data-virtuo

MASTERING DATA ENGINEERING

Learn to build scalable data pipelines and process big data with Python, PySpark, SQL, Hive, and ETL.

Curriculum

- Introduction to Data Engineering
- SQL for Data Analysis
- Python for Data Engineering
- PySpark for Big Data Processing
- Azure Cloud for Data Engineering
- ETL Concepts and Frameworks
- Capstone Project
- Bonus : Interview prep

AZURE DATA ENGINEERING

Detailed course topics

Module 1: Introduction to Data Engineering
- Overview of Data Engineering and its importance.
- Key tools and technologies in the data engineering ecosystem.
- Understanding the data lifecycle.

Module 2: SQL for Data Analysis
- Basics of SQL: SELECT, JOIN, GROUP BY, etc.
- Advanced SQL: CTEs, window functions, and optimization techniques.
- Writing SQL queries for ETL workflows.

Module 3: Python for Data Engineering
- Python fundamentals for data manipulation.
- Working with data structures.
- Writing reusable scripts for data processing.

Module 4: Azure cloud
- Introduction to Azure cloud and distributed computing.
- Setting up and using ADF.
- DataSets and Linked Services
- Activities within ADF.
- scenarios on ADF workflows

Module 5: PySpark for Big Data Processing
- Fundamentals of DataBricks
- Writing transformations and actions in PySpark.
- Integrating in ADF and scheduling pipeline
- Optimizing jobs.

Module 6: ETL Concepts and Frameworks in Databricks
- Understanding ETL (Extract, Transform, Load) processes.
- Complex scenarios on PySpark on Databricks
- Designing efficient ETL pipelines.
- Best practices in ETL development.

Module 7: Real-Time Project
- Real-world data engineering project.
- Building end-to-end pipelines.
- Data processing, storage, and reporting

Bonus: Interview Prep and Career Guidance.

This outline ensures participants gain theoretical knowledge and practical skills in data engineering.