BO·TT·EGA
IT minds

Program szkolenia:

Tworzenie platform Big Data z wykorzystaniem technologii z rodziny Apache

Informacje:

Nazwa:	**Tworzenie platform Big Data z wykorzystaniem technologii z rodziny Apache**
Kod:	**BigDataML-Apache**
Kategoria:	BigData, streaming i Machine Learning
Odbiorcy:	architekci, analitycy, developerzy
Czas trwania:	3 dni
Forma:	50% wykłady 50% ćwiczenia

Celem szkolenia jest zdobycie praktycznej wiedzy w rozwiązaniach Big Data.

Nauczymy się wykorzystywać popularne technologie Big Data (Apache Spark, Apache Kafka, Apache Airflow oraz Apache Druid). Dowiemy się jak zbudować złożone systemy Big Data od zera. Warsztaty praktyczne stanowią główny punkt szkolenia

BOTTEGA IT minds

Szczegółowy program:

1. Przegląd rozwiązań Big Data z rodziny Apache oraz wprowadzenia do przetwarzania danych

- 1.1. Przegląd rozwiązań Big Data z rodziny Apache
- 1.2. Scala for Big Data
  - 1.2.1. Case Class, Traits
  - 1.2.2. Tuples
  - 1.2.3. Lazy evaluation
  - 1.2.4. Interpolacja ciągów
  - 1.2.5. Pattern matching
  - 1.2.6. Companion object
  - 1.2.7. Kolekcje i przekształcenia
  - 1.2.8. For comprehension, mapowania
  - 1.2.9. Try / Either/ Option
  - 1.2.10. Implicits
- 1.3. Apache Spark - wprowadzenia
  - 1.3.1. RDD, DataFrame, Dataset
  - 1.3.2. Lazy evaluation
  - 1.3.3. Transformacje i akcje
  - 1.3.4. Spark vs Hadoop
  - 1.3.5. DataFrame vs Dataset API

2. Przetwarzanie danych z wykorzystaniem Apache Spark oraz nowoczesna hurtownia danych - Apache Druid

- 2.1. Warsztaty: Spark - jak wzbogacić swoje dane?
- 2.2. Apache Spark - architektura i optymalizacje
  - 2.2.1. Architektura (driver, worker, executor...)

BO·TT·EGA
IT minds

2.2.2. Optymalizacja jobów i parametrów

2.2.3. Deployment

2.2.4. Shuffling

2.2.5. Typowe błędy - key-skew,serializacja, OOM

2.2.6. Broadcast, repartition, caching, execution plans, optymalizacja

2.2.7. Spark internals - joins, group by

2.3. Apache Druid

2.3.1. Architektura

2.3.2. Struktury danych

2.3.3. Zarządzanie komponentami

2.3.4. Druid i platformy Big Data oparte na Apache Hadoop

2.3.5. Przetwarzanie real-time i batch

3. Streaming i orkiestracja

3.1. Apache Kafka

3.1.1. Wzorzec Pub/Sub. Różnica pomiędzy modelami push oraz pull

3.1.2. Architektura

3.1.3. Topicki

3.1.4. Producent i konsument Kafkowy

3.1.5. Analiza skalowalności systemu opartego o Apache Kafka

3.1.6. Grupy konsumenckie

3.1.7. Replikacja i retencja

3.1.8. Zookeeper

3.2. Apache Airflow

3.2.1. Automatyzacja przetwarzania

BO·TT·EGA
IT minds

3.2.2. Tworzenie data pipeline - Definiowanie Acyklicznych Grafów Skierowanych Przetwarzania (DAG)

3.2.3. Architektura