**Data Engineering with Databricks**

**Duration: 40 Hours**

**Prerequisites:**

* Familiarity with Python programming.
* Basic understanding of SQL.
* General knowledge of data processing and analytics.

**Scope of the Training:**

* Provide participants with foundational skills for building data engineering solutions.
* Enable participants to work effectively with Databricks and Spark for ETL workflows.
* Build confidence in handling real-world data engineering challenges.

**Course Content**

**🔹 Module 1: Databricks Fundamentals and Platform Overview**

**Estimated Duration**: 3.5 hours

**Topics**:

* What is Databricks? Overview of the Lakehouse Platform
* Key components: Workspace, Clusters, Notebooks, Jobs, Repos
* Understanding DBUs and cluster pricing
* Navigating the Databricks UI
* Introduction to PySpark and notebooks

**Hands-On**:

* Launching a Databricks workspace
* Creating and configuring a cluster
* Writing and executing PySpark code in notebooks
* Using widgets, markdown, and visualizations

**🔹 Module 2: Data Engineering with Delta Lake**

**Estimated Duration**: 6 hours

**Topics**:

* Spark architecture essentials: DAGs, partitions, and execution
* Loading and transforming data with PySpark
* File formats: CSV, JSON, Parquet, Delta
* Delta Lake: ACID compliance, versioning, schema enforcement
* Partitioning and optimization techniques

**Hands-On**:

* Load and transform datasets using PySpark
* Save outputs as Delta tables
* Convert Parquet to Delta
* Handle missing or bad data during transformation

**🔹 Module 3: Real-time Processing with Structured Streaming & Delta Live Tables**

**Estimated Duration**: 6 hours

**Topics**:

* Difference between batch and streaming processing
* Structured Streaming architecture and triggers
* Delta Live Tables (DLT): concepts, syntax, and benefits
* Bronze–Silver–Gold streaming architecture

**Hands-On**:

* Stream data from files into Delta Lake
* Apply window functions and aggregations
* Create a basic DLT pipeline with quality expectations
* Monitor DLT pipelines and results

**🔹 Module 4: Workflow Orchestration and Production Pipelines**

**Estimated Duration**: 4 hours

**Topics**:

* Introduction to Databricks Jobs and Workflows
* Task dependencies, retries, parameters
* CI/CD concepts in Databricks context
* Overview of deployment best practices

**Hands-On**:

* Build a multi-step workflow using notebooks
* Schedule pipelines and handle errors
* Trigger workflows on events or schedules

**🔹 Module 5: DevOps and CI/CD Essentials for Data Engineering**

**Estimated Duration**: 4 hours

**Topics**:

* Introduction to Git and Databricks Repos
* Branching strategies, pull requests, and merges
* Using Databricks CLI for automation

**Hands-On**:

* Link a notebook to a Git repo
* Perform basic Git operations (commit, push, branch)
* Use CLI to deploy and run notebooks

**🔹 Module 6: Monitoring and Performance Optimization**

**Estimated Duration**: 4 hours

**Topics**:

* Understanding Spark UI: stages, tasks, DAG visualization
* Common issues: skewed data, wide transformations
* Performance tips: broadcast joins, caching, checkpointing
* Using cluster pools and autoscaling
* Usage and cost monitoring with DBU tracking

**Hands-On**:

* Analyze job performance using Spark UI
* Apply optimizations (e.g., partitioning, broadcast join)
* Configure autoscaling and spot instances
* Review usage metrics and identify cost drivers

**🔹 Module 7: Platform Administration and Governance**

**Estimated Duration**: 5 hours

**Topics**:

* Admin console and workspace settings
* User and group management (including SCIM and SSO overview)
* Cluster policies and access control
* Token and secret management
* Service principals and workspace audit logs

**Hands-On**:

* Configure user permissions and groups
* Apply cluster access policies
* Review and interpret audit logs
* Manage workspace secrets and tokens

**🔹 Module 8: Data Privacy and Governance with Unity Catalog**

**Estimated Duration**: 4.5 hours

**Topics**:

* Introduction to Unity Catalog: catalogs, schemas, and tables
* Access control: user, group, table, and column-level security
* Auditing and logging of data access
* Overview of encryption, masking, and compliance frameworks (GDPR, HIPAA)

**Hands-On**:

* Set up Unity Catalog with managed access
* Apply table- and column-level permissions
* Perform audit queries and view lineage
* Secure data for compliance scenarios

**🔹 Module 9: AI Enablement using Databricks**

**Estimated Duration**: 4 hours

**Topics**:

* Overview of ML lifecycle and MLflow
* Experiment tracking and metrics
* Model registry and deployment basics
* Feature Store and AutoML preview

**Hands-On**:

* Train a sample ML model and track with MLflow
* Register a model and run batch inference
* View run history and performance