multisoft
Systems

Databricks Certified Data Engineer Professional Training

COURSE CONTENT

GET IN TOUCH

Multisoft Systems
B - 125, Sector - 2, Noida

(+91) 9810-306-956

info@multisoftsystems.com

www.multisoftsystems.com

About Multisoft

Train yourself with the best and develop valuable in-demand skills with Multisoft Systems. A leading certification training provider, Multisoft collaborates with top technologies to bring world-class one-on-one and certification trainings. With the goal to empower professionals and business across the globe, we offer more than 1500 training courses, which are delivered by Multisoft's global subject matter experts. We offer tailored corporate training; project Based Training, comprehensive learning solution with lifetime e-learning access, after training support and globally recognized training certificates.

About Course

The Databricks Certified Data Engineer Professional Training by Multisoft Systems is designed to equip professionals with the advanced knowledge and hands-on expertise required to excel in data engineering. This training focuses on building, optimizing, and managing large-scale data pipelines using Databricks and its integrated ecosystem.

Module 1: Databricks Tooling

- ✓ Explain how Delta Lake uses the transaction log and cloud object storage to guarantee atomicity and durability
- ✓ Describe how Delta Lake's Optimistic Concurrency Control provides isolation, and which transactions might conflict
- ✓ Describe basic functionality of Delta clone.
- ✓ Apply common Delta Lake indexing optimizations including partitioning, zorder, bloom filters, and file sizes
- ✓ Implement Delta tables optimized for Databricks SQL service
- ✓ Contrast different strategies for partitioning data (e.g. identify proper partitioning columns to use)

Module 2: Data Processing (Batch processing, Incremental processing, and Optimization)

- ✓ Describe and distinguish partition hints: coalesce, repartition, repartition by range, and rebalance
- ✓ Articulate how to write Pyspark dataframes to disk while manually controlling the size of individual part-files.
- ✓ Articulate multiple strategies for updating 1+ records in a spark table
- ✓ Implement common design patterns unlocked by Structured Streaming and Delta Lake.
- ✓ Explore and tune state information using stream-static joins and Delta Lake
- ✓ Implement stream-static joins
- ✓ Implement necessary logic for deduplication using Spark Structured Streaming
- ✓ Enable CDF on Delta Lake tables and re-design data processing steps to process CDC output instead of incremental feed from normal Structured Streaming read
- ✓ Leverage CDF to easily propagate deletes
- ✓ Demonstrate how proper partitioning of data allows for simple archiving or deletion of data

- ✓ Articulate, how "smalls" (tiny files, scanning overhead, over partitioning, etc) induce performance problems into Spark queries

Module 3: Data Modeling

- ✓ Describe the objective of data transformations during promotion from bronze to silver
- ✓ Discuss how Change Data Feed (CDF) addresses past difficulties propagating updates and deletes within Lakehouse architecture
- ✓ Design a multiplex bronze table to avoid common pitfalls when trying to productionalize streaming workloads.
- ✓ Implement best practices when streaming data from multiplex bronze tables.
- ✓ Apply incremental processing, quality enforcement, and deduplication to process data from bronze to silver
- ✓ Make informed decisions about how to enforce data quality based on strengths and limitations of various approaches in Delta Lake
- ✓ Implement tables avoiding issues caused by lack of foreign key constraints
- ✓ Add constraints to Delta Lake tables to prevent bad data from being written
- ✓ Implement lookup tables and describe the trade-offs for normalized data models
- ✓ Diagram architectures and operations necessary to implement various Slowly Changing Dimension tables using Delta Lake with streaming and batch workloads.
- ✓ Implement SCD Type0, 1, and 2 tables

Module 4: Security & Governance

- ✓ Create Dynamic views to perform data masking
- ✓ Use dynamic views to control access to rows and columns

Module 5: Monitoring & Logging

- ✓ Describe the elements in the Spark UI to aid in performance analysis, application debugging, and tuning of Spark applications
- ✓ Inspect event timelines and metrics for stages and jobs performed on a cluster

- ✓ Draw conclusions from information presented in the Spark UI, Ganglia UI, and the Cluster UI to assess performance problems and debug failing applications.
- ✓ Design systems that control for cost and latency SLAs for production streaming jobs
- ✓ Deploy and monitor streaming and batch jobs

Module 6: Testing & Deployment

- ✓ Adapt a notebook dependency pattern to use Python file dependencies
- ✓ Adapt Python code maintained as Wheels to direct imports using relative paths
- ✓ Repair and rerun failed jobs
- ✓ Create Jobs based on common use cases and patterns
- ✓ Create a multi-task job with multiple dependencies
- ✓ Configure the Databricks CLI and execute basic commands to interact with the workspace and clusters
- ✓ Execute commands from the CLI to deploy and monitor Databricks jobs
- ✓ Use REST API to clone a job, trigger a run, and export the run output