

**Building Batch Data Analytics Solutions on AWS**

**Course Number:** AWS-154  
**Duration:** 1 day

**Overview**

This Building Batch Data Analytics Solutions on AWS training teaches attendees how to construct batch data analytics solutions using Amazon EMR, a cluster framework that simplifies running big data frameworks like Apache Spark and Hadoop. Participants also learn how Amazon EMR integrates with open-source projects such as Hive, Hue, and HBase, as well as with other AWS services such as AWS Glue and AWS Lake Formation.

Accelebrate is an AWS Training Partner (ATP) and this hands-on official AWS Classroom Training course is taught by an accredited Amazon Authorized Instructor (AAI).

**Prerequisites**

Students should have a minimum of one year of experience managing open-source data frameworks such as Apache Spark or Apache Hadoop. It is recommended that attendees complete [AWS Technical Essentials](file:////training/aws-technical-essentials) or [Architecting on AWS](file:////training/aws-architecture). It is also recommended that students complete [Building Data Lakes on AWS](file:////training/building-data-lakes-on-aws).

**Materials**

All AWS training students receive comprehensive courseware.

**Software Needed on Each Student PC**

A modern web browser and an Internet connection that allows connections by SSH or Remote Desktop (RDP) into AWS virtual machines.

**Objectives**

* 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

**Outline**

* Overview of Data Analytics and the Data Pipeline
  + Data analytics use cases
  + Using the data pipeline for analytics
* Introduction to Amazon EMR
  + Using Amazon EMR in analytics solutions
  + Amazon EMR cluster architecture
  + Launching an Amazon EMR cluster
  + Cost management strategies
* Data Analytics Pipeline Using Amazon EMR: Ingestion and Storage
  + Storage optimization with Amazon EMR
  + Data ingestion techniques
* 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
  + Connect to an EMR cluster and perform Scala commands using the
  + Spark shell
  + Transformation, processing, and analytics
  + Using notebooks with Amazon EMR
  + Low-latency data analytics using Apache Spark on Amazon EMR
* Processing and Analyzing Batch Data with Amazon EMR and Apache Hive
  + Using Amazon EMR with Hive to process batch data
  + Transformation, processing, and analytics
  + Batch data processing using Amazon EMR with Hive
  + Introduction to Apache HBase on Amazon EMR
* Serverless Data Processing
  + Serverless data processing, transformation, and analytics
  + Using AWS Glue with Amazon EMR workloads
  + Orchestrate data processing in Spark using AWS Step Functions
* Security and Monitoring of Amazon EMR Clusters
  + Securing EMR clusters
  + Client-side encryption with EMRFS
  + Monitoring and troubleshooting Amazon EMR clusters
  + Reviewing Apache Spark cluster history
* Designing Batch Data Analytics Solutions
  + Batch data analytics use cases
  + Designing a batch data analytics workflow
* Developing Modern Data Architectures on AWS
  + Modern data architectures