

KOENIG SOLUTIONS • COURSE OUTLINE

Comprehensive Data Engineering with Python and Azure Databricks (Microsoft)

Learn to engineer data pipelines with Python, SQL, PySpark and Azure Databricks, from Delta Lake to interactive dashboards.

15 Days

DURATION

16

MODULES

126

TOPICS

12

HANDS-ON ACTIVITIES

COURSE OVERVIEW

This course begins with Python fundamentals, covering control flow, functions, collections, NumPy and Pandas, before introducing data visualisation with Matplotlib and Seaborn. It then moves into PySpark, SQL fundamentals, joins, aggregations and integrating SQL with Python. The course continues into Azure Databricks, covering the workspace, runtime architecture, jobs, workflows and scheduling, followed by landmark Azure architecture and Delta Lake concepts including ACID transactions, reliability features and the medallion architecture. It closes with performance tuning, monitoring and building interactive dashboards in Databricks SQL.

WHO THIS COURSE IS FOR

Data Engineers

Python Developers

Data Analysts

BI Developers

Database Administrators

LEARNING OBJECTIVES

- Apply Python control flow, functions and collections
- Analyse data using NumPy arrays and Pandas DataFrames
- Create visualisations with Matplotlib and Seaborn
- Compare RDDs and DataFrames when transitioning to PySpark
- Write SQL queries using filtering, sorting and joins
- Apply aggregations and best practices in SQL
- Configure Databricks clusters, jobs and scheduled workflows
- Implement Delta Lake features such as time travel and MERGE INTO

16 Modules, 126 Topics

01

Day 01: Python Basics & Control Flow

TOPICS

Python Basics · Variables and Data Types · Input/Output Operations · Basic Operators · Conditional Statements · Loops (for, while) · Control Flow Exercises

02

Day 02: Functions and Collections in Python

TOPICS

Defining and Calling Functions · Arguments and Return Values · Lists, Tuples, Sets, Dictionaries · Python Functions, Python Collection Objects · Iteration and Comprehension Exercises

03

Day 03: NumPy and Pandas Series

TOPICS

Introduction to NumPy Arrays · Mathematical Operations · Creating Pandas Series · Indexing and Filtering · Descriptive Statistics

04

Day 04: Pandas DataFrames and Visualization

TOPICS

Creating and Modifying DataFrames · Filtering, Grouping, Aggregating · Handling Missing Data · Plotting with Matplotlib · Customizing Charts · Statistical Plots with Seaborn

05

Day 05: Transition to PySpark

TOPICS

Python Code Evaluation · Writing Efficient Functions · Setting Up PySpark · RDDs vs DataFrames · Running Simple Transformations · Practical Comparison with Pandas

06

Day 06: SQL Fundamentals and Filtering

TOPICS

SQL Syntax and Datatypes · SELECT Statements · WHERE Clause for Filtering · LIKE, IN, BETWEEN · Practice with Sample Tables · Filtering and Pattern Matching

07

Day 07: SQL Sorting and Joins

TOPICS

ORDER BY Clause · Sorting Multiple Columns · INNER JOIN, LEFT JOIN, RIGHT JOIN · Joining Multiple Tables · Join Conditions and Filtering Joins

08

Day 08: Aggregations and SQL Best Practices

TOPICS

COUNT, SUM, AVG, MIN, MAX · GROUP BY with Aggregates · HAVING Clause · NULL Handling in Aggregations · Nested Aggregates · SQL Coding Best Practices

09

Day 09: SQL + Python Integration and Practice

TOPICS

Writing Advanced SQL Queries · Multi-table Joins + Aggregations · Query Optimization Hints · Reading SQL Data in Python · Executing SQL from Python

10

Day 10: Python Practice & Getting Started with Azure Databricks

TOPICS

Python Recap with Advanced Practice · Introduction to Azure Databricks · Databricks Workspace Overview · Cluster Creation and Management · Uploading and Exploring Datasets

11**Hands-on Lab**

Day 11: Understanding Databricks Runtime

OVERVIEW OF DATABRICKS RUNTIME ARCHITECTURE

- Runtime Versions and Compatibility
- Performance and Cost Considerations
- Selecting the Right Runtime (Standard, ML, Photon, etc.)
- Deep Dive into Databricks Components

INTRODUCTION TO JOBS AND TASKS

- Job Clusters vs Interactive Clusters
- Workflows and Pipelines in Databricks
- Job Monitoring and Logs
- Task Dependencies and Alerting
- Using Tags for Resource Management

SCHEDULING AND AUTOMATION

- Job Scheduler Basics
- Time-based Scheduling and Triggers
- Using Parameters and Task Values
- Notebook-based vs SQL-based Tasks
- Alerts and Notifications

12**2 Labs**

Day 12: Hands-on with Workflow Creation and Management

CREATING MULTI-TASK JOBS IN THE UI

Using Databricks Workflows with Notebooks · Debugging Failed Runs · Dependency Trees and Retry Logic

MONITORING AND TROUBLESHOOTING JOBS

Viewing Job Histories · Interpreting Logs and Metrics · Error Resolution Techniques

13

Landmark Architecture on Azure

OVERVIEW OF CURRENT REFERENCE ARCHITECTURES

- Role of Databricks in the Modern Azure Stack
- Integration with Azure Data Factory, ADLS, Synapse
- Data Lakehouse Concepts
- Governance, Security, and Networking in Azure

14

3 Demos

Day 13: Introduction to Delta Lake

WHAT IS DELTA LAKE?

- Role of Delta in Azure Databricks
- **Evolution:** From Parquet to Delta
- Delta Lake Architecture Overview
- **Key Benefits:** Reliability, Performance, and Scalability
- Batch vs. Incremental Data Processing

DEFINITION AND USE CASES

- Comparing Batch, Micro-batch, and Streaming
- When to use Incremental Data Loads
- Real-world examples in Azure Data Lake / Databricks

ACID TRANSACTIONS IN DELTA

- What are ACID properties?
- How Delta Lake ensures ACID compliance
- Transaction Logs and Checkpoints

DELTA LAKE RELIABILITY FEATURES

- Schema Enforcement and Evolution
- Data Versioning (Time Travel)
- File Compaction (Optimize Command)
- Handling Deletes, Updates, and Merges (MERGE INTO)

IMPLEMENTING DELTA WORKFLOWS AND MEDALLION ARCHITECTURE

- **Use Case:** ETL Pipeline with Bronze -> Silver -> Gold Layers
- Load Raw Data to Bronze
- Clean and Transform to Silver
- Aggregate and Publish to Gold
- ACID transaction behaviour during processing

15

Demo

Day 14: Performance Tuning and Best Practices

TOPICS

OPTIMIZE, ZORDER, and Vacuum · Partitioning Strategy · File Sizes and Auto Compaction · Delta vs. Parquet Performance Comparison

MONITORING AND DATA LINEAGE

Delta Lake Tables in Unity Catalog · Auditing and Monitoring Delta Tables · Reading Delta Logs and Transaction History · Integration with Azure Purview for Lineage · Troubleshooting common Delta failures

Day 15: Databricks Visualization and Dashboard

INTRODUCTION TO DATA VISUALIZATION IN DATABRICKS

- Importance of Visualization in Analytics
- Overview of Databricks SQL for Visualization
- Visual Layer vs. Query Layer
- Tour of Databricks SQL Editor and Dashboard Interface
- Best Practices for Designing Dashboards

WRITING QUERIES FOR VISUALIZATIONS

- Writing SQL Queries for Data Exploration
- Filtering and Aggregation for Charts
- Handling NULLs and Outliers
- Formatting Query Results for Charts

VISUALIZATION TYPES AND USE CASES

- Bar Charts, Line Charts, Area Charts
- Pie and Donut Charts
- Scatter Plots and Bubble Charts
- Choosing the Right Chart Type Based on Data

CREATING INTERACTIVE DASHBOARDS

- Adding and Arranging Visualizations
- Dashboard Filters and Global Parameters
- Time Filters and Refresh Controls
- Adding Text, Titles, and Descriptions

HANDS-ON ACTIVITIES

Practise What You Learn

#	MODULE	ACTIVITY
11	Day 11: Understanding Databricks Runtime	Hands-on Lab: Creating a Scheduled Workflow
12	Day 12: Hands-on with Workflow Creation and Management	LAB ■ Create a Realistic Multi-stage Workflow ■ Analyze Failed Jobs and Fix Issues
14	Day 13: Introduction to Delta Lake	DEMO ■ Loading Batch and Incremental Data ■ Handling Concurrent Writes and Read Isolation ■ Using Delta for Reliable Ingestion & Transformations
15	Day 14: Performance Tuning and Best Practices	Demo: Optimizing a Delta Table for Performance
16	Day 15: Databricks Visualization and Dashboard	LAB ■ Build a Base Query for Revenue by Region ■ Create Visualizations for Business KPIs ■ Build a Complete Sales Dashboard with Filters EVALUATION AND TASK ■ Practical Evaluation based on Content Covered ■ Post Test

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training

Live Online Classes (ILO)

Fly-Me-A-Trainer (FMAT)

Flexi (Self-Paced)

1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com

✉ info@koenig-solutions.com