



DATA ENGINEERING TRAINING MODULE

Total Duration: 40 Hours

Mode: Instructor-led, Practical &
Project-based

Introduction to Data Engineering [3 Hours]

→ What is Data Engineering?

Understanding the role of data engineering in modern data-driven organizations. How it bridges data science, analytics, and software engineering.

→ Role and Responsibilities of a Data Engineer

Designing, building, and maintaining data infrastructure; working with large datasets; ensuring reliability, scalability, and security.

→ Data Lifecycle and Architecture Overview

From raw data ingestion to data warehousing and analytics.

→ ETL Concepts and Data Pipelines

Deep dive into Extract, Transform, Load [ETL] processes, ELT vs ETL, data ingestion methods [batch & streaming], and pipeline orchestration.

→ Key Tools Overview

Overview of tools used by data engineers - Airflow, Kafka, Spark, Hadoop, etc.

Databases & SQL [6 Hours]

→ Introduction to Databases

Understanding different types of databases - Relational [MySQL, PostgreSQL] and Non-relational [MongoDB, Cassandra].

→ SQL Fundamentals

Writing complex SQL queries using SELECT, WHERE, GROUP BY, ORDER BY, and aggregate functions.

→ Joins and Subqueries

Mastering different types of joins [INNER, LEFT, RIGHT, FULL], and using subqueries effectively.

→ Views, Indexing, and Stored Procedures

Creating reusable SQL components, optimizing queries for better performance, and automating operations.

→ **Data Modeling and Normalization**

Designing efficient database schemas, normalization [1NF to 3NF], denormalization for analytics, and understanding star & snowflake schemas.

→ **Hands-on Practice**

Building and querying a sample e-commerce or financial database.

Big Data Tools & Ecosystem [6 Hours]

→ **Understanding Big Data**

Characteristics of Big Data [Volume, Velocity, Variety], challenges, and solutions.

→ **Hadoop Ecosystem Overview**

Introduction to HDFS, MapReduce, Hive, and Pig. Understanding how Hadoop stores and processes large data sets.

→ **Apache Spark Fundamentals**

Spark architecture, RDDs, DataFrames, SparkSQL, and PySpark basics.

→ **Distributed File Systems**

Concepts of distributed storage and computation. Comparing HDFS, Amazon S3, and Azure Blob Storage.

→ **Real-world Use Cases**

How big companies use Hadoop and Spark for large-scale analytics and ETL.

Cloud Data Platforms [6 Hours]

→ **Introduction to Cloud in Data Engineering**

Benefits of cloud-based data pipelines: scalability, reliability, and cost-effectiveness.

→ **AWS for Data Engineers**

- ◆ S3: Data storage and versioning.
- ◆ Redshift: Building data warehouses and performing analytical queries.
- ◆ Glue: Managed ETL service for automating workflows.
- ◆ Athena: Understanding and usage of Query Engine

→ **Azure Data Services**

- ◆ Introduction to Azure Data Factory, Synapse Analytics, and Blob Storage.

→ **Hands-on Activities**

Building a simple ETL pipeline using AWS Glue and Redshift or Google BigQuery.

Python for Data Engineering [6 Hours]

→ **Python Essentials for Data Engineering**

Working with data structures, loops, functions, and libraries.

→ **Connecting Python with Databases**

Using libraries like psycopg2, SQLAlchemy, and pandas to interact with SQL and NoSQL databases.

→ **Working with APIs**

Fetching real-time data from REST APIs and integrating it into pipelines.

→ **Data Wrangling and Cleaning**

Using pandas, numpy, and regex to clean and transform raw data into usable formats.

→ **Automation and Scheduling**

Writing Python scripts for data automation, logging, and error handling.

→ **Hands-on Exercise**

Automating data extraction from a public API and loading it into a database.

Data Pipeline Projects [9 Hours]

- End-to-End Real-Time Data Pipeline
- Introduction to streaming concepts.
- Building a real-time data ingestion system using Kafka.
- Processing data with Apache Spark or Python scripts.
- Visualizing output through dashboards [Power BI / Tableau / Grafana].
- ETL Pipeline from Source to Data Warehouse
- Designing an ETL pipeline using Airflow for orchestration.
- Extracting data from APIs or databases.
- Transforming data using Python / Spark.
- Loading into AWS Redshift or BigQuery.

Mini Project [4 Hours]

- Students will build a complete data pipeline project from scratch - from ingestion to analysis.
- Capstone Project Discussion & Review
Presentation of final project work, discussion on real-world challenges, and industry best practices.

Key Takeaways

- Gain hands-on experience with modern data tools like Spark, Airflow, Kafka, AWS Glue, and Redshift.
- Learn to design and implement end-to-end data pipelines.
- Build strong SQL and Python foundations specifically for data workflows.
- Understand cloud data architecture and how to manage large-scale data efficiently.
- Become job-ready for roles such as Data Engineer, ETL Developer, or Cloud Data Specialist.