

# Python - datová analýza III (BigData Spark Analysis)

Kód kurzu: PYTHON\_DATAN3

Školení pro analýzu velkých dat pomocí Apache Spark zahrnuje přehled základních a pokročilých témat, praktická cvičení a diskuse k posílení znalostí o analýze velkých dat. Spark je výkonný nástroj pro zpracování velkých dat, který umožňuje rychlé provádění analýz a podporuje různé úkoly, včetně dávkového zpracování, streamování, interaktivních dotazů a strojového učení.

## Pro koho je kurz určen:

- Data Scientist, datoví analytici, zejména v Big Data prostředí, jsou primárním auditoriem pro tento intenzivní kurz.
- Software vývojáři, kteří ovládají jazyk Python alespoň na střední až pokročilé úrovni a kteří mají za cíl vytvářet data-intenzivní aplikace pomocí enginu SPARK v prostředí Big Data (Cloud).
- Datoví architekti

Požadované vstupní znalosti:

- Znalosti jazyka Python a datové analýzy na úrovni kurzu PYTHON\_ADV a PYTHON\_DATAN2

Metody výuky:

- Odborný výklad s praktickými ukázkami, cvičení na počítačích.

Studijní materiály:

- Prezentace probírané látky v tištěné nebo online formě.

Osnova: Úvod do Apache Spark a ekosystému

- Úvod do velkých dat a jejich význam.
- Přehled ekosystému Apache Spark a jeho porovnání s jinými technologiemi velkých dat.
- Instalace a konfigurace Apache Spark a příprava vývojového prostředí.
- Základy RDD (Resilient Distributed Dataset) a jeho operace.
- Praktické cvičení: Vytvoření prvního Spark aplikace s využitím RDD.
- Diskuse o výhodách a nevýhodách RDD.
- Úvod do Datasetů a DataFrames pro efektivnější práci s daty.

Pokročilé zpracování dat s Apache Spark

- Podrobný pohled na DataFrames a operace s nimi.
- SQL dotazy ve Sparku a práce s Spark SQL.
- Praktické cvičení: Transformace dat a agregace pomocí Spark SQL a DataFrames.
- Úvod do zpracování streamových dat s Apache Spark Streaming.
- Praktické cvičení: Jednoduchá streamová aplikace.

Strojní učení a pokročilá analýza dat ve Sparku

- Přehled MLlib (Machine Learning Library) ve Sparku.
- Budování a evaluace modelů strojního učení.
- Praktické cvičení: Klasifikace, regrese a shlukování s MLlib.
- Integrace Sparku s jinými úložišti dat (např. HDFS, Amazon S3).

Optimalizace a tuning výkonu Spark aplikací

- Monitorování a ladění Spark aplikací.
- Práce s Spark UI pro analýzu výkonu aplikací.
- Optimalizace výkonu pomocí particionování a persistence.
- Praktické tipy a triky pro efektivní zpracování velkých dat.

Škálování a nasazení Spark aplikací

- Architektura Spark clusteru a jeho konfigurace.
- Skalování Spark aplikací vertikální a horizontální.
- Nasazení Spark aplikací v produkčním prostředí.
- Best practices pro práci s Apache Spark.
- Závěrečná diskuse, odpovědi na otázky a zpětná vazba od účastníků.

## GOPAS Praha

Na Strži 2097/63  
140 00 Praha 4 - Krč  
Tel.: +420 226 201 390  
[info@gopas.cz](mailto:info@gopas.cz)

## GOPAS Brno

Nové sady 996/25  
602 00 Brno  
Tel.: +420 530 513 590  
[info@gopas.cz](mailto:info@gopas.cz)

## GOPAS Bratislava

Dr. Vladimíra Clementisa 10  
Bratislava, 821 02  
Tel.: +421 902 903 132  
[info@gopas.sk](mailto:info@gopas.sk)



Copyright © 2026 GOPAS, a.s.,  
All rights reserved