Course Duration: 20 hours (3 Days)

Databricks on AWS for Professionals

This course provides a comprehensive understanding of Databricks on AWS, covering big data processing, building pipelines, and data engineering. Participants will learn how to leverage Databricks to process large datasets, run analytics workloads, and build automated pipelines within an AWS cloud environment.

Course objectives

By the end of this course, participants will be able to:

- Understand the Databricks architecture on AWS and its key components.
- Set up and configure Databricks with AWS services (S3, Glue, IAM, etc.).
- Use Apache Spark for big data processing, transformations, and optimizations.
- Build and optimize data pipelines for ETL workflows and real-time analytics.

Prerequisites

- Completed AWS Cloud Practitioner Essentials, or AWS Technical Essentials.
- Programming experience using Python.
- Data Engineering experience using Spark.
- Ability to write and interpret SQL Queries.
- Valid AWS Account.
- Valid Databricks Account (free Databricks Account is not sufficient).

Target Audience

- DevOps Engineers
- Data Engineers
- Testers
- Cloud Professionals

KOENIG
step forward

Course outline

Module 1: Fundamentals of Data Engineering

- Understanding Data Engineering Concepts and Principles
    - Core Concepts and Principles in Data Engineering
    - Overview of Data Pipelines, Data Integration, and Data Transformation
- Overview of Databricks
    - Databricks as a Unified Analytics Platform
    - Key Features and Benefits of Using Databricks for Data Engineering
    - Overview of Databricks Architecture and Components
    - Components of Databricks Architecture
    - Understanding the Databricks Workspace and Its Functionalities
    - Introducing Databricks Notebooks and Its Role in Data Engineering

Module 2: Setting Up Databricks Environment and Workspace

- Databricks Utilities
    - File System Utilities (dbutils.fs)
    - Library Utility (dbutils.library)
    - Notebook Utility (dbutils.notebook)
    - Secrets Utility (dbutils.secrets)
    - Widgets Utility (dbutils.widgets)
- Configuring Databricks Clusters
    - Selection Criteria for Different Workloads
    - Databricks Runtime Versions
    - Cluster Sizing
    - Cluster Sizing Considerations
    - Cluster Sizing Examples
- Differences Between Standard and High Concurrency Clusters

- SetUp Databricks Workspace on AWS
- Databricks UI walkthrough
- Configure Databricks cluster on AWS
- Bring your data to Databricks UI Dashboard
- Creating Databricks Notebooks
- Develop Spark application using notebook

Module 3: Delta Lakes and Delta Tables

- Understanding the Benefits of Delta Tables in Databricks
- Enhanced Data Reliability and Consistency with Delta Tables
- Efficient Data Processing and Query Optimization
- Overview of Delta Lake Architecture and Concepts
    - Introduction to Delta Lake
    - Delta Lake Architecture

Module 4: Data Ingestion and Extraction

- Ingesting Data into Databricks
- Understanding the Importance of Data Ingestion
- Choosing the Appropriate Data Ingestion Method
- Identifying Key Considerations for Data Ingestion
- Implementing Data Ingestion Best Practices

Module 5: Data Pipelines with Databricks

- Overview of the ETL Process
- Reading Data from Different Sources in Databricks
- Using Pre-built Connectors in ETL Pipeline Tool
- Building Scalable Data Transformation Pipelines
    - Understanding the Importance of Data Transformation in Data Engineering
    - Designing Scalable and Efficient Data Transformation Pipelines

- Techniques for Data Transformation in Databricks
- Applying ETL Methodologies
  - Overview of the ETL Process and Its Key Components
- Optimizing Data Processing and Performance in Databricks
  - Strategies for Optimizing Data Processing in Databricks
  - Leveraging Databricks Runtime Configurations for Performance Improvements
  - Performance Tuning for Data Transformation Operations in Databricks

Module 6: Data Orchestration and Workflow Management

- Implementing Workflow Automation with Databricks
  - Overview of Workflow Automation
  - Introduction to Databricks Jobs and Notebooks for Workflow Automation
  - Designing and Implementing Automated Data Pipelines in Databricks
- Managing Dependencies and Scheduling Data Pipelines
  - Understanding Dependencies Between Data Pipelines and Tasks
  - Techniques for Managing Dependencies in Databricks Workflows
  - Scheduling and Orchestrating Data Pipelines Using Databricks Jobs
  - Best Practices for Handling Complex Workflows and Task Dependencies
- Monitoring and Error Handling in Workflow Execution
  - Strategies for Monitoring Workflow Execution
  - Implementing Logging and Alerting Mechanisms for Error Detection
  - Techniques for Handling Workflow Failures and Retries
  - Utilizing Databricks Monitoring and Debugging Tools for Workflow Optimization

Module 7: Data Security and Governance

- Unity Catalog
  - Understanding the Unity Catalog in Databricks
  - Overview of Metadata Management and Data Discovery in Unity Catalog

- Leveraging Unity Catalog for Efficient Data Governance and Metadata Management
- Ensuring Data Privacy and Compliance in Databricks
  - Importance of Data Privacy and Compliance
  - Implementing Privacy Measures and Techniques in Databricks
  - Ensuring Compliance with Data Protection Regulations
  - Techniques for Anonymization, Pseudonymization, and Data Masking
- Implementing Access Controls and Data Encryption
  - Overview of Access Controls and Authorization in Databricks
  - Designing and Implementing Access Policies for Data Protection
  - Techniques for Encrypting Data at Rest and in Transit in Databricks
  - Implementing Key Management and Secure Credential Storage Practices
- Data Governance Best Practices
  - Importance of Data Governance
  - Techniques for Data Lineage, Metadata Management, and Data Cataloging
  - Best Practices for Data Documentation, Stewardship, and Data Lifecycle Management