SZKOLENIE ŚREDNIO ZAAWANSOWANE

Apache Spark z wykorzystaniem języka Scala

BIGDATA/SPARK

Czas trwania: 3 dni (24h)

Praktyczne wykorzystanie platformy Apache Spark w kontekście przetwarzania Big Data

Cele szkolenia

- Zapoznanie z platformą Spark oraz jej API w języku Scala
- Pokazanie metod analizy dużej ilości danych

Zalety

- Kompleksowe wprowadzenie do platformy Spark
- Przedstawienie praktycznych przykładów oraz praktyk związanych z analizą dużej ilości danych
- Praktyka przed teorią - wszystkie szkolenia technologiczne prowadzone są w formie warsztatowej. Konieczna teoria jest wyjaśniana na przykładzie praktycznych zadań
- Konkretne umiejętności - w ramach każdego szkolenia rozwijamy praktyczne umiejętności związane z daną technologią i tematyką
- Nauka z praktykami - wszyscy trenerzy na co dzień pracują w projektach, gwarantuje to dostęp do eksperckiej wiedzy i praktycznego know-how

Dla kogo?

- Analitycy i programiści, którzy znają podstawy Big Data i chcą rozpocząć przygodę z wykorzystaniem platformy Spark oraz językiem Scala

Wymagania

- Dobra znajomość: języka SQL, relacyjnego modelu danych oraz hurtowni danych
- Podstawowa znajomość obiektowych języków programowania np.: Java, Python lub Scala
- Znajomość zagadnień Big Data, platformy Hadoop oraz powiązanych z nią narzędzi
- Zalecany jest wcześniejszy udział w szkoleniu: Big Data i platforma Hadoop - wprowadzenie (BIGDATA/BASE)

Program

1. Podstawy języka Scala
    a. Zmienne, kontrola statyczna i wnioskowanie typów

b. Instrukcje sterujące
c. Skala jako język obiektowy
-Klasy
-Dziedziczenie
-Singletony
-Klasy przypadków
-Metody klas
-Hierarchia klas
d. Skala jako język funkcyjny
-Cechy funkcji
-Przekazywanie parametrów
-Domyślne wartości parametrów
-Funkcje ze zmienną liczbą parametrów
-Funkcje wyższego rzędu
-Funkcje anonimowe
e. Złożone typy danych
-Tablice
-Krotki
-Kolekcje
f. Zagadnienia uzupełniające
-Pattern matching
-Option
-Closure
-Obsługa ciągów znaków
-Języki domenowe
g. Warsztat

2. Wprowadzenie do Apache Spark
a. Historia
b. Architektura
c. Typy konfiguracji
d. Terminologia - aplikacje, zadania, etapy, jednostki
e. Jak to wszystko działa?
f. Struktura programu
g. Środowiska REPL - spark-shell
h. Dlaczego Scala?
i. Co dalej?
j. Warsztat

3. Przetwarzanie RDD
a. Wprowadzenie do RDD
b. Transformacje
c. Akcje
d. Agregacja i redukcja
e. Warsztat

4. RDD typu klucz-wartość

a. Typy RDD
b. PairRDDFunctions i jego znaczenie
c. Tworzenie RDD par
d. Metody przetwarzające pojedyncze RDD par
e. Łączenie RDD par
f. Warsztat

5. Spark SQL - DataFrames
   a. Wprowadzenie do Spark SQL
   b. DataFrames vs. Dataset
   c. Wczytywanie danych, źródła danych
   d. Schemat danych
   e. Przetwarzanie danych
      -Transformacje (typed vs untyped)
      -Grupowanie
      -Akcje
      -Wykorzystanie SQL
   f. Typy danych - konsekwencje
   g. Warsztat
6. Spark SQL - Dataset
   a. Wydajność Spark SQL
      -Catalyst
      -Plany zapytań
      -Tungsten
   b. Dataset
      -Tworzenie
      -Metody
      -Grupowanie
      -KeyValueGroupedDataset
      -Agregacja
   c. Profilowanie danych
   d. Czyszczenie danych
   e. Podsumowanie: RDD, DataFrames, Dataset
   f. Warsztat - projekt
7. RDD - wydajność
   a. Wprowadzenie - opóźnienia, czas dostępu do danych
   b. Przesyłanie danych - konsekwencje
   c. Partycjonowanie danych
   d. Wąskie i szerokie zależności
      -Wpływ na wydajność
      -Wpływ na obsługę awarii
   e. Zmienne rozgłoszeniowe
   f. Akumulatory
   g. Warsztat
8. Biblioteka Delta Lake

a. Wprowadzenie
b. Zasilanie Delta Lake
c. Odczyt i zapis
d. Obsługa modyfikacji
e. Elementy zaawansowane
    -Narzędzia
    -Wersjonowanie
    -Ograniczenia
    -Kontrola współbieżnego dostępu
f. Warsztat

9. Biblioteka ML
    a. Wprowadzenie
    b. Prosta statystyka
    c. Algorytmy uczenia maszynowego
    d. Spark ML
    e. Przykłady
    f. Warsztat - regresja liniowa, klasyfikator