

# Data Engineering Course Syllabus

## Module 1: Introduction to Data Engineering

- Role of a Data Engineer
- Data Engineering vs. Data Science vs. Data Analytics
- Overview of Data Pipelines

## Module 2: Python for Data Engineering

- Python Basics & Advanced Concepts
- Data Manipulation with Pandas
- Working with APIs & JSON
- Handling Large Datasets in Python

## Module 3: SQL & Databases

- SQL Basics & Advanced Queries
- Normalization & Indexing
- Transactions & Stored Procedures
- NoSQL Databases (MongoDB, Cassandra)

## Module 4: Big Data Technologies (Hadoop & PySpark)

- Hadoop Ecosystem (HDFS, MapReduce, Hive)
- Introduction to Spark & PySpark
- RDDs, DataFrames & Datasets
- Optimizing Spark Performance

## Module 5: Cloud Data Engineering (AWS/GCP)

- AWS Services (S3, Redshift, Glue, Lambda)
- GCP Services (BigQuery, Dataflow, Pub/Sub)
- Serverless Data Processing

## Module 6: ETL & ELT Pipeline Design

- ETL vs. ELT Concepts
- Data Warehousing Principles
- Data Modeling & Schema Design
- Building Scalable Data Pipelines

## Module 7: Apache Airflow

- Introduction to Apache Airflow
- DAGs & Task Scheduling
- Workflow Automation & Monitoring

# Data Engineering Course Syllabus

- Best Practices for Airflow

## **Module 8: Real-World Projects & Case Studies**

- End-to-End Data Pipeline Implementation
- Data Lake vs. Data Warehouse Design
- Performance Optimization & Monitoring