AllCloud

Building Batch Data Analytics Solutions on AWS

Course Benefits & Agenda

Overview

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, cataloging, 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.

Course Benefits

This course teaches you how to:

- Compare the features and benefits of data warehouses, data lakes, and modern data architectures
- Design and implement a batch data analytics solution
- Identify and apply appropriate techniques, including compression, to optimize 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 visualization mechanisms needed to gain actionable business insights
- Secure data at rest and in transit
- Monitor analytics workloads to identify and remediate problems
- Apply cost management best practices

AllCloud

Agenda

Day 1

Module	Topic
Module A	Overview of Data Analytics and the Data Pipeline
Module 1	Introduction to Amazon EMR
Module 2	Data Analytics Pipeline Using Amazon EMR: Ingestion and Storage
Module 3	High-Performance Batch Data Analytics Using Apache Spark on Amazon EMR
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
Lab 2	Batch data processing using Amazon EMR with Hive
Module 5	Serverless Data Processing
Lab 3	Orchestrate data processing in Spark using AWS Step Functions
Module 6	Security and Monitoring of Amazon EMR Clusters
Module 7	Designing Batch Data Analytics Solutions
Module B	Developing Modern Data Architectures on AWS
Module C	Course Wrap-Up