BO·TT·EGA
IT minds

Program szkolenia:

Przetwarzanie danych Big Data z wykorzystaniem Apache Spark

Informacje:

Nazwa:	**Przetwarzanie danych Big Data z wykorzystaniem Apache Spark**
Kod:	**BigDataML-BigData**
Kategoria:	BigData, streaming i Machine Learning
Odbiorcy:	
Czas trwania:	2 dni
Forma:	40% wykłady 60% ćwiczenia

W trakcie szkolenia nauczymy się jak używać Apache Spark framework do szybkiego przetwarzania dużych ilości danych.

Kurs obejmuje wprowadzenie do architektury Apache Spark. Zobaczymy jak wygląda Spark API i będziemy pisać Spark Joby ilustrując typowe jak i specyficzne problemy.

BOTTEGA IT minds

Szczegółowy program:

1. Wprowadzenie do Apache Spark

- 1.1. Architektura Apache Spark
  - 1.1.1. Czym różni się executor od driver
  - 1.1.2. Kiedy wykorzystywać executor a kiedy driverach
  - 1.1.3. Konfiguracja ilości pamięci ram i liczby executorów
  - 1.1.4. Jak dobrać odpowiednie parametry zależnie od ilości wejściowych danych
- 1.2. Lazy Evaluation - Transformations and Actions
  - 1.2.1. Jak zbudowany jest graf wykonywania transformacji
  - 1.2.2. Kiedy graf wykonywania jest wykorzystywany
  - 1.2.3. Jak re-używać wcześniej stworzone RDD z wcześniej wykonanymi transformacjami
  - 1.2.4. Jak i kiedy używać cache() i persist()
- 1.3. Shuffling - Przesyłanie danych między maszynami
  - 1.3.1. Które transformacje wymagają shufflingu (wide i narrow dependencies)
  - 1.3.2. Przewaga transformacji operujące na parach i metody typu reduceByKey() na PariRDD nad metodami typu reduce i operacjach na pojedynczych elementach
- 1.4. Partitioning Data
  - 1.4.1. Typy
    - 1.4.1.1. HashPartitioner
    - 1.4.1.2. RangePartitioner
    - 1.4.1.3. najlepsze praktyki wyboru
  - 1.4.2. Kiedy warto wykonać metodę repartition() a kiedy coalesce()
  - 1.4.3. Zasady dzielenia wyników na partycje - zapis w hdfs
- 1.5. Joining
  - 1.5.1. Typy joinów

BOTTEGA IT minds

1.5.2. Stosowalność

1.5.3. Wykorzystanie Join aby zminimalizować przesyłanie danych pomiędzy maszynami (reduce shuffling)

1.6. GroupBy

1.6.1. Kiedy używać (w ostateczności)

1.6.2. Kiedy wystarczy reduce na wyższym poziomie

1.6.3. Jak tworzyć przetwarzanie aby unikać groupBy

1.6.4. Alternatywy: CombineByKey

1.7. Spark API

1.7.1. DataFrame

1.7.2. DataSet

1.7.3. RDD

2. Pisanie przetwarzania BIG DATA używając APACHE SPARK

2.1. Tworzenie projektu który używa Apache Spark:

2.1.1. Jak skonfigurować skrypt uruchomieniowy joba

2.1.2. Jak pisać kod jobów aby był łatwo testowalny

2.2. Spark Context jako wejście do joba

2.2.1. Jak uzyskać sparkContextu

2.2.2. Jak stworzyć HiveContext (SQlContext) gdy pobieramy dane z SQL

2.2.3. Zasada jednego SparkContext

2.3. Możliwości pobierania danych przez Spark (baza danych, hdfs, avro, text file, csv, json …)

2.3.1. Tworzenie loadera do ustrukturyzowanych danych avro

2.3.2. Tworzenie loadera do nieustrukturyzowanych danych csv

2.3.3. Wczytywanie danych z hdfs

2.4. Implementowanie logiki w paradygmacie Map/Reduce

BO·TT·EGA
IT minds

2.4.1. Wykonywanie kodu przez driver vs executor

2.4.2. Operowanie na RDD i pairRDD

2.5. Testowanie Spark jobów

2.5.1. Jak tworzyć poprawnie test używający SparkContext aby unikać race conditions pomiędzy testami

2.5.2. Jak poprawnie mockować źródła danych do testu