Azure Databricks Data Engineering

Duration : 40 hours

📅 Day 1 – Databricks and Lakehouse Foundations

1. Introduction to the Lakehouse Architecture

- What is a Lakehouse and how it unifies lakes and warehouses
- Lakehouse vs. Data Lake vs. Data Warehouse
- Databricks control plane vs. data plane architecture

2. Navigating the Databricks Environment

- Databricks UI: Notebooks, Repos, Jobs, Data, Compute
- All-purpose vs. Job clusters
- Multi-language support (SQL, Python, Scala)
- Git integration for notebook version control

3. Introduction to Delta Lake

- What is Delta Lake and how it enhances the data lake
- Creating Delta tables (managed vs. external)
- ACID operations: insert, update, delete, merge
- Table versioning and schema evolution

4. Databricks SQL Essentials

- Executing SQL queries in the SQL workspace
- Views, temporary views, CTEs
- Visualizations and dashboards (intro level)

🧪 Case Study: Analyze operational data, implement version control, and explore ACID behavior using Delta tables.

🗓️ Day 2 – ELT with Spark SQL and Python (PySpark)

5. Working with Spark DataFrames and Spark SQL

- Spark DataFrames and SQL
- Load Dataset from Data Catalog
- Load Data into DataFrames Using PySpark
- Load Dataset Using Spark SQL and Create Table

6. Table Management and Schema Handling

- Create Temporary and Permanent Tables Using a Notebook
- Demonstrate Custom Schema
- Select and Index/Slice Columns in PySpark DataFrames

7. Developing ELT Pipelines

- ELT Pipeline Development with DataFrames and Spark SQL
- Demonstrate Data Cleaning and Preprocessing
- Demonstrate Data Wrangling and Transformation
- Demonstrate Data Wrangling Using Filters and Sorting
- Writing transformed outputs to Delta

🧪 Case Study: Create an ELT pipeline to clean and transform customer and order data using Spark SQL and PySpark DataFrames.

🗓️ Day 3 – Incremental Loads and Scalable Architecture

8. Structured Streaming and Incremental Processing

- Structured Streaming concepts
- Real-time ingestion with Auto Loader
- Deduplication using MERGE
- Watermarking, late-arriving data, checkpointing

9. Implementing Medallion Architecture

- Bronze, Silver, and Gold layer design
- Building modular and scalable pipelines

10. Delta Live Tables (DLT)

- DLT concepts and pipeline creation
- Triggered vs. continuous pipelines
- APPLY CHANGES INTO for CDC
- Debugging and monitoring DLT pipelines

🧪 Case Study: Stream sales logs into a Medallion architecture and create a Gold layer analytics view with deduplication and DLT.

🗓️ Day 4 – Optimization, and Monitoring

11. Optimizing Delta Lake Performance

- Z-ordering, file compaction, and VACUUM
- Schema evolution and constraints
- ON VIOLATION clause and conflict handling
- Best practices for table size and layout

12. Workflow Orchestration and Monitoring

- Creating multi-task workflows using Jobs
- Setting dependencies, retries, scheduling
- Logs, alerts, execution history

13. Monitoring and Logging Capabilities

- Monitoring pipeline health
- Integration with workspace logs and cluster metrics
- Logging strategies in notebooks and jobs

14. Cost Optimization and Performance Tuning

- Cluster tuning and scaling

- Job execution planning
- Storage optimization (cache, Auto Loader checkpointing)
- Demo: Schedule Automation and Optimization

🧪 Case Study: Build, optimize, and monitor an e-commerce analytics pipeline with performance tuning and alerting.

🗓️ Day 5 – Governance, Business Reporting, and Deployment

15. Databricks SQL for Business Intelligence

- Writing analytical queries
- Building interactive dashboards and reports
- Sharing dashboards securely and scheduling refreshes

16. Unity Catalog and Access Control

- Catalogs, schemas, and tables in Unity Catalog
- Role-based access control (RBAC)
- Granting/revoking permissions (USAGE, SELECT)
- Cluster modes and data isolation with service principals

17. Deployments in Azure Databricks using Asset Bundles

- Deployments with Databricks Asset Bundles
- Introduction to Asset Bundles for CI/CD
- Packaging notebooks, pipelines, and jobs for deployment
- Automating deployment using workspace Git

🧪 Case Study: Deploy a full-featured data pipeline to QA and Production using asset bundles, access controls, and automated scheduling.

✅ Program Outcome

By the end of this 5-day program, participants will be able to:

- Build robust, scalable ELT pipelines in Databricks using PySpark and Spark SQL
- Leverage Delta Lake and Delta Live Tables for both batch and streaming data
- Optimize pipeline performance and cost for real-world workloads
- Monitor and orchestrate data pipelines effectively
- Implement secure deployments using Databricks Asset Bundles and Unity Catalog