

Program szkolenia:

Pyspark i Jupyter - niezbędny analityków i developerów Big Data

Informacje:

Nazwa:	Pyspark i Jupyter - niezbędny analityków i developerów Big Data
Kod:	data-science-jupyter
Kategoria:	Analiza danych
Grupa docelowa:	analitycy developerzy
Czas trwania:	3 dni
Forma:	20% wykłady / 80% warsztaty

W trakcie warsztatów uczestnicy poznają od podstaw API silnika Spark w języku Python i nauczą się wykorzystywać go praktycznie tworząc notatniki w aplikacji Jupyter.

Szkolenie jest dedykowane dla analityków, którzy znają podstawy Pythona i chcieliby rozszerzyć swoje umiejętności na optymalne przetwarzanie dużych zbiorów danych.

Zalety szkolenia:

- problemy o realnej złożoności
- sprawdzone narzędzia
- dobór modelu do klasy problemu

Szczegółowy program:

1. Wprowadzanie

- 1.1. Rys historyczny pracy w świecie Big Data - jak rozwijały się Hadoop, Hive i Spark
- 1.2. Budowa Sparka, możliwości w zakresie dostępu do danych składowanych w różnych systemach
- 1.3. Dlaczego Spark jest lepszy/gorszy niż rozwiązania SQL'owe na Hadoopie?
- 1.4. Metodologia pracy: akcje, transformacje, frameworki dostępne w PySparku (DF, RDD)

2. Praca z notebookami Jupyter

- 2.1. Konfiguracja środowiska
- 2.2. Zarządzanie bazą notatników, wersjonowanie w repozytorium
- 2.3. Wizualizacje - wykresy, mapy
- 2.4. Rozszerzenia umożliwiające interaktywną pracę

3. PySpark

- 3.1. Uzyskiwanie dostępu do danych składowanych w różnych formatach
- 3.2. Oglądanie danych
- 3.3. Filtrowanie i projekcja (odpowiedniki WHERE i SELECT)
- 3.4. Konwersje pomiędzy DF i RDD
- 3.5. Agregacje (grupowanie) danych
- 3.6. Łączenie zbiorów danych (odpowiednik JOIN)
- 3.7. Wykorzystywanie języka Python do pracy w API Dataframe (User Defined Functions)
- 3.8. Praca na oknach danych, mierzenie wydajności różnych podejść
- 3.9. Zapisywanie danych w optymalny sposób, repartycjonowanie

4. Różne podejścia do eksploracji danych

- 4.1. Cachowanie
- 4.2. Ekosystem sparka (thriftserver, historyserver)

4.3. Ładowanie "nietypowych" danych - zapytania do API REST

5. **Uczenie maszynowe w MLlib**

5.1. Regresja liniowa i logistyczna

5.2. Praca z tekstem

5.3. Klastrowanie danych

5.4. Association rule learning