Kod szkolenia: **HADOOP**

Tytuł szkolenia: **Projektowanie rozwiązań Big Data z wykorzystaniem Apache Hadoop & Family**

Dni: 5

Opis:



**Adresaci szkolenia:**

Szkolenie jest adresowane do programistów, architektów oraz administratorów aplikacji, którzy chcą tworzyć lub utrzymywać systemy, w których wolumen przetwarzanych danych ma najwyższy priorytet i przekracza możliwości tradycyjnych architektur i systemów takich jak relacyjne bazy danych czy nawet hurtownie danych. Szkolenie jest także kierowane do osób, które chcą uzupełnić swoją wiedzę o pojęcia związane z Big Data, MapReduce, NoSQL oraz ich realizacją z wykorzystaniem oprogramowania Apache Hadoop & Family.

**Cel szkolenia:**

Uczestnicy szkolenia zdobędą przekrojową wiedzę dotyczącą takich pojęć jak algorytm MapReduce, poznają założenia Big Data, BigTable, rozproszone systemy plikowe DFS, bazy danych typu NoSQL. Dzięki temu będą mogli wybrać właściwy zestaw narzędzi i technik dla swoich projektów. Szkolenie, poza ogólnym wprowadzeniem do pojęć teoretycznych, skupia się na stosie produktowym wybudowanym wokół Apache Hadoop.

**Mocne strony szkolenia:**

Program obejmuje zarówno ogólne wprowadzenie w tematykę Big Data jak i całościowe przedstawienie stosu produktowego wokół Apache Hadoop. Szkolenie jest unikalne, gdyż

tematyka poruszana w jego trakcie nie jest wyczerpująco ujęta w dostępnej literaturze, a wiedza na ten temat jest mocno rozproszona. Program jest ciągle uaktualniany ze względu na szybki rozwój rozwiązań, których dotyczy szkolenie.

**Wymagania:**

Od uczestników wymagana jest podstawowa znajomość baz danych, podstawowa umiejętność programowania w języku Java.

**Parametry szkolenia:**

5*8 godzin (5*7 netto) wykładów i warsztatów, z wyraźną przewagą warsztatów. W trakcie warsztatów, oprócz prostych ćwiczeń, uczestnicy rozwiązują problemy przetwarzania danych implementując własne algorytmy z wykorzystaniem paradygmatu MapReduce, modelują struktury danych bazy NoSQL, wykonują podstawowe czynności administracyjne. Wielkość grupy: maks. 8-10 osób

## Program szkolenia:

1. Wstęp do BigData
    I. Definicja
    II. Czym jest BigData?
        i. Geneza i historia BigData
        ii. Strony w projektach BigData
    III. Problemy BigData
    IV. Typy przetwarzania BigData
        i. Wsadowe
        ii. Strumieniowe
    V. Dystrybucje Big Data
    VI. Rozwiązania w chmurze
2. Apache Hadoop
    I. HDFS
        i. Wprowadzenie do rozproszonego systemu plików
        ii. Zarządzanie za pomocą linii komend
        iii. Dostęp przez WWW
        iv. Korzystanie za pomocą API
        v. Importowanie i eksportowanie danych
    II. MapReduce
        i. Wprowadzenie do paradygmatu MapReduce
        ii. Formaty wejścia i wyjścia, tworzenie własnych formatów
        iii. Wbudowane i własne typy danych
        iv. Partitioner i Combiner, kiedy i jak używać
        v. Liczniki danych
        vi. Konfiguracja zadań za pomocą parametrów
        vii. Łańcuchy zadań MapReduce
        viii. Wykorzystanie kompresji dla zmniejszenia liczby danych

        ix. Optymalizacja zadań MapReduce
    III. YARN
        i. Wprowadzenie
        ii. Uruchamianie i zarządzanie zadaniami uruchomionymi w oparciu o architekturę YARN
3. Apache Pig
    I. Wstęp
    II. PigLatin w szczegółach
    III. Funkcje wbudowane
    IV. Funkcje użytkownika (UDF)
    V. Wydajność
    VI. Testowanie i diagnostyka
4. Apache Hive
    I. Czym jest Hive
    II. Architektura
    III. Unikalne cechy Hive
    IV. HiveCLI
    V. HiveQL
    VI. PigLatin vs HiveQL
    VII. Tabele w Hive
5. Apache HBase
    I. Wstęp
        i. Wprowadzenie do baz danych NoSQL
        ii. Przyczyna powstania baz chmurowych
        iii. Spójność, Dostępność, Odporność na partycjonowanie
        iv. Twierdzenie CAP
        v. Co różni bazy NoSQL od baz relacyjnych
        vi. Podstawowe parametry baz NoSQL
        vii. Klasyfikacja i przegląd baz NoSQL
        viii. Unikalne cechy HBase
    II. Architektura HBase
        i. Elementy składowe
            A. Master Servers
            B. Regiony i Region Servers
            C. Zookeeper
        ii. Zasada działania
    III. Model danych
        i. Model koncepcyjny a fizyczny
        ii. Przestrzeń nazw
        iii. Tabela
        iv. Wiersz
        v. Kolumna
        vi. Wersja
        vii. Komórka
    IV. Wykorzystanie HBase
        i. HBase API

ii. Z poziomu platformy Apache Hadoop i zadań MapReduce
iii. Za pomocą API zewnętrznych - REST API, Apache Thrift etc.
iv. Testowanie aplikacji HBase
V. Zarządzanie
i. Optymalizacja i konfiguracja
ii. Dobre praktyki korzystania z bazy
iii. Diagnostyka
iv. Snapshoty i backup danych
v. Podstawowe operacje administracyjne
vi. Bezpieczeństwo
VI. Apache HBase w porównaniu do innych baz danych NoSQL
6. Apache Spark
I. Wstęp
i. Historia
ii. Spark a Hadoop
iii. Rozproszone kolekcje obiektów Resilient Distributed Datasets (RDDs)
iv. Przetwarzanie w pamięci a z dysku
v. Architektura
vi. Warianty uruchomienia klastra
A. Własny klaster Spark
B. Apache Mesos
C. Apache YARN
II. Spark Core
i. Wstęp
ii. Java vs Spark vs Python
iii. RDD vs Dataset vs DataFrame
iv. Łączenie z klastrem
v. Rozproszone dane
vi. Operacje RDD
vii. Transformacje
viii. Akcje
ix. Współdzielone zmienne
x. Uruchomienie i testowanie
xi. Dostrajanie zadań
xii. Serializacja
xiii. Pamięć
III. Spark SQL
i. Wstęp
ii. Spark SQL a Hive
iii. Zasada działania
iv. Dane i schematy
v. Zapytania
vi. Integracja z Hive
vii. Uruchomienie i testowanie
IV. Spark Streaming
i. Wstęp

        ii. Zasada działania
        iii. Strumienie
        iv. Wejście
        v. Transformacja
        vi. Wyjście
        vii. Uruchomienie i testowanie
    V. Spark MLlib
        i. Wstęp
        ii. Dostępne algorytmy
        iii. Transformery i estymatory
        iv. Dostępne transformacje
        v. Budowa pipeline'u
        vi. Uczenie modeli
7. Apache Kafka
    I. Wprowadzenie
        i. Historia
        ii. Zastosowania
        iii. Terminologia
        iv. Porównanie z innymi narzędziami typu producent konsument
    II. Korzystanie z API
        i. Wysyłanie wiadomości
        ii. Odbieranie wiadomości
        iii. Serializacja
        iv. Konfiguracja producentów i konsumentów
        v. Projektowanie rozwiązań w oparciu o Apache Kafka
        vi. Integracja z Hadoop i Spark
    III. Zarządzanie
        i. Instalacja
        ii. Konfiguracja
        iii. Replikacja
        iv. Kompresja danych
8. Apache Oozie
    I. Akcje HDFS
    II. Akcje MapReduce
    III. Akcje Spark
    IV. Akcje Pig
    V. Akcje Hive
    VI. Akcje Subworkflow
9. Zarządzanie i monitoring infrastrukturą Apache Hadoop & Family
    I. Apache Ambari
10. Przegląd Apache Hadoop & Family