BO·TT·EGA
IT minds

Program szkolenia:

Rozwiązywanie problemów big data używając narzędzi z ekosystemu Hadoop

Informacje:

Nazwa:	**Rozwiązywanie problemów big data używając narzędzi z ekosystemu Hadoop**
Kod:	**BigDataML-hadoop**
Kategoria:	BigData, streaming i Machine Learning
Grupa docelowa:	
Czas trwania:	3 dni
Forma:	30% wykłady / 70% warsztaty

Szkolenie demonstruje praktyczne użycie technologii z ekosystemu Hadoop do rozwiązywania codziennych problemów Big Data.

Podczas warsztatów praktycznych uczestnicy nauczą się takich narzędzi jak HDFS, Kafka, Spark, Streaming, HBase. Podczas każdej z sekcji będziemy brać konkretny problem biznesowy z danej domeny i przy użyciu odpowiednich narzędzi Big Data będziemy rozwiązywać go w sposób efektywny.

Zalety szkolenia:

- Dobór narzędzi do klasy problemu
- Realne przykłady o realistycznym poziomie złożoności
- Najlepsze praktyki i typowe pułapki

BOTTEGA IT minds

Szczegółowy program:

1. Umówienie i poznanie narzędzi z ekosystemu hadoop

- 1.1. Hadoop i HDFS
- 1.2. YARN - schedulowanie jobow
- 1.3. Apache Hive - sql interface na HDFS
- 1.4. Apache Kafka - message queue
- 1.5. Hadoop Hbase - baza danych zbudowana na Hadoop
- 1.6. Apache Spark i Spark MLLib - biblioteka to przetwarzania big data
- 1.7. Spark GraphX - biblioteka do przetwarzania grafów w sposób rozproszony
- 1.8. Spark Streaming - biblioteka do przetwarzania streamingowego

2. Analiza Streamu płatności w sposób Streamingowy

- 2.1. pisanie przetwarzania w Spark Streaming
- 2.2. trzymanie rezultatów w Apache HBase

3. Odfiltrowywanie botów w kontekście Ad-Targeting

- 3.1. przetwarzanie w Spark
- 3.2. zapisywanie danych na Hadoop
- 3.3. udostępnianie danych przez interfejs Hive

4. Analiza transakcji

- 4.1. enrichowanie transakcji w sposób streamingowy
- 4.2. agregacje na streamach danych i znajdywanie TopSeller w danym oknie czasowym

5. Customer Churn Analysis

- 5.1. Analiza batchowa danych w Apache Spark

6. Internet of things (IoT)

- 6.1. Zapisywanie danych z sensorów w sposób Streamingowy w Apache Hbase

BOTTEGA IT minds

6.2. Skanowanie i liczenie danych w Hbase

6.3. Obliczanie statystyk z danych przetrzymywanych w Hbase w sposób batchowy. Zapisywanie rezultatów w Hbase

7. Używanie Grafów do rozwiązywania problemów

7.1. Wstęp do Spark GraphX

7.2. Nauka API GraphX

7.3. counting degree of a Graph

7.4. Connected Components

7.5. Page Rank

8. Detekcja Anomalii - budowanie wykrywania anomalii bazując na ruchu HTTP używając Spark MLLib

8.1. Spark MLlib

8.2. K-Means clustering

8.3. Wykrywanie anomalii

9. Analiza text - znajdywanie autora postu bazując tylko na treści postu

9.1. Wyciąganie feature vector z nie ustrukturyzowanego tekstu

9.2. Suprevised Learning - Logistic Regression

9.3. Unsuprevised Learning - GMM

10. Cloudera Sandbox - uzywanie cloudera sandbox z wszystkimi narzędziami z Hadoop Ecosystem

10.1. ładowanie danych do HDFS używając Sqoop

10.2. analiza danych używając Hive i interfejsu graficznego Hue

11. Personalizacja - budowanie silnika rekomendacji używając Spark MLLib i Collaborative Filtering

11.1. budowanie silnika bazując na danych o preferencjach użytkowników oceniających filmy