

Building Batch Data Analytics Solutions on AWS – Intensive Training («AWSB05»)

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.

Duration: 1 day

Price: 900.–

Course documents: Digital original AWS courseware

Content

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.

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

Module 7: Designing Batch Data Analytics Solutions

- Batch data analytics use cases
- Activity: Designing a batch data analytics workflow

Module B: Developing Modern Data Architectures on AWS

- Modern data architectures

Key Learnings

- Comparing the features and benefits of data warehouses, data lakes, and modern data architectures
- Designing and implementing a batch data analytics solution
- Identifying and applying appropriate techniques, including compression, to optimize data storage
- Selecting and deploying appropriate options to ingest, transform, and store data
- Choosing the appropriate instance and node types, clusters, auto scaling, and network topology for a particular business use case
- Understanding how data storage and processing affect the analysis and visualization mechanisms needed to gain actionable business insights
- Securing data at rest and in transit
- Monitoring analytics workloads to identify and remediate problems
- Applying cost management best practices

Methodology & didactics

This course includes presentations, interactive demos, practice labs, discussions, and class exercises.

Target audience

This course is intended for the following job roles:

- Data Engineer

Requirements

We recommend that attendees of this course have the following prerequisites:

- a minimum one-year experience managing open-source data frameworks such as Apache Spark or Apache Hadoop
- the [AWS Hadoop Fundamentals](#) course for those that need a refresher on Apache Hadoop

and have attended the following course (or have equivalent knowledge):

- [Architecting on AWS – Intensive Training \(«AWSA01»\)](#)
- [AWS Technical Essentials – Intensive Training \(«AWSE01»\)](#)
- [Building Data Lakes on AWS – Intensive Training \(«AWSB04»\)](#)

Further courses

- [Building Streaming Data Analytics Solutions on AWS – Intensive Training \(«AWSB08»\)](#)

Any questions?

We are happy to advise you on +41 44 447 21 21 or info@digicomp.ch. You can find detailed information about dates on www.digicomp.ch/courses-digital-transformation-technologies/cloud/amazon-web-services-aws/aws-data-engineer/course-building-batch-data-analytics-solutions-on-aws-intensive-training