Databricks Advance Training Course

Curriculum 5 Days

Day - 1

- **Review of Databricks Fundamentals & Key Concepts**
  - Recap of Databricks basics: Databricks Notebooks, Clusters, Jobs, and Workflows
  - Overview of Apache Spark architecture and Databricks' optimizations
  - Key components: Delta Lake, MLflow, Unity Catalog, and Databricks SQL.
- **Advanced Spark Performance Optimization**
  - Spark Tuning: Partitioning strategies, memory management, and Spark configurations
  - Catalyst Optimizer and Tungsten for performance improvements
  - Optimizing shuffling and join strategies: Shuffle partitions, Broadcast joins, and Optimizing joins
  - Hands-on Lab: Implementing Spark performance optimizations in Databricks.
- **Delta Lake Advanced Features**
  - Understanding Delta Lake: ACID transactions, time travel, and schema evolution
  - Delta Table Optimizations: Z-Ordering, OPTIMIZE, and VACUUM
  - Change Data Capture (CDC) with Delta Lake
  - Hands-on Lab: Using Delta Lake for efficient data storage and optimized queries in Databricks.
- **Advanced Data Engineering Workflows**
  - Orchestrating complex data workflows in Databricks Workflows
  - Scheduling and dependency management of Databricks Jobs
  - MLflow integration for managing machine learning pipelines within data workflows
  - Hands-on Lab: Building an end-to-end data engineering pipeline using Databricks Workflows

Day - 2

- **Advanced Machine Learning Techniques with Databricks**
  - Introduction to MLlib for scalable machine learning models
  - Building custom machine learning models with SparkML and MLflow
  - Hyperparameter tuning with GridSearch and RandomSearch in Databricks
  - Using Spark's distributed computing for training large models
  - Hands-on Lab: Building and tuning a classification model using SparkML.
- **Model Training and Experimentation with MLflow**
  - Managing machine learning experiments using MLflow: Logging metrics, parameters, and models
  - Model versioning and comparison in MLflow Model Registry
  - Model deployment using MLflow's Model Serving feature
  - Hands-on Lab: Managing experiments and deploying a machine learning model using MLflow.
- **Advanced Spark Streaming and Real-Time Data Pipelines**
  - Overview of Structured Streaming vs. DStreams in Apache Spark
  - Real-time processing with Spark Streaming and Delta Lake
  - Integrating Databricks with Apache Kafka for real-time data streaming
  - Hands-on Lab: Setting up a real-time streaming job using Structured Streaming and Kafka.
- **Advanced Data Governance and Security**
  - Using Unity Catalog for fine-grained data governance and security policies
  - Managing access control: Users, groups, and permissions
  - Best practices for ensuring data privacy and security in Databricks
  - Hands-on Lab: Setting up Unity Catalog and applying access controls for data governance.

Day - 3

- **Scaling Databricks for Big Data**
  - Scaling Databricks clusters for high-performance computing: Autoscaling, managing large jobs, and cost optimization
  - Spark Cluster Management: Managing resources and distributing workloads efficiently
  - Data Partitioning: Optimizing read and write performance for massive datasets
  - Hands-on Lab: Scaling a large-scale data processing job in Databricks.
- **Building and Optimizing ETL Pipelines at Scale**
  - Designing ETL pipelines using Databricks Workflows for large-scale data processing
  - Integrating Delta Lake with ETL jobs for real-time data pipelines
  - Job scheduling, monitoring, and error handling in Databricks
  - Hands-on Lab: Building a scalable ETL pipeline using Delta Lake and Databricks Workflows
- **Advanced Databricks Integration with Cloud Services**
  - Integrating Databricks with cloud storage (AWS S3, Azure Data Lake Storage, Google Cloud Storage)
  - Working with cloud data warehouses (e.g., Redshift, BigQuery) in Databricks
  - Using Databricks Connect for seamless integration with other data processing platforms
  - Hands-on Lab: Integrating Databricks with AWS S3 and querying data from a cloud data warehouse.
- **Databricks MLOps and Model Deployment at Scale**
  - Introduction to MLOps in Databricks: Automating ML workflows and managing model lifecycles
  - Model deployment at scale using MLflow and Databricks Model Serving
  - Managing continuous integration and continuous deployment (CI/CD) for machine learning models
  - Hands-on Lab: Implementing MLOps for an ML model using Databricks.

Day - 4

- **Scaling Spark for Large-Scale Data Processing**
  - Scaling Databricks clusters for big data processing: Auto-scaling and cluster optimization
  - Dynamic Allocation of resources for performance optimization
  - Managing large-scale jobs using Databricks Jobs
  - Hands-on Lab: Scaling Spark jobs for large datasets in Databricks.
- **Advanced Data Management and Governance**
  - Managing data access and security with Unity Catalog
  - Fine-grained access control and auditing in Databricks
  - Implementing data governance with Delta Lake and Unity Catalog
  - Hands-on Lab: Configuring Unity Catalog for data governance in Databricks
- **Data Versioning, Time Travel, and Change Data Capture (CDC)**
  - Understanding CDC and implementing it with Delta Lake
  - Time Travel with Delta Lake: Querying historical data versions
  - Hands-on Lab: Implementing CDC and using time travel in Delta Lake.
- **Cost Optimization Strategies in Databricks**
  - Optimizing cost and resource usage in Databricks
  - Techniques for job optimization, cluster sizing, and auto-scaling
  - Managing costs with Databricks Jobs and Workflows
  - Hands-on Lab: Implementing cost optimization strategies for large-scale Databricks jobs.

Day - 5

- **Scaling Databricks for Big Data**
  - Scaling Databricks clusters for high-performance computing: Autoscaling, managing large jobs, and cost optimization
  - Spark Cluster Management: Managing resources and distributing workloads efficiently
  - Data Partitioning: Optimizing read and write performance for massive datasets
  - Hands-on Lab: Scaling a large-scale data processing job in Databricks.
- **Building and Optimizing ETL Pipelines at Scale**
  - Designing ETL pipelines using Databricks Workflows for large-scale data processing
  - Integrating Delta Lake with ETL jobs for real-time data pipelines
  - Job scheduling, monitoring, and error handling in Databricks
  - Hands-on Lab: Building a scalable ETL pipeline using Delta Lake and Databricks Workflows
- **Advanced Databricks Integration with Cloud Services**
  - Integrating Databricks with cloud storage (AWS S3, Azure Data Lake Storage, Google Cloud Storage)
  - Working with cloud data warehouses (e.g., Redshift, BigQuery) in Databricks
  - Using Databricks Connect for seamless integration with other data processing platforms
  - Hands-on Lab: Integrating Databricks with AWS S3 and querying data from a cloud data warehouse.
- **Databricks MLOps and Model Deployment at Scale**
  - Introduction to MLOps in Databricks: Automating ML workflows and managing model lifecycles
  - Model deployment at scale using MLflow and Databricks Model Serving
  - Managing continuous integration and continuous deployment (CI/CD) for machine learning models
  - Hands-on Lab: Implementing MLOps for an ML model using Databricks.