

Trainings

Develop Apache Spark Applications (Spark v 2.1)



Zielgruppe: Software-Entwickler, -Architekten

Dauer: 3 Tage

Abstract

Dieses Training vermittelt alle Kenntnisse, die notwendig sind, um Big-Data-Applikationen auf Basis von Apache Spark 2.1 zu entwickeln. Die Teilnehmer lernen zunächst den Umgang mit der Spark-Shell, um Datensätze aus verschiedenen Quellen und Formaten zu laden und diese interaktiv zu analysieren. Darauf aufbauend entwickeln die Teilnehmer eine eigenständige Spark-Anwendung, um Daten in Form von Datasets und DataFrames lokal oder auf einem Rechencluster zu verarbeiten.. Abgeschlossen wird das Training durch eine Einführung in Spark-Streaming zur Verarbeitung von Datenströmen, GraphFrame zur Analyse von Graphen sowie der Machine-Learning-Bibliothek MLlib.

Notwendige Vorkenntnisse:

- Grundlegendes Hadoop-Wissen
- Grundlegende Linux-Kenntnisse (einschließlich Befehlszeilenoptionen wie ls, cd, cp und su)
- gute Java oder Scala-Kenntnisse
- gute SQL-Kenntnisse

Dieses von MapR offiziell akkreditierte Training befähigt die Teilnehmer zur Teilnahme an der Prüfung zum MapR Certified Spark Developer (MCSD).

Agenda:

1. Apache Spark Grundlagen (DEV 360)

- Apache Spark Features
- Komponenten des Spark Frameworks
- Use Cases

2. Datasets erstellen

- Datenquellen, -strukturen und -schemata definieren
- Arbeiten mit Datasets und DataFrame
- DataFrames in Datasets umwandeln

Praxis:

- Daten laden und Datasets mittels Reflection erstellen
- Einfacher Usecase: Wortzählung mit Datasets (Optional)

3. Operationen für Datasets

- grundlegende Operationen auf Datasets
- Caching von Datasets
- Benutzerdefinierter Funktionen (UDFs)
- Partitioning von Datasets

Praxis:

- Analyse von SFPD-Daten
- UDFs erstellen und anwenden
- Daten mit Hilfe von UDF und Queries analysieren

4. Eine einfache Apache-Spark-Anwendung entwickeln (DEV 361)

- Spark Application Lifecycle
- Verwendung von SparkSession
- Starten von Spark-Anwendungen

Praxis:

- Anwendungsdateien importieren und konfigurieren
- Anwendungen bauen, deployen und starten

5. Apache-Spark-Anwendungen monitoren

- Logische und physische Ablaufpläne von Spark
- Spark Web UI zum Monitoring von Spark-Anwendungen
- Spark Anwendungen debuggen und tunen

Praxis:

- Spark UI nutzen
- Spark-Systemeigenschaften interpretieren

6. Apache Spark Streaming Anwendung erstellen (DEV 362)

- Einführung in die Spark Streaming-Architektur
- Entwicklung von Spark Structured-Streaming-Anwendungen
- Operationen auf Streaming DataFrames anwenden
- eigene Windows-Anwendungen entwickeln

Praxis:

- Mit der Spark Shell Daten laden und analysieren
- Spark Streaming in der Spark Shell
- Streaming-Anwendung mit SQL bauen und ausführen
- Streaming-Anwendung mit Windows-Funktionen und SQL bauen und ausführen

7. Apache Spark GraphFrames verwenden

- Einführung in GraphFrame
- Reguläre, gerichtete und Property Graphen definieren
- Property Graphen erstellen
- Durchführung von Operationen auf Graphen

Praxis:

- Graphanalyse mit GraphFrames

8. Apache Spark MLlib verwenden

- Einführung in Apache Spark MLlib (Machine Learning Library)
- Collaborative Filtering zur Generierung von Produktempfehlungen

Praxis:

- Datenanalyse mit der Spark Shell
- Entwicklung einer Spark-Applikation für Film-Empfehlungen
- Analyse eines einfachen Flug-Systems mit Entscheidungsbäumen

Hinweis:

- Die Kursgebühr beinhaltet Schulungsunterlagen, Mittagessen, Getränke und Snacks.
- Die Teilnehmer müssen ein eigenes Notebook zum Training mitbringen.

Sollten Sie Fragen zum Training haben, eine Hotelempfehlung benötigen oder ein anderes Anliegen rund um unsere Trainings haben, kontaktieren Sie uns bitte unter trainings@inovex.de oder unseren Academy-Leiter Collin Rogowski unter Tel. +49 (0)172 5673497. Vielen Dank! Wir freuen uns auf Sie!