Big Data Training: Data Lakes – Hands On

Was hat ein Data Lake mit dem Turm im Reschensee zu tun?

- **Das Datenvolumen wächst unaufhaltsam, so dass wesentliche Strukturen verschwinden, neue müssen geschaffen werden.**
- **Neue Hilfsmittel sind notwendig, um Daten zu analysieren und strukturieren.**
- **Tools und Frameworks sind bereits erfolgreich im Einsatz.**

Was sollte einen Date Lake vom Reschensee unterscheiden?

- **Mit der passenden Organisation bleibt der Überblick erhalten.**
- **Sorgfältig ausgewählte Werkzeuge ermöglichen die schlagkräftige Analyse der Datenflut.**

Viele Open Source Tools versprechen ihre Dienste zur Big Data Analyse. Die Tools sind rasch heruntergeladen und auf einem Laptop installiert. Die APIs (Applications Programming Interfaces) zur Datenanalyse sehen einfach aus, oftmals wird das bekannte SQL eingesetzt, das erste Erfolgserlebnis stellt sich bald ein. Und damit kommen viele Fragen:

- Welche Tools und Frameworks zur Organisation eines Data Lakes gibt es?
- Wodurch unterscheiden sich die einzelnen Frameworks?
- Welche Vor- und Nachteile haben sie?
- Wie werden die Tools in die bestehende Systemlandschaft integriert?
- Welche Fragen sollten in einer Evaluation eines Data Lakes gestellt werden?
- Wie sieht der passende PoC (Proof of Concept) für meinen Anwendungsfall aus?

Dieses Training setzt bei diesen Fragen ein.

Die Teilnehmenden lernen die Grundlagen der verteilten Datenspeicherung- und Datenanalyse kennen und sind nach dem Kurs in der Lage, die Evaluation der Tools für ihren Anwendungsfall zu planen und einen Proof of Concept zielgerichtet durchzuführen.

Zielpublikum

Informatikerinnen und Informatiker, die

… sich für Big Data Analytics in Data Lakes interessieren und hinter die Kulissen der APIs blicken möchten;

… die ein Grundverständnis für die Komplexität der verteilten Big-Data-Systeme aufbauen möchten;

… grundlegende Organisationsstrukturen für Data Lakes kennen möchten;

… die anhand eines soliden Verständnisses von Konzepten in der Lage sein wollen, selbständig Details zu vertiefen.

Lernziele

Die Teilnehmenden ...

… können erläutern, warum verteilte Systeme für das ausfallfreie Rechnen mit großen Datenbeständen eingesetzt werden;
… kennen die Trade-Offs, die verteilte Systeme mit sich bringen;
… kennen die markantesten Architekturen verteilter Systeme;
… können die wichtigsten Open-Source-Frameworks für die Organisation von Data Lakes benennen und in Kategorien einteilen;
… kennen Herangehensweisen, wie einzelne Frameworks zur Big-Data-Analytics eingesetzt werden.
… können das Vorgehen zur Evaluation eines Data Lakes Ihren spezifischen Use-Case planen und kennen die wichtigsten Fragen, die ihr PoC klären sollte.

Notwendige Vorkenntnisse

- Erfahrene Informatikerinnen und Informatiker, insbesondere Programmiererinnen und Programmierer, Systemverantwortliche, Enterprise Architektinnen und Architekten.
- Grundkenntnisse Linux sind von Vorteil.
- Hands-On Übungen werden mit den Python-APIs und mit SQL durchgeführt.

Dozentin

Ursula Deriu führt an verschiedenen Fachhochschulen Vorlesungen zu den Themenkreisen Data Management, Big Data und Data Science durch. Sie ist Autorin des Buchs Stream Analytics Pipelines.

Sprache

- Deutsch
- Englisch auf Anfrage
- Unterlage in deutscher Sprache

Form und Dauer

- Online Training auf einer geschlossenen Cloud-Plattform.
- In-House-Training auf Anfrage.
- Das Training dauert zwei Tage

Buchung und Daten

- Auskunft und Buchung über die Homepage von https://tirsus.com/data-lakes-hands-on/