

Building Batch Data Analytics Solutions on AWS

AWS 課堂培訓

課程說明

在本課程中，您將學習使用企業等級 Apache Spark 和 Apache Hadoop 受管服務 Amazon EMR 建置批次資料分析解決方案。您將學習 Amazon EMR 如何與 Apache Hive、Hue 和 HBase 等開放原始碼專案整合，以及如何與 AWS Glue 和 AWS Lake Formation 等 AWS 服務整合。本課程討論 Spark 和 Hadoop 環境下的資料收集、擷取、編目、儲存和處理元件。您將學習使用 EMR Notebooks 支援分析和機器學習工作負載。您還將學習如何在 Amazon EMR 操作中運用安全、效能和成本管理最佳實務。

- 課程等級：中級
- 授課時長：1 天

活動

本課程包含簡報、互動式示範、練習實驗室、討論和課堂練習。

課程目標

在此課程中，您將學習如何：

- 比較資料倉儲、資料湖和現代資料架構的功能和優勢
- 設計和實作批次資料分析解決方案
- 識別和運用適當的技術 (包括壓縮)，將資料儲存最佳化
- 選取和部署適當選項以擷取、轉換和儲存資料
- 為特定商業使用案例選擇適當的執行個體和節點類型、叢集、自動擴展和網際網路拓撲
- 了解資料儲存和處理如何影響必要的分析和視覺化機制，藉此取得可執行的業務洞察
- 保護靜態和傳輸中資料的安全
- 監控分析工作負載以識別並補救問題
- 運用成本管理最佳實務

目標對象

本課程適合：

- 資料平台工程師

Building Batch Data Analytics Solutions on AWS

AWS 課堂培訓

- 建置和管理資料分析管道的架構師和操作員

先決條件

至少具有一年管理開放原始碼資料架構 (例如 Apache Spark 或 Apache Hadoop) 的學員將從本課程中獲益。

我們建議需要複習 Apache Hadoop 的學員進修 [AWS Hadoop Fundamentals](#) 課程。

我們建議參加本課程的學員具備：

- 完成下列課程其中之一：[AWS Technical Essentials](#) 或 [Architecting on AWS](#)
- 完成下列課程其中之一：[Building Data Lakes on AWS](#) 或 [Getting Started with AWS Glue](#)

Building Batch Data Analytics Solutions on AWS

AWS 課堂培訓

課程大綱

單元 A：資料分析和 Data Pipeline 概觀

- 資料分析使用案例
- 使用 Data Pipeline 進行分析

單元 1：Amazon EMR 簡介

- 在分析解決方案中使用 Amazon EMR
- Amazon EMR 叢集架構
- 互動式示範 1：啟動 Amazon EMR 叢集
- 成本管理策略

單元 2：使用 Amazon EMR 的資料分析管道：擷取和儲存

- 使用 Amazon EMR 達成儲存最佳化
- 資料擷取技術

單元 3：使用 Amazon EMR 上的 Apache Spark 達成高效能批次資料分析

- Amazon EMR 上的 Apache Spark 使用案例
- 為何採用 Amazon EMR 上的 Apache Spark
- Spark 觀念
- 互動式示範 2：使用 Spark shell 連線到 EMR 叢集並執行 Scala 命令
- 轉換、處理和分析
- 搭配 Amazon EMR 使用筆記本
- 練習實驗室 1：使用 Amazon EMR 上的 Apache Spark 達成低延遲的資料分析

單元 4：使用 Amazon EMR 和 Apache Hive 處理和分析批次資料

- 使用安裝了 Hive 的 Amazon EMR 處理批次資料
- 轉換、處理和分析
- 練習實驗室 2：使用安裝了 Hive 的 Amazon EMR 進行批次資料處理
- Apache HBase on Amazon EMR 簡介

單元 5：無伺服器資料處理

- 無伺服器資料處理、轉換和分析
- 使用 AWS Glue 搭配 Amazon EMR 工作負載

Building Batch Data Analytics Solutions on AWS

AWS 課堂培訓

- 練習實驗室 3：使用 AWS Step Functions 協調 Spark 中的資料處理

單元 6：Amazon EMR 叢集的安全和監控

- 保護 EMR 叢集
- 互動式示範 3：使用 EMRFS 進行用戶端加密
- Amazon EMR 叢集的監控和疑難排解
- 示範：檢閱 Apache Spark 叢集歷史記錄

單元 7：設計批次資料分析解決方案

- 批次資料分析使用案例
- 活動：設計批次資料分析工作流程

單元 B：在 AWS 上開發現代化資料架構

- 現代化資料架構