Advanced Big Data Analytics with AWS Databricks
Duration: 5 Days

Objective: The primary goal of this training is to enable data engineers, data professionals and data analysts understand AWS offerings to Databricks features, how they really work, architecture and components that make the ecosystem work. And importantly practically use those technologies which would help to solving enterprise problems related data movement, analytics and engineering.

At the end of this course, participants would be able to understand

- AWS Technologies and Service Offerings
- AWS Data Lake Store
- Spark
- ADB with Python
- AWS Databricks – Deep-dive

Detailed Course Contents:

Day 1

S3 Introduction
Working with S3 buckets

Spark Introduction
Data Bricks
DataBricks & Cloud Architecture

AWS S3 and Data Lake Store Offerings
Introduction
Key Capabilities
Securing data in AWS Data Lake Store
Applications compatible with AWS Data Lake Store
What is AWS Data Lake Store file system (adl://)?
How do I start using AWS Data Lake Store?

Using AWS Data Lake Store for big data requirements
Ingest data into Data Lake Store
Process data stored in Data Lake Store
Download data from Data Lake Store
Visualize data in Data Lake Store

Apache Spark
DataFrames and Datasets
Introduction to DataFrames - Python
Introduction to DataFrames - Scala
Introduction to Datasets

Day 2

- Complex and Nested Data
- Aggregators
- Structured Streaming
- Introductory Notebooks
- Streaming Data Sources and Sinks
- Structured Streaming in Production
- Examples
- Spark Streaming (Legacy)

SQL

- SQL Language Manual
- Spark SQL Examples
- Compatibility with Apache Hive

Day 3

What is AWS Databricks?

- Create Databricks workspace - Portal
- Create Databricks workspace - Resource Manager template
- Create Databricks workspace - Virtual network

Get started with AWS Databricks

- Data overview
- AWS Databricks concepts
- AWS Databricks datasets

Runtime overview

- Databricks Runtime

Workspaces

- Explore the Databricks workspace
- Workspace assets
- Work with workspace objects
- Get workspace, cluster, notebook, and job identifiers

Clusters

- Clusters overview
- Create a cluster
- Manage clusters
- Configure clusters
- Initialize cluster nodes
- Custom containers
- GPU-enabled clusters
- Types of Clusters
    - Interactive
    - High Concurrency
    - Job Clusters
- Creating and Managing Clusters with Spark Configurations
- Terminating and Stopping Clusters

- Administering Clusters with Reusable Configurations

Pools

- Pools overview
- Display pools
- Create a pool
- Configure a pool
- Edit a pool
- Delete a pool
- Use a pool

Day 4

Databricks Jobs and Clusters

- Introduction to Jobs and Cluster
- Create Cluster on AWS Databricks
- Request to increase CPU Quota on Azure
- Creating Job on Databricks using Notebook
- Submitting Jobs using Job Cluster
- Create Pool in Databricks
- Running Job using Interactive Cluster attached to Pool
- Running Job using Job Cluster attached to Pool
- Exercise - Submit the application as job using interactive cluster

Notebooks

- Notebooks overview
- Manage notebooks
- Use notebooks

Dashboards - Overview

Notebook workflows

Package cells

Jobs

Day 5:

- Databases and tables
- Datasources
- Delta Lake
- UDF
- Meta Data Server
- SQL databases using JDBC
- AWS SQL Data Warehou**se**