**SZKOLENIE ŚREDNIO ZAAWANSOWANE**

# Big Data i platforma Hadoop - wprowadzenie

BIGDATA/BASE

**Czas trwania: 3 dni**

Szkolenie skupia się na platformie Hadoop, jej podstawowych komponentach oraz najważniejszych narzędziach

## Cele szkolenia

- Wprowadzenie do koncepcji Big Data
- Zapoznanie z platformą Hadoop oraz dostępnymi narzędziami

## Zalety

- Praktyczne wprowadzenie do tematyki Big Data
- Warsztatowy charakter zajęć, umożliwiający przyswojenie wiedzy, niezbędnej do przeprowadzania analizy dużych ilości danych
- Praktyka przed teorią - wszystkie szkolenia technologiczne prowadzone są w formie warsztatowej. Konieczna teoria jest wyjaśniana na przykładzie praktycznych zadań
- Konkretne umiejętności - w ramach każdego szkolenia rozwijamy praktyczne umiejętności związane z daną technologią i tematyką
- Nauka z praktykami - wszyscy trenerzy na co dzień pracują w projektach, gwarantuje to dostęp do eksperckiej wiedzy i praktycznego know-how

## Dla kogo?

- Analitycy i programiści, którzy chcą rozpocząć przygodę z analizą dużych zbiorów danych

## Wymagania

- Podstawy SQL oraz hurtowni danych
- Podstawowa umiejętność programowania, najlepiej w: Java, Python lub Scala

## Program

1. Wprowadzenie do Big Data
    a. Czym jest Big Data?
        -Definicja
        -Geneza i historia
        -Problemy Big Data
        -Zastosowania i przypadki użycia
    b. Typy przetwarzania w Big Data
        -Przetwarzanie wsadowe
        -Przetwarzanie strumieni danych
    c. Dystrybucje Big Data
    d. Rozwiązania w chmurze
2. Apache Hadoop
    a. Wprowadzenie do platformy Hadoop
        -Rynek Big Data na świecie
        -Rynek Big Data w Polsce
        -Hadoop a RDBMS
        -Historia Hadoop
        -Wprowadzenie do komponentów Hadoop
        -Podstawowe narzędzia
    b. MapReduce
        -Podstawy przetwarzania MapReduce
        -Podstawowe pojęcia
        -Przepływ danych
        -Przykłady
        -MapReduce "Classic", czyli Java
        -Optymalizacja przetwarzania i elementy zaawansowane
        -Hadoop Streaming, czyli Python, PHP, i...
        -Czy to takie "group by"?
        -Warsztaty MapReduce
    c. HDFS
        -Wprowadzenie do rozproszonego systemu plików
        -Podstawowe cechy i pojęcia
        -Architektura
        -Zarządzanie za pomocą linii komend
        -Dostęp przez WWW
        -Korzystanie za pomocą API
        -Importowanie i eksportowanie danych
        -Formaty plików popularne w Big Data
        -Wykorzystanie kompresji danych
    d. YARN
        -Wprowadzenie

-Zasada działania i podstawowa konfiguracja
-Sposoby szeregowania zadań
-Podstawowe operacje
-Uruchamienie zadań MapReduce
-Zarządzanie zadaniami uruchomionymi w oparciu o architekturę YARN
-Warsztaty HDFS i YARN

3. Apache Pig
   a. Wprowadzenie
   b. Architektura
   c. PigLatin w szczegółach
   d. Uruchamianie zadań
   e. Różne źródła danych
   f. Funkcje wbudowane
   g. Biblioteki, makra
   h. Funkcje użytkownika (UDF)
   i. Warsztaty Pig
4. Apache Hive
   a. Czym jest Hive
   b. Model danych w Hive
   c. Formaty przechowywania danych
      -Format wierszowy vs. kolumnowy
      -ORCFile
   d. HiveSQL
      -Źródła danych
      -Selekcja, projekcja, łączenie, grupowanie
      -DML
      -Rozszerzenia grupowania i funkcje analityczne
   e. Uruchamianie zadań
   f. Różne źródła danych
   g. Korzystanie w konsoli
   h. Interfejsy użytkownika
   i. Funkcje wbudowane
   j. Funkcje użytkownika (UDF)
   k. Wykorzystanie Apache Tez i optymalizacja zadań
   l. Warsztaty Hive
5. Wprowadzenie do baz danych NoSQL
   a. Historia
   b. Podstawowe cechy
   c. Przyczyny sukcesu
   d. Problem spójności
      -Własności BASE vs. ACID
      -Własności CAP
      -Twierdzenie CAP
   e. Przegląd modeli NoSQL

f. Powiązane technologie

6. HBase
   a. Wprowadzenie
   b. Case Study
   c. Organizacja danych
   d. Widoki danych: koncepcyjny i fizyczny
   e. Architektura
   f. Jak to wszystko działa?
   g. Interfejsy
      -HBase shell
      -Phoenix - JDBC
   h. Warsztaty HBase
   i. Warsztaty HBase z zewnętrznych narzędzi: Pig i Hive