Tytuł szkolenia: Spark dla programistów

Kod szkolenia: SPARK

Wprowadzenie

Szkolenie z analizy i przetwarzania danych w oparciu o Spark.

Adresaci szkolenia

Cel szkolenia

• Wprowadzenie w zagadnienia BigData;
• Poznanie architektury przetwarzania danych w klastrze;
• Zdefiniowanie rozproszonych zbirów danych;
• Poznanie technik pracy z danymi, transformacje i agregacje;
• Poznanie w praktyce języka Spark SQL;
• Nabycie praktyki w wykorzystywaniu RDDs, w przetwarzaniu strumieniowym;
• Wykorzystanie biblioteki MLib do zagadnień uczenia maszynowego.

Czas i forma szkolenia

- 14 godzin (2 dni x 7 godzin), w tym wykłady i warsztaty praktyczne.

Plan szkolenia

1. Spark

- Wprowadzenie do BigData
- Architektura i zastosowania Apache Spark
- Środowisko i uruchamianie aplikacji, cykl życia aplikacji

2. Structured API

- Charakterystyka rozproszonych zbirów danych, Datasets, DataFrames, SQL Tables
- Architektura przetwarzani danych w klastrze
- Praca z danymi, transformacje, typy, schematy, rekordy, kolumny, agregacje, łączenia
- Praca ze źródłami danych, pliki, CSV, JSNO, bazy danych SQL
- Wykorzystanie Spark SQL

3. RDDs

- Charakterystyka i przypadki użycia Low-Level API
- Współpraca z DataFrames i Datasets
- Praca z RDDs, transformacje, akcje, agregacje
- Broadcast variables i współdzielenie

4. Stream Processing

- Charakterystyka przetwarzania strumieniowego w Spark
- Praca ze Streaming API, Structured Streaming
- Przetwarzanie Event-Time i Stateful

5. Machine Learning

- Charakterystyka procesu zaawansowanej analizy w Spark
- Charakterystyka mechanizmu
- Machine Learning
- Praca z biblioteką MLib