

## ***AW272 Building Batch Data Analytics Solutions on AWS***

### **Kurzbeschreibung:**

In dem Kurs „Building Batch Data Analytics Solutions on AWS“ erhalten die Teilnehmer einen Einblick in die Verwendung von Amazon EMR in Verbindung mit Open Source Projekten und lernen verschiedene Komponenten kennen.

### **Zielgruppe:**

- Data Engineers
- Architekten und Operators, die Datenanalyse-Pipelines aufbauen und verwalten

### **Voraussetzungen:**

Um an dem Kurs „Building Batch Data Analytics Solutions on AWS“ bei qSkills teilnehmen zu können, sollten Sie folgende AWS Trainings besucht haben:

- "AWS Technical Essentials"  
oder
- "Architecting on AWS"
- "Building Data Lakes on AWS"

Zusätzlich sollten die Teilnehmer folgende Voraussetzungen erfüllen:

Mindestens 1 Jahr Erfahrung im Umgang mit Open-Source-Daten-Frameworks wie Apache Spark oder Apache Hadoop.

### **Sonstiges:**

**Dauer:** 1 Tage

**Preis:** 750 Euro plus Mwst.

### **Ziele:**

- Data Warehouse, Data Lakes und moderne Datenarchitekturen vergleichen
- Batch-Datenanalyiselösung entwerfen und implementieren
- Geeignete Techniken für Komprimierung und Optimierung der Datenspeicherung einsetzen
- Passende Instanz- und Knotenpunkte, Cluster, automatische Skalierung und Netzwerktopologie für einen Anwendungsfall
- Auswirkungen der Datenspeicherung und -verarbeitung auf Analyse- und Visualisierungsmechanismen
- Sichern der Daten im Ruhezustand und während der Übertragung
- Mit Hilfe von Analyse-Workloads Probleme erkennen und beheben
- Best Practices für das Kostenmanagement

## Inhalte/Agenda:

- ♦ Das eintägige Training „Building Batch Data Analytics Solutions on AWS“ vermittelt die Erstellung von Batch-DatenanalySELösungen unter Verwendung von Amazon EMR. Dabei handelt es sich um einen verwalteten Apache Spark und Apache Hadoop-Service der Enterprise-Klasse.

Die Teilnehmer lernen Amazon EMR in Open Source Projekte wie Hive, Hue oder HBase zu integrieren und mit AWS-Services wie AWS Glue und AWS Lake Formation zu kombinieren. Darüber hinaus werden verschiedene Komponenten der Datensammlung, -aufnahme, -katalogisierung, -speicherung und -verarbeitung in Verbindung mit Spark und Hadoop vorgestellt. Der Einsatz von EMR Notebooks für Analyse und Machine Learning sowie Best-Practices zu den Themen Sicherheit, Leistung und Kostenmanagement runden den Kurs ab.

Dieser Kurs setzt sich aus einer Präsentation, Übungen und interaktiven Demos zusammen, um das Erlernte praktisch anzuwenden.

Die Kursunterlagen (E-Book) sind in englischer Sprache, die Kurssprache ist deutsch.

♦

### ♦ **Modul A: Überblick über Datenanalyse und die Datenpipeline**

#### ♦ Modul 1: Einführung in Amazon EMR

- ♦ ♦ Verwendung von Amazon EMR in Analyselösungen
- ♦ ♦ Architektur von Amazon EMR-Clustern
- ♦ ♦ Interaktive Demo 1: Starten eines Amazon EMR-Clusters
- ♦ ♦ Strategien zur Kostenverwaltung

#### ♦ Modul 2: Datenanalyse-Pipeline mit Amazon EMR: Ingestion und Speicherung

- ♦ ♦ Speicheroptimierung mit Amazon EMR
- ♦ ♦ Techniken zur Datenübernahme

#### ♦ Modul 3: Leistungsstarke Batch-Datenanalyse mit Apache Spark auf Amazon EMR

- ♦ ♦ Anwendungsfälle für Apache Spark auf Amazon EMR
- ♦ ♦ Warum Apache Spark auf Amazon EMR
- ♦ ♦ Spark-Konzepte
- ♦ ♦ Interaktive Demo 2: Interaktive Analytik mit Apache Spark auf Amazon EMR
- ♦ ♦ Transformation, Verarbeitung und Analyse
- ♦ ♦ Verwendung von Notebooks mit Amazon EMR
- ♦ ♦ Übungslabor 1: Datenanalyse mit niedriger Latenz mit Apache Spark auf Amazon EMR

#### ♦ Modul 4: Verarbeiten und Analysieren von Batch-Daten mit Amazon EMR und Hive

- ♦ ♦ Verwendung von Amazon EMR mit Hive zur Verarbeitung von Stapeldaten
- ♦ ♦ Transformation, Verarbeitung und Analyse
- ♦ ♦ Practice Lab 2: Batch-Datenverarbeitung mit Amazon EMR und Hive
- ♦ ♦ Einführung in HBase auf Amazon EMR

#### ♦ Modul 5: Serverlose Datenverarbeitung

- ♦ ♦ Serverlose Datenverarbeitung, -transformation und -analytik
- ♦ ♦ Verwendung von AWS Glue mit Amazon EMR-Arbeitslasten
- ♦ ♦ Praxisübung 3: Orchestrierung der Datenverarbeitung in Spark mit AWS Step Functions

#### ♦ Modul 6: Sicherheit und Überwachung von Amazon EMR-Clustern

- ♦ ♦ Sichern von EMR-Clustern
- ♦ ♦ Interaktive Demo 3: Verschlüsselung von Daten im Ruhezustand in Amazon EMR
- ♦ ♦ Überwachung und Fehlersuche bei EMR-Clustern
- ♦ ♦ Demo: Überprüfung des Verlaufs von Apache Spark-Clustern
- ♦ ♦ Überwachung und Fehlerbehebung von Amazon EMR-Clustern

#### ♦ Modul 7: Entwerfen von Batch-DatenanalySELösungen

- ♦ ♦

- ◇ Anwendungsfälle der Batch-Datenanalyse
- ◇ Aktivität: Entwerfen eines Arbeitsablaufs für die Batch-Datenanalyse

- ◆ **Modul B: Entwickeln moderner Datenarchitekturen auf AWS**

- ◆ ◇ Moderne Datenarchitekturen

