DYNAMICDNA
We develop and empower

aws partner network
Training Partner

Building Batch data Analytics Solutions on AWS

Duration	Delivery Method	Level
1 day	Online / Instructor Led	Intermediate

Introduction:

In this course, you will learn to build batch data analytics solutions using Amazon EMR, an enterprise-grade Apache Spark and Apache Hadoop managed service. You will learn how Amazon EMR integrates with open-source projects such as Apache Hive, Hue, and HBase, and with AWS services such as AWS Glue and AWS Lake Formation.

The course addresses data collection, ingestion, cataloguing, storage, and processing components in the context of Spark and Hadoop. You will learn to use EMR Notebooks to support both analytics and machine learning workloads. You will also learn to apply security, performance, and cost management best practices to the operation of Amazon EMR.

Audience Profile

This course is intended for data warehouse engineers, data platform engineers, and architects and operators who build and manage data analytics pipelines.

Pre-requisite

- Completed either AWS Technical Essentials or Architecting on AWS
- Completed either Building Data Lakes on AWS or Getting Started with AWS Glue

Learning Objectives and Outcomes:

- Compare the features and benefits of data warehouses, data lakes, and modern data architectures
- Design and implement a data warehouse analytics solution
- Identify and apply appropriate techniques, including compression, to optimise data storage
- Select and deploy appropriate options to ingest, transform, and store data
- Choose the appropriate instance and node types, clusters, auto scaling, and network topology for a particular business use case
- Understand how data storage and processing affect the analysis and visualisation mechanisms needed to gain actionable business insights
- Secure data at rest and in transit remediate problems
- Apply cost management best practices

Course outline

DYNAMICDNA
We develop and empower

Module 1: Introduction to Amazon EMR

- Using Amazon EMR in analytics solutions
- Amazon EMR cluster architecture
- Interactive Demo 1: Launching an Amazon EMR cluster
- Cost management strategies

Module 2: Data Analytics Pipeline using Amazon EMR

- Storage optimisation with Amazon EMR
- Data ingestion techniques

Module 3: High performance batch data analytics using apache spack in Amazon EMR

- Apache Spark on Amazon EMR use cases
- Why Apache Spark on Amazon EMR
- Spark concepts
- Interactive Demo 2: Connect to an EMR cluster and perform Scala commands using the Spark shell
- Transformation, processing, and analytics
- Using notebooks with Amazon EMR
- Practice Lab 1: Low-latency data analytics using Apache Spark on Amazon EMR

Module 4: Processing and analysing batch data with amazon EMR and apache hive

- Apache Spark on Amazon EMR use cases
- Why Apache Spark on Amazon EMR
- Spark concepts
- Interactive Demo 2: Connect to an EMR cluster and perform Scala commands using the Spark shell
- Transformation, processing, and analytics
- Using notebooks with Amazon EMR
- Practice Lab 1: Low-latency data analytics using Apache Spark on Amazon EMR

Module 5: Severless data Processing

- Serverless data processing, transformation, and analytics
- Using AWS Glue with Amazon EMR workloads
- Practice Lab 3: Orchestrate data processing in Spark using AWS Step Functions

Module 6: Security and Monitoring of amazon EMR Cluster

- Securing EMR clusters
- Interactive Demo 3: Client-side encryption with EMRFS
- Monitoring and troubleshooting Amazon EMR clusters
- Demo: Reviewing Apache Spark cluster history