Python - data analysis III (BigData Spark Analysis)

Course code: PYTHON_DATAN3

Big Data Analytics with Apache Spark training includes an overview of basic and advanced topics, hands-on exercises, and discussions to reinforce your knowledge of Big Data Analytics. Spark is a powerful big data processing engine that enables rapid analysis and supports a variety of tasks, including batch processing, streaming, interactive queries, and machine learning.

For whom the course is intended

- Data Scientist, data analysts, especially in a Big Data environment, are the primary audience for this intensive course.
- Software developers who know the Python language at least at an intermediate and advanced level and who aim to create data-intensive applications using the SPARK engine in a Big Data (Cloud) environment.
- Data architects

Required skills

- Knowledge of Python and data analysis at the PYTHON_ADV and PYTHON_DATAN2 course level

Course outline

Introduction to Apache Spark and the ecosystem

- Introduction to big data and its importance
- An overview of the Apache Spark ecosystem and how it compares to other big data technologies
- Installing and configuring Apache Spark and preparing the development environment
- Basics of RDD (Resilient Distributed Dataset) and its operations
- Practical exercise: Creating the first Spark application using RDD
- Discussion of the advantages and disadvantages of RDD
- Introduction to Datasets and DataFrames for more efficient work with data

Advanced data processing with Apache Spark

- Detailed view of DataFrames and operations with them
- SQL queries in Spark and working with Spark SQL
- Hands-on: Data transformation and aggregation using Spark SQL and *taFrames
- Introduction to stream data processing with Apache Spark Streaming
- Hands-on exercise: A simple streaming application

Machine learning and advanced data analysis in Spark

- Overview of MLlib (Machine Learning Library) in Spark
- Building and evaluating machine learning models
- Hands-on exercise: Classification, regression and clustering with MLlib
- Spark integration with other data stores (eg HDFS, Amazon S3)

Optimizing and tuning the performance of Spark applications

- Monitoring and debugging Spark applications
- Working with Spark UI for application performance analysis
- Performance optimization using partitioning and persistence
- Practical tips and tricks for efficient processing of big data

Scaling and deploying Spark applications

- Spark cluster architecture and its configuration
- Scaling Spark applications vertically and horizontally
- Deploying Spark applications in a production environment
- Best practices for working with Apache Spark
- Final discussion, answers to questions and feedback from participants

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