Power BI) for effective data visualization and reporting.
13. Develop a comprehensive understanding of the modern data stack ecosystem and its evolving trends, including data mesh and data fabric concepts.

Organizational Benefits

- By implementing robust data cleansing and validation processes, organizations can ensure higher accuracy in their data assets, leading to more reliable business insights.
- Efficient ETL/ELT pipelines facilitate faster data integration and delivery, providing stakeholders with timely access to critical information for agile decision-making.

- Automation of data workflows reduces manual effort, minimizes errors, and frees up valuable resources, allowing teams to focus on strategic initiatives.
- Adopting modern data stack principles enables organizations to handle growing data volumes and evolving data types, ensuring their data infrastructure remains relevant and performant.
- Leveraging cloud-native solutions and optimized data processing techniques can lead to significant cost savings in infrastructure and resource allocation.
- Well-structured and clean data pipelines are the bedrock for successful machine learning models, predictive analytics, and other advanced data initiatives.
- Organizations with a strong data engineering foundation can react faster to market changes, identify new opportunities, and deliver data-driven products and services.

Target Audience

1. Data Engineers
2. Data Analysts.
3. Business Intelligence Developers
4. Software Developers.
5. Database Administrators.
6. Solution Architects.
7. IT Professionals.
8. Aspiring Data Professionals

Course Outline

Module 1: Introduction to Data Pipelines and the Modern Data Stack

- Understanding the evolution from traditional ETL to modern ELT.
- Key components of a modern data stack: data sources, ingestion, storage, transformation, and consumption.
- The role of data engineers in today's data-driven organizations.
- Differences between data warehouses, data lakes, and data lakehouses.
- Case Study: Analyzing a company's transition from an on-premise data warehouse to a cloud-based data lakehouse.

Module 2: Data Ingestion Fundamentals

- Overview of various data sources: relational databases, APIs, flat files (CSV, JSON, Parquet), streaming data.
- Batch vs. Real-time data ingestion techniques.
- Tools for data extraction: Fivetran, Stitch, custom scripts (Python).
- Data partitioning and indexing strategies for efficient ingestion.
- Case Study: Ingesting customer transaction data from an e-commerce platform's PostgreSQL database into a cloud storage bucket.

Module 3: Cloud Storage for Data Lakes

- Deep dive into cloud object storage (Amazon S3, Azure Data Lake Storage, Google Cloud Storage).
- Data lake architecture and best practices for organizing data.
- Data formats for efficiency: Parquet, Avro, ORC.
- Data lake vs. Data warehouse: choosing the right storage.

- Case Study: Designing a scalable data lake for a media company to store various media types (images, videos, text).

Module 4: SQL for Data Transformation

- Advanced SQL techniques for data manipulation (joins, window functions, CTEs).
- Data cleansing and validation using SQL.
- SQL for data aggregation and summarization.
- Best practices for writing performant SQL queries for large datasets.
- Case Study: Transforming raw sales data in a data lake into a structured format suitable for business intelligence reports using SQL.

Module 5: Python for Data Transformation (Pandas & PySpark)

- Introduction to Pandas for in-memory data manipulation.
- Data cleaning, transformation, and feature engineering with Pandas.
- Introduction to PySpark for distributed data processing.
- Scaling Python transformations with Spark (Databricks, EMR).
- Case Study: Developing a Python script using Pandas to clean and standardize product catalog data from multiple vendors.

Module 6: Data Orchestration with Apache Airflow

- Introduction to Directed Acyclic Graphs (DAGs) and Airflow concepts.
- Building and scheduling data pipelines with Airflow.
- Operators, sensors, and task dependencies in Airflow.
- Monitoring, logging, and error handling in Airflow.
- Case Study: Orchestrating an end-to-end data pipeline that extracts data from a CRM, transforms it, and loads it into a data warehouse using Airflow.

Module 7: Cloud Data Warehousing (Snowflake/Redshift/BigQuery)

- Deep dive into modern cloud data warehouses architectures.
- Loading data into Snowflake (or equivalent): Snowpipe, Copy Into.
- Optimizing query performance in cloud data warehouses.
- Data security and access control in the data warehouse.
- Case Study: Migrating an existing on-premise data warehouse to Snowflake and optimizing its data loading processes.

Module 8: Analytics Engineering with dbt (data build tool)

- Introduction to Analytics Engineering principles.
- Transforming data in the data warehouse using dbt.
- Version control, testing, and documentation with dbt.
- Building data models and views for analytics.
- Case Study: Using dbt to create robust and testable data models for financial reporting within a cloud data warehouse.

Module 9: Streaming Data Pipelines with Apache Kafka

- Introduction to real-time data processing and stream analytics.
- Apache Kafka fundamentals: topics, producers, consumers.

- Building basic streaming data pipelines with Kafka.
- Integrating Kafka with other data stack components.
- Case Study: Implementing a real-time analytics dashboard for website user activity by ingesting clickstream data via Kafka.

Module 10: Data Quality and Observability

- Strategies for ensuring data quality: data profiling, validation rules.
- Implementing data quality checks in ETL/ELT pipelines.
- Data observability concepts: monitoring, alerting, lineage.
- Tools for data quality and observability (e.g., Great Expectations, Datadog).
- Case Study: Setting up data quality checks for a critical marketing attribution pipeline to identify and resolve data discrepancies.

Module 11: Performance Optimization and Scalability

- Techniques for optimizing data extraction and loading performance.
- Strategies for scaling data transformations (horizontal vs. vertical scaling).
- Cost optimization in cloud data pipelines.
- Parallel processing and distributed computing concepts.
- Case Study: Identifying and resolving performance bottlenecks in a large-scale data ingestion pipeline that processes billions of records daily.

Module 12: Data Governance and Security in Pipelines

- Data governance frameworks and their importance in data pipelines.
- Implementing data masking, encryption, and access controls.
- Compliance with data privacy regulations (GDPR, CCPA).
- Auditing and logging data pipeline activities.
- Case Study: Designing a secure data pipeline for sensitive customer health records, ensuring compliance with HIPAA regulations.

Module 13: Advanced ETL/ELT Patterns and Best Practices

- Change Data Capture (CDC) and its applications.
- Reverse ETL: syncing data from data warehouses back to operational systems.
- Data Mesh and Data Fabric architectures.
- Infrastructure as Code (IaC) for data pipelines (e.g., Terraform).
- Case Study: Implementing a CDC mechanism to track changes in a master data management system and propagate them to downstream analytics systems.

Module 14: Project Work and Capstone

- Designing an end-to-end ETL/ELT pipeline for a hypothetical business problem.
- Implementing the pipeline using learned tools and techniques.
- Troubleshooting and refining the pipeline for production readiness.
- Presenting the solution and discussing challenges encountered.
- Case Study: Building a comprehensive data platform for a ride-sharing company, from raw driver/rider data to analytical dashboards.

Module 15: Emerging Trends in Data Engineering

- The rise of data contracts and data product thinking.
- Low-code/No-code ETL/ELT tools.
- Impact of AI/ML on data pipeline automation.
- Career paths and future outlook in data engineering.
- Case Study: Discussing the application of generative AI to automate schema detection and data mapping in complex integration scenarios.

Training Methodology

This training program employs a blended learning approach, combining theoretical concepts with extensive hands-on practice to ensure a comprehensive and practical learning experience.

- **Instructor-Led Sessions:** Interactive lectures, discussions, and Q&A sessions to cover core concepts and best practices.
- **Hands-on Labs:** Practical exercises and coding challenges using industry-standard tools and platforms (e.g., cloud environments, Python, SQL, Airflow).
- **Real-World Case Studies:** In-depth analysis and problem-solving based on practical scenarios faced by leading organizations.
- **Live Coding Demonstrations:** Instructors will demonstrate how to build and debug data pipelines step-by-step.
- **Group Projects:** Collaborative projects where participants apply their learned skills to design and implement end-to-end data solutions.
- **Mentorship and Support:** Access to instructors for guidance, feedback, and troubleshooting during and after the training.
- **Assessment:** Quizzes, assignments, and a final capstone project to evaluate understanding and practical application.

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