Enterprise Data Engineering and Big Data Analytics

Course Description

This masterclass equips participants with robust technical expertise and hands-on experience to design, develop, and manage data pipelines for large-scale structured and unstructured data. Learners gain mastery over core programming, advanced SQL, Apache Spark, Hadoop, cloud platforms (including Azure), orchestration frameworks, and DevOps practices.

Real-world scenarios and industry best practices are blended with modern tooling to develop skills addressing enterprise-grade big data challenges.

Course Duration

- Total Duration: **88 Hours**

Learning Objectives

- Build scalable data engineering pipelines from scratch.
- Work efficiently with distributed storage and compute systems.
- Design and optimize both relational and NoSQL databases.
- Utilize cloud computing platforms and native services for modern data workflows.
- Orchestrate, automate, and ensure data quality in enterprise environments.
- Develop and deploy pipelines using CI/CD and DevOps methodologies.
- Implement effective collaboration and version control strategies.
- Integrate advanced data processing and modeling best practices.

Pre-requisites

- Basic understanding of programming concepts.
- Familiarity with databases and SQL queries.
- Exposure to data processing concepts (recommended).
- Comfortable working in a command-line environment.
- Prior big data experience is beneficial but not mandatory—skills are progressively built throughout the course.

Content Coverage :

Module 01: Programming Fundamentals

- Variables
- Data Types
- Operators
- Control Structures (if-else)
- Loops
- Functions
- Modular Programming
- Code Reusability

Module 02: Python for Data Engineering

- Python Syntax and Best Practices
- Data Structures (lists)
- Data Structures (dictionaries)
- Data Structures (sets)
- Data Structures (tuples)
- File Handling
- Exception Management
- Essential Libraries for Data Processing (pandas)
- Essential Libraries for Data Processing (numpy)
- Essential Libraries for Data Processing (os)
- Essential Libraries for Data Processing (sys)

Module 03: SQL for Data Management

- Relational Database Concepts
- Writing Optimized Queries (SELECT)
- Writing Optimized Queries (WHERE)

- Writing Optimized Queries (GROUP BY)
- JOINs
- Subqueries
- Set Operations
- Aggregations
- Window Functions
- Data Manipulation (INSERT)
- Data Manipulation (UPDATE)
- Data Manipulation (DELETE)
- Database Optimization
- Normalization
- Indexing
- Advanced SQL (Stored Procedures)
- Advanced SQL (Triggers)
- Advanced SQL (CTEs)
- Advanced SQL (Analytical Queries)

Module 04: Data Modeling & Cloud Platforms

- Introduction to Data Modeling Concepts and Techniques
- Entity-Relationship Diagrams (ERDs)
- Star Schema
- Snowflake Schema
- Data Vault Modeling Basics
- Use of Cloud Platforms in Data Engineering Workflows (Overview of Public Clouds)
- Cloud-native Data Architectures
- Cost Optimization Strategies
- AWS services for data engineering
- Azure services for data engineering
- GCP services for data engineering

Module 05: Azure Ecosystem Core Training

- Introduction to Azure Data Engineering
- Azure Storage (Blob)
- Azure Storage (Data Lake)
- Azure Synapse Analytics Basics
- Azure Data Factory (ETL Pipelines)
- Azure Databricks Integration
- Data Movement between Azure Services (Hands-on Labs)

Module 06: PySpark Application Development

- Introduction to Spark Ecosystem
- RDDs
- DataFrames
- Datasets
- Transformations
- Actions
- Performance Tuning
- Building PySpark Applications
- Deploying PySpark Applications
- Monitoring PySpark Applications

Module 07: Data Processing Methodologies

- Batch Processing Architectures
- Streaming Processing Architectures
- Design Patterns for Modern Data Workflows

Module 08: Big Data Storage Systems

- Distributed File Storage (HDFS)

- Distributed File Storage (S3)
- Distributed File Storage (Azure Data Lake)
- NoSQL Databases (Cassandra)
- NoSQL Databases (MongoDB)
- Data Lakes
- Data Warehouses
- Use Cases and Trade-offs

Module 09: CI/CD and DevOps for Data Pipelines

- Introduction to DevOps Principles in Data Engineering
- CI/CD Strategies for ETL/data pipelines
- Azure DevOps
- Jenkins
- GitHub Actions
- Automated Testing for Pipelines
- Automated Build and Deployment

Module 10: Branching Strategies & Version Control

- Fundamentals of Version Control (Git)
- Branching Strategies (Feature)
- Branching Strategies (Release)
- Branching Strategies (Hotfix)
- Merging Strategies
- Managing Data Engineering Source Code
- Code Review Best Practices
- Collaboration Best Practices

Module 11: Scala for Apache Spark

- Scala Language Essentials for Data Engineers

- Spark Transformations in Scala

Module 12: Apache Spark Advanced Concepts

- Spark SQL
- Real-Time Streaming
- MLlib Overview
- Application Performance Tuning

Module 13: Data Fabric Architecture

- Principles of Data Fabric
- Key Components of Data Fabric
- Design Patterns for Data Fabric
- Integrating Heterogeneous Data Sources

Module 14: Apache Airflow for Orchestration

- Airflow DAG Design
- Scheduling with Airflow
- Monitoring with Airflow

Module 15: Hadoop Ecosystem Overview

- Hadoop Core Architecture
- HDFS
- MapReduce Processing Model
- YARN Resource Management
- Hive
- Pig
- HBase

Module 16: Data Testing & Quality Assurance

- Data Quality Principles
- Data Quality Frameworks
- Unit Testing for Data Pipelines
- Integration Testing for Data Pipelines
- Regression Testing for Data Pipelines
- Data Testing Tools (pytest)
- Data Testing Tools (Great Expectations)
- Implementing Data Validations in ETL