Building Batch Data Analytics Solutions on AWS

Amazon

- **Nível:**
- **Duração:** 7h

Sobre o curso

Do you need to operate, and scale your big data environments by automating time-consuming tasks?
The *Building Batch Data Analytics Solutions on AWS* course will help you learn how to build batch data analytics solutions using Amazon EMR to optimize cost and performance.

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.

Destinatários

This course is intended for:

- Data platform engineers
- Architects and operators who build and manage data analytics pipelines

---

Objetivos

In this course, you will learn 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 aparticular 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

Condições

Após a formação, é possível adquirir, o exame de certificação do parceiro oficial com 10% de desconto. Oferta válida até 6 meses após a conclusão do curso.

Pré-requisitos

- Students with a minimum one-year experience managing open-source data frameworks such as Apache Spark or Apache Hadoop will benefit from this course
- We recommend that attendees of this course have:
    - Completed either AWS Technical Essentials or Architecting on AWS
    - Completed Building Data Lakes on AWS

Programa

- Overview of Data Analytics and the Data Pipeline
- Introduction to Amazon EMR
- Data Analytics Pipeline Using Amazon EMR: Ingestion and Storage
- High-Performance Batch Data Analytics Using Apache Spark on Amazon EMR
- Processing and Analyzing Batch Data with Amazon EMR and Apache Hive
- Serverless Data Processing
- Security and Monitoring of Amazon EMR Clusters
- Designing Batch Data Analytics Solutions
- Developing Modern Data Architectures on AWS

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
- Interactive Demo 1: 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
- 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

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

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

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

Designing Batch Data Analytics Solutions

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

Developing Modern Data Architectures on AWS

- Modern data architectures