Python - dátová analýza III (BigData Spark Analysis)

Kód kurzu: PYTHON_DATAN3

Školenie pre analýzu veľkých dát pomocou Apache Spark zahŕňa prehľad základných a pokročilých tém, praktické cvičenia a diskusie na posilnenie znalostí o analýze veľkých dát. Spark je výkonný nástroj na spracovanie veľkých dát, ktorý umožňuje rýchle vykonávanie analýz a podporuje rôzne úlohy, vrátane dávkového spracovania, streamovania, interaktívnych dotazov a strojového učenia.

Pre koho je kurz určený

- Data Scientist, dátoví analytici, najmä v Big Data prostredí sú primárnymi poslucháčmi tohto intenzívneho kurzu
- Softvéroví vývojári, ktorí ovládajú jazyk Python aspoň na strednej až pokročilej úrovni a ktorí majú za cieľ vytvárať data-intenzívne aplikácie pomocou enginu SPARK v prostredí Big Data (Cloud)
- Dátoví architekti

Požadované vstupné znalosti

- Znalosti jazyka Python a dátovej analýzy na úrovni kurzu
- PYTHON_ADV
- a
- PYTHON_DATAN2

Osnova kurzu

Úvod do Apache Spark a jeho ekosystému

- Úvod do veľkých dát a ich význam
- Prehľad ekosystému Apache Spark a jeho porovnanie s inými technológiami veľkých dát
- Inštalácia a konfigurácia Apache Spark a príprava vývojového prostredia
- Základy RDD (Resilient Distributed Dataset) a jeho operácie
- Praktické cvičenie: Vytvorenie prvej Spark aplikácie s využitím RDD
- Diskusia o výhodách a nevýhodách RDD
- Úvod do Datasetov a DataFrames pre efektívnejšiu prácu s dátami

Pokročilé spracovanie dát s Apache Spark

- Podrobný pohľad na DataFrames a operácie s nimi
- SQL dotazy v Spark a práca so Spark SQL
- Praktické cvičenie: Transformácia dát a agregácia pomocou Spark SQL a DataFrames
- Úvod do spracovania prúdových dát s Apache Spark Streaming
- Praktické cvičenie: Jednoduchá prúdová aplikácia

Strojové učenie a pokročilá analýza dát v Spark

- Prehľad MLlib (Machine Learning Library) v Sparku
- Vytváranie a hodnotenie modelov strojového učenia
- Praktické cvičenie: Klasifikácia, regresia a zhlukovanie s MLlib
- Integrácia Sparku s inými úložiskami dát (napr. HDFS, Amazon S3)

Optimalizácia a ladenie výkonu Spark aplikácií

- Monitorovanie a ladenie Spark aplikácií
- Práca so Spark UI pre analýzu výkonu aplikácií
- Optimalizácia výkonu pomocou particionovania a perzistencie
- Praktické tipy a triky pre efektívne spracovanie veľkých dát

Škálovateľnosť a nasadenie Spark aplikácií

- Architektúra Spark clustera a jeho konfigurácia
- Škálovanie Spark aplikácií vertikálne a horizontálne
- Nasadenie Spark aplikácií v produkčnom prostredí
- Najlepšie postupy pre prácu s Apache Spark

GOPAS Praha
Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk

- Záverečná diskusia, odpovede na otázky a spätná väzba od účastníkov

GOPAS Praha
Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk