Building Batch Data Analytics Solutions on AWS

Course outline

Module A: Overview of Data Analytics and the Data Pipeline

• Data analytics use cases

• Using the data pipeline for analytics

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: Ingestion and Storage

• Storage optimization with Amazon EMR

• Data ingestion techniques

Module 3: High-Performance Batch Data Analytics Using Apache Spark on 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 Analyzing Batch Data with Amazon EMR and Apache Hive

• Using Amazon EMR with Hive to process batch data

• Transformation, processing, and analytics

• Practice Lab 2: Batch data processing using Amazon EMR with Hive

• Introduction to Apache HBase on Amazon EMR

Module 5: Serverless 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 Clusters

• Securing EMR clusters

• Interactive Demo 3: Client-side encryption with EMRFS

• Monitoring and troubleshooting Amazon EMR clusters

• Demo: Reviewing Apache Spark cluster history

• Batch data analytics use cases

• Activity: Designing a batch data analytics workflow

Module B: Developing Modern Data Architectures on AWS

• Modern data architectures