# DP-3027 : Implement a Data Engineering Solution with Azure Databricks

**Course Description:**

Learn how to harness the power of Apache Spark and scalable clusters running on the Azure Databricks platform to manage and execute large-scale data engineering workloads in the cloud. This learning path covers real-time streaming, Delta Live Tables, performance tuning, CI/CD automation, data governance, and integration with other Azure services.

**Duration:** 8 hours

**Learning Objectives:**

- Implement incremental and streaming data processing with Spark and Delta Live Tables
- Optimize performance and manage costs in data pipelines
- Automate workflows using Databricks Jobs and CI/CD practices
- Govern data access, quality, and compliance with Unity Catalog
- Use SQL Warehouses for query-based analytics
- Integrate Azure Databricks with Azure Data Factory for end-to-end orchestration

**Content Coverage :**

**Module 1: Perform Incremental Processing with Spark Structured Streaming**

- Introduction
- Set up real-time data sources
- Optimize Delta Lake for incremental processing
- Handle late data and out-of-order events
- Monitoring and performance tuning
- Exercise: Real-time ingestion and processing with Delta Live Tables

**Module 2: Implement Streaming Architecture Patterns with Delta Live Tables**

- Introduction
- Event-driven architectures
- Ingest data with structured streaming
- Maintain data consistency and reliability
- Scale streaming workloads
- Exercise: End-to-end streaming pipeline

**Module 3: Optimize Performance with Spark and Delta Live Tables**

- Introduction

- Optimize performance with Spark and Delta
- Perform cost-based optimization and query tuning
- Use change data capture (CDC)
- Use enhanced autoscaling
- Implement observability and data quality metrics
- Exercise: Optimize data pipelines

**Module 4: Implement CI/CD Workflows in Azure Databricks**

- Introduction
- Version control and Git integration
- Unit testing and integration testing
- Environment configuration
- Rollback and roll-forward strategies
- Exercise: CI/CD workflows

**Module 5: Automate Workloads with Azure Databricks Jobs**

- Introduction
- Implement job scheduling and automation
- Optimize workflows with parameters
- Handle dependency management
- Implement error handling and retry logic
- Best practices and guidelines
- Exercise: Automate data ingestion and processing

**Module 6: Manage Data Privacy and Governance with Azure Databricks**

- Introduction
- Data encryption techniques
- Access control
- Data masking and anonymization
- Compliance frameworks and secure sharing
- Data lineage and metadata management
- Governance automation
- Exercise: Unity Catalog implementation

**Module 7: Use SQL Warehouses in Azure Databricks**

- Introduction
- Get started with SQL Warehouses
- Create databases and tables
- Create queries and dashboards
- Exercise: Use a SQL Warehouse

**Module 8: Run Azure Databricks Notebooks with Azure Data Factory**

- Introduction
- Understand notebooks and pipelines
- Create linked services
- Use notebook activities in pipelines
- Use parameters in notebooks
- Exercise: Run notebooks with Azure Data Factory