

KOENIG SOLUTIONS • COURSE OUTLINE

Mastering Azure Databricks-From Foundations to Production Excellence

Learn to build, govern and operationalise scalable data pipelines on Azure Databricks, from Delta Lake to CI/CD and cost optimisation.

5 Days

40 HOURS

14

MODULES

77

TOPICS

6

HANDS-ON ACTIVITIES

COURSE OVERVIEW

This immersive 5-day training equips data engineers, analysts, and architects with expert-level skills to build, optimize, and operationalize scalable data pipelines on Azure Databricks. Participants dive deep into Spark-based ETL/ELT, Delta Lake, streaming, governance, orchestration, and testing strategies, enabling end-to-end deployment of enterprise-grade data solutions. Through hands-on labs, real-world scenarios, and best practices, you'll master cost-efficient, high-performance Databricks workflows in the Azure ecosystem.

WHO THIS COURSE IS FOR

Data Engineers

Data Analysts

Data Architects

DevOps Engineers

QA Engineers

PREREQUISITES

- Basic knowledge of Python or Scala programming
- Familiarity with SQL and data concepts
- Experience with Azure fundamentals (e.g., storage, compute)
- Understanding of ETL/ELT processes (recommended)

LEARNING OBJECTIVES

By the end of this course, participants will be able to:

- Architect and implement scalable data pipelines using Azure Databricks and Delta Lake
- Design ETL/ELT workflows with Spark, including structured streaming and medallion architecture
- Govern data assets with Databricks SQL and Unity Catalog
- Orchestrate jobs, implement CI/CD pipelines, and optimize for performance and cost
- Build and test data quality frameworks, integrations, and UAT-ready ETL jobs
- Apply production best practices for reliable, efficient Databricks deployments

14 Modules, 77 Topics

01

Introduction to Azure Databricks

TOPICS

- Overview of Databricks platform architecture and Azure integration
- **Key components:** clusters, notebooks, jobs, and workspaces
- **Databricks vs. open-source Spark:** managed services and benefits
- Setting up Azure Databricks workspace and first cluster
- Role-based access and security fundamentals
- Use cases in data engineering, analytics, and ML

02

Hands-on

Delta Lake Fundamentals

TOPICS

- **Core concepts:** ACID transactions, schema enforcement, and time travel
- Delta Lake architecture on Databricks
- Creating and managing Delta tables
- OPTIMIZE and Z-ORDER for performance
- **Delta Lake vs. Parquet:** advantages and migration strategies

03

Lab

ETL/ELT with Spark on Databricks

TOPICS

- Spark DataFrames and Datasets in Databricks
- **Reading/writing data:** formats, partitioning, and caching
- **Transformations:** joins, aggregations, and UDFs
- Spark SQL for ELT pipelines
- Error handling and fault tolerance in Spark jobs

04

Databricks ETL Architecture & Ingestion Patterns

TOPICS

- **Common ingestion patterns:** batch, micro-batch, streaming
- Auto Loader for incremental data ingestion
- **Partner integrations:** Azure Data Factory, Event Hubs, Blob Storage
- Scalable ingestion architectures
- Handling schema evolution and late data
- Best practices for reliable data pipelines

05**Lab**

Structured Streaming Testing

TOPICS

- **Structured Streaming model:** sources, sinks, and triggers
- Watermarking, stateful operations, and foreachBatch
- **Testing streaming queries:** unit, integration, and chaos testing
- Monitoring streaming metrics in Databricks
- Fault recovery and exactly-once semantics

06**Hands-on**

Data Modeling & Delta Lake (Bronze to Silver to Gold)

TOPICS

- **Medallion architecture:** Bronze (raw), Silver (curated), Gold (aggregated)
- Implementing layers with Delta Live Tables
- Data quality checks at each layer
- Change Data Capture (CDC) patterns
- Versioning and auditing across layers

07

Databricks SQL & Unity Catalog Governance

TOPICS

- **Databricks SQL:** warehouses, queries, and dashboards
- **Unity Catalog:** three-level namespace and metastore
- Data lineage, sharing, and access controls
- **Governance policies:** tagging, auditing, and compliance
- Query federation and external tables
- Securing sensitive data with dynamic views

08

Lab

Orchestration (Databricks Jobs, Workflows, Triggers)

TOPICS

- **Databricks Jobs:** task graphs, dependencies, and scheduling
- Workflows with multi-task jobs and parameters
- **Triggers:** file arrival, cron, and external APIs
- Delta Live Tables for declarative pipelines
- Error handling, retries, and alerts

09

CI/CD & DevOps for Databricks

TOPICS

- Git integration and notebook versioning
- Databricks Asset Bundles (DABs) for IaC
- CI/CD pipelines with Azure DevOps or GitHub Actions
- Environment promotion: dev/staging/prod
- Automated testing in pipelines
- Branching strategies for collaborative development

10

Performance Testing on Databricks

TOPICS

- Cluster sizing, autoscaling, and spot instances
- **Spark UI analysis:** stages, tasks, and spills
- Benchmarking queries with Delta Lake optimizations
- Photon engine for accelerated performance
- Load testing tools and frameworks
- Profiling and tuning labs

11

Cost Optimization Testing

TOPICS

- **Cost monitoring:** clusters, DBUs, and usage reports
- Auto-termination, scaling policies, and job clusters
- Predicting costs with Databricks tools
- **Optimization techniques:** caching, broadcast joins
- Testing scenarios for cost vs. performance trade-offs
- Governance for budget controls

12

Hands-on

Data Quality & Validation Frameworks

TOPICS

- Great Expectations and Delta Expectations integration
- Custom validation rules in Spark/Delta Live Tables
- Anomaly detection and data profiling
- Automated quality gates in pipelines
- Reporting and alerting on quality metrics

13

End-to-End Cloud Integration Testing

TOPICS

- **Integrating with Azure services:** Synapse, Purview, Sentinel
- **Security testing:** RBAC, encryption, and network isolation
- API testing for Databricks endpoints
- Mocking external dependencies
- Cross-service data flows and latency testing
- Full E2E test automation

14

ETL Jobs Testing & UAT Preparation

TOPICS

- Unit testing Spark code with PyTest/ScalaTest
- Integration and end-to-end ETL testing
- **UAT checklists:** data accuracy, performance SLAs
- Mock data generation and regression testing
- Deployment readiness and rollback strategies

HANDS-ON ACTIVITIES

Practise What You Learn

#	MODULE	ACTIVITY
02	Delta Lake Fundamentals	Hands-on: Building initial Delta tables from data sources
03	ETL/ELT with Spark on Databricks	Lab: End-to-end ETL pipeline with sample datasets
05	Structured Streaming Testing	Lab: Real-time streaming pipeline with testing

#	MODULE	ACTIVITY
06	Data Modeling & Delta Lake (Bronze to Silver to Gold)	Hands-on: Multi-layer pipeline build
08	Orchestration (Databricks Jobs, Workflows, Triggers)	Lab: Multi-step workflow orchestration
12	Data Quality & Validation Frameworks	Hands-on: Building a DQ framework

WHAT IS INCLUDED

Everything You Need to Succeed

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