BO·TT·EGA
IT minds

Program szkolenia:

Apache Spark i Hadoop dla analityków danych

Informacje:

Nazwa:	**Apache Spark i Hadoop dla analityków danych**
Kod:	**data-science-spark**
Kategoria:	Analiza danych
Grupa docelowa:	developerzy architekci analitycy
Czas trwania:	3 dni
Forma:	30% wykłady / 70% warsztaty

Szkolenie przeznaczone jest dla uczestników pracujących w R z małymi zbiorami danych i chcących się nauczyć SQLa oraz przetwarzania danych w sposób rozproszony z użyciem Hadoopa i Sparka.

Szkolenie rozpoczyna się od krótkiego wprowadzenia (w razie potrzeb) do SQL i API Sparka oraz narzędzi: łączenie RStudio i Hadoopa.

Kolejne ćwiczenia pozwalają nabyć konkretne umiejętność w zakresie wizualizacji danych i uczenia maszynowego w trakcie budowania kompletnej aplikacji.

W trakcie ćwiczeń możemy wybrać dowolną klasę problemu ML.

Zalety szkolenia:

- Przetwarzania dużych zbiorów danych
- Dobór narzędzia i modelu do klasy problemu
- Zrozumienie pryncypiów obliczeń rozproszonych

BOTTEGA IT minds

Szczegółowy program:

1. Wprowadzenie

1.1. Krótka historia Big Data - skąd potrzeba na przechowywanie i przetwarzanie danych w sposób rozproszony

1.2. Hadoop i Apache Spark - architektura i możliwości

1.3. Różne metodyki pracy z Big Data i R

1.4. Dlaczego format i sposób zapisu danych ma znaczenie?

2. Przetwarzanie danych w sposób rozproszony: SparkSQL oraz SparkR (każde z ćwiczeń wykonujemy najpierw w SQL, a potem w SparkR)

2.1. Ładowanie danych do kontekstu, oglądanie danych

2.2. Filtrowanie (WHERE) i projekcja (SELECT)

2.3. Agregacje jednego zbioru danych (GROUP BY i HAVING)

2.4. Łączenie zbiorów danych (JOIN)

2.5. Różne algorytmy łączenia danych w świecie rozproszonym: SortMergeOuterJoin i BroadcastHashJoin

2.6. Praca z oknami danych i funkcje analityczne (lag, row_number)

2.7. Zapis danych i pobieranie wyników do RStudio

3. Wykorzystanie języka R w sposób rozproszony

3.1. Funkcja R'a jako UDF (dapply/gapply)

3.2. Łączenie się do klastra Hadoopa bezpośrednio z RStudio i konwersje między R'owymi i Sparkowymi ramkami danych (Dataframes)

4. Wizualizacja danych

4.1. Przetwarzanie danych w Sparku i obrazowanie wyników w R (ggplot)

4.2. Apache Superset - otwartoźródłowy system do intuicyjnego tworzenia dashboardów z wykorzystaniem SQL (Presto SQL oraz Hive na Hadoopie) (łącznie 1.5h)

5. Uczenie maszynowe na dużych danych

5.1. Regresja logistyczna w Spark ML Lib

BOTTEGA IT minds

5.2. Przekrój pozostałych algorytmów dostępnych w Sparku. Hybrydowy sposób pracy z ML między Sparkiem i R

6. Budujemy kompletną aplikację wykorzystującą Sparka, algorytmy rekomendacyjne i R'a - łącznie 3h

6.1. Analiza możliwych podejść do danych

6.2. Porównanie wydajności różnych sposób

6.3. Optymalizacja parametrów modelu

6.4. Jak debugować problemy z przetwarzaniem

7. Porównanie pracy z Big Data w środowisku chmurowym (na przykładzie AWS) z klasycznymi dystrybucjami - 1h*

8. Przegląd komercyjnych narzędzi do wizualizacji danych, porównanie funkcjonalności - 1h*