Databricks Certified Associate Developer for Apache Training

Curriculum 3 Days

Day - 1

- **Introduction to Apache Spark and Databricks**
  - Overview of Apache Spark: Architecture, components, and use cases
  - Introduction to Databricks: Why Databricks for Spark? Key features, architecture, and tools
  - Understanding Databricks Notebooks, Clusters, and Jobs
  - Hands-on Lab: Setting up a Databricks workspace and creating clusters.
- **Spark Basics and Core Concepts**
  - RDDs (Resilient Distributed Datasets) and DataFrames: Spark’s core abstractions
  - Spark SQL and its integration with DataFrames and Datasets
  - Actions and Transformations in Spark: Key operations for distributed data processing
  - Understanding the Spark execution model: Lazy evaluation and stages
  - Hands-on Lab: Writing simple Spark SQL queries and basic transformations with DataFrames.
- **Spark Programming with RDDs and DataFrames**
  - Working with RDDs: Creating, transforming, and performing actions on RDDs
  - Converting between RDDs and DataFrames
  - Advantages of DataFrames over RDDs in terms of performance and ease of use
  - Hands-on Lab: Implementing Spark RDD operations and DataFrame transformations.
- **Spark Session and SparkSQL**
  - Introduction to SparkSession: The entry point for Spark in Databricks
  - Using SparkSQL for querying structured data
  - Connecting to external data sources (e.g., JDBC, Hive, Parquet files)
  - Hands-on Lab: Running SparkSQL queries on structured data stored in Databricks

Day - 2

- **Data Sources and I/O Operations**
  - Understanding how Spark interacts with different data sources
  - Reading and writing data from/to CSV, Parquet, JSON, JDBC, and Hive
  - Handling data partitions and optimizing I/O operations
  - Hands-on Lab: Working with different data sources and performing read/write operations.

- **Spark Transformations and Performance Optimization**
  - Transformations: Narrow vs. wide transformations
  - Optimizing shuffling in Spark: Reducing data movement for better performance
  - Working with caching and persisting DataFrames to speed up iterative queries
  - Understanding partitioning strategies: How to improve performance with partitioning
  - Hands-on Lab: Implementing and optimizing Spark transformations and caching strategies.

- **Working with Spark Streaming**
  - Introduction to Spark Streaming and its use in real-time data processing
  - DStreams vs. Structured Streaming: When to use each
  - Streaming DataFrames and event-time processing
  - Hands-on Lab: Setting up a basic Structured Streaming job in Databricks to process streaming data.

- **Spark Session and SparkSQL**
  - Introduction to MLlib for scalable machine learning in Spark
  - Overview of machine learning algorithms: Classification, Regression, Clustering
  - Feature engineering and model evaluation in Spark
  - Hands-on Lab: Building a machine learning pipeline with Spark MLlib

Day - 3

- **Advanced Spark Concepts**
  - Spark SQL optimization: Catalyst Optimizer and Tungsten
  - Understanding the Broadcast Join and Shuffle Join strategies
  - Tuning Spark Jobs: Memory management, task scheduling, and Spark UI
  - Hands-on Lab: Implementing advanced optimization techniques using Spark SQL and joins.

- **Distributed Processing with Spark**
  - Understanding task execution and data shuffling in a distributed environment
  - Spark Executors: Managing resources and parallelism
  - Fault tolerance in Spark: Checkpointing and handling task failures
  - Hands-on Lab: Executing and debugging Spark jobs on a cluster

- **Job Orchestration with Databricks**
  - Orchestrating Spark jobs using Databricks Workflows
  - Scheduling, managing dependencies, and ensuring fault tolerance
  - Using MLflow for managing machine learning experiments and models
  - Hands-on Lab: Creating and orchestrating jobs with Databricks Workflows and tracking with MLflow.

- **Spark Session and SparkSQL**
  - Review of exam topics and best practices for Databricks Certified Associate Developer for Apache Spark Exam
  - Test-taking strategies: Time management and question approach
  - Hands-on Mock Exam: A series of multiple-choice and scenario-based questions
  - Answer review and detailed explanations of the mock exam results.