Training: AWS
Building Batch Data Analytics Solutions on AWS

TRAINING GOALS:

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 objectives

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 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

Intended audience

This course is intended for:

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

CONSPECT:

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- 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

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70

BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Module B: Developing Modern Data Architectures on AWS
  - Modern data architectures

REQUIREMENTS:

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 suggest the AWS Hadoop Fundamentals course for those that need a refresher on Apache Hadoop.

We recommend that attendees of this course have:

- Completed either AWS Technical Essentials or Architecting on AWS
- Completed either Building Data Lakes on AWS or Getting Started with AWS Glue

Difficulty level



CERTIFICATE:

The participants will obtain certificates signed by AWS (course completion).

This course also helps you prepare for the AWS Certified Data Engineer - Associate DEA-C01 exam and this way gain the AWS Certified Data Engineer - Associate title – associate level. AWS certification exams are offered at Pearson Vue test centers worldwide
https://home.pearsonvue.com/Clients/AWS.aspx

TRAINER:

AWS Authorized Instructor (AAI)

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007