

Building Batch Data Analytics Solutions on AWS



Odbiorcy szkolenia

Ten kurs jest przeznaczony dla:

- Inżynierów platform danych
- Architektów i operatorów, którzy projektują i zarządzają rurociągami analityki danych.



Korzyści

W tym kursie nauczysz się:

- Porównywać cechy i korzyści z magazynów danych, jezior danych oraz nowoczesnych architektur danych
- Projektować i wdrażać rozwiązania analityki danych w trybie batch
- Identyfikować i stosować odpowiednie techniki, w tym kompresję, w celu optymalizacji przechowywania danych
- Wybierać i wdrażać odpowiednie opcje do pozyskiwania, przekształcania i przechowywania danych
- Wybierać odpowiednie instancje, typy węzłów, klastry, automatyczne skalowanie oraz topologię sieci w zależności od konkretnego przypadku biznesowego
- Rozumieć, jak przechowywanie i przetwarzanie danych wpływają na mechanizmy analizy i wizualizacji niezbędne do uzyskania użytecznych informacji biznesowych
- Zabezpieczać dane w spoczynku oraz w trakcie przesyłania
- Monitorować obciążenia analityczne w celu identyfikacji i usuwania problemów
- Stosować najlepsze praktyki zarządzania kosztami.



Program szkolenia

Moduł A: Przegląd analityki danych i pipeline'u danych

- Przykłady zastosowań analityki danych
- Wykorzystanie pipeline'u danych w analizach

Moduł 1: Wprowadzenie do Amazon EMR

- Wykorzystanie Amazon EMR w rozwiązaniach analitycznych
- Architektura klastra Amazon EMR
- Interaktywna demonstracja 1: Uruchamianie klastra Amazon EMR
- Strategie zarządzania kosztami

Moduł 2: Pipeline analityki danych z użyciem Amazon EMR: Ingestia i przechowywanie

- Optymalizacja przechowywania danych z użyciem Amazon EMR
- Techniki ingestii danych

Moduł 3: Wydajna analityka wsadowa z użyciem Apache Spark na Amazon EMR

- Przykłady zastosowań Apache Spark na Amazon EMR
- Dlaczego Apache Spark na Amazon EMR
- Podstawowe pojęcia Spark
- Interaktywna demonstracja 2: Łączenie z klastrzem EMR i wykonywanie poleceń Scala za pomocą powłoki Spark
- Transformacja, przetwarzanie i analityka
- Korzystanie z notatników z Amazon EMR
- Praktyczne laboratorium 1: Analityka danych o niskiej latencji z użyciem Apache Spark na Amazon EMR

Moduł 4: Przetwarzanie i analiza danych wsadowych z użyciem Amazon EMR i Apache Hive

- Wykorzystanie Amazon EMR z Apache Hive do przetwarzania danych wsadowych
- Transformacja, przetwarzanie i analityka
- Praktyczne laboratorium 2: Przetwarzanie danych wsadowych z użyciem Amazon EMR i Apache Hive
- Wprowadzenie do Apache HBase na Amazon EMR

Moduł 5: Przetwarzanie danych w trybie serverless

- Serverless przetwarzanie danych, transformacja i analityka
- Wykorzystanie AWS Glue z obciążeniami Amazon EMR
- Praktyczne laboratorium 3: Orkiestracja przetwarzania danych w Spark z użyciem AWS Step Functions

Moduł 6: Bezpieczeństwo i monitoring klastrów Amazon EMR

- Zabezpieczanie klastrów EMR
- Interaktywna demonstracja 3: Szyfrowanie po stronie klienta z EMRFS
- Monitoring i rozwiązywanie problemów z klastrami Amazon EMR
- Demonstracja: Przeglądanie historii klastra Apache Spark

Moduł 7: Projektowanie rozwiązań analityki danych wsadowych

- Przykłady zastosowań analityki danych wsadowych
- Aktywność: Projektowanie workflow analityki danych wsadowych

Moduł B: Rozwój nowoczesnych architektur danych na AWS

- Nowoczesne architektury danych



Oczekiwane przygotowanie uczestnika

Studenci z co najmniej rocznym doświadczeniem w zarządzaniu frameworkami open-source do przetwarzania danych, takimi jak Apache Spark czy Apache Hadoop, będą czerpać największe korzyści z tego kursu.



Szkolenie obejmuje

Kurs obejmuje prezentacje, interaktywne pokazy, laboratoria praktyczne, dyskusje oraz ćwiczenia klasowe.



Czas trwania

1 dni / 7 godzin

Język

- Szkolenie: polski
- Materiały: angielski