Kod szkolenia: **HADOOP/P**

Tytuł szkolenia: **Hadoop dla programistów**

Dni: 3

Partner merytoryczny SigDelta

Opis:

Adresaci szkolenia

Szkolenie jest adresowane do programistów, którzy chcą rozwijać systemy służące do składowania i/lub analizowania dużych zbiorów danych z wykorzystaniem platformy Apache Hadoop. Szkolenie jest dedykowane zarówno początkującym użytkownikom tej platformy jak i takim którzy mają już pierwsze kroki za sobą i chcą rozwinąć bądź ugruntować swoją wiedzę.

Cel szkolenia

Uczestnicy szkolenia zdobędą wiedzę niezbędną do rozpoczęcia pracy z systemem Apache Hadoop, w tym jak implementować wydajne algorytmy w oparciu o MapReduce oraz jak składować i importować dane do systemu. Przedstawione zostaną wzorce projektowe oraz tak zwane dobre praktyki programistyczne. Szkolenie kładzie nacisk zarówno na aspekty teoretyczne jak i przede wszystkim praktyczne.

Mocne strony szkolenia

Program obejmuje zarówno ogólne wprowadzenie w tematykę Big Data jak i szczegółowe przedstawienie narzędzi Apache Hadoop na poziomie pozwalającym zacząć pracę w tym środowisku. Szkolenie jest unikalne, gdyż tematyka poruszana w jego trakcie nie jest wyczerpująco ujęta w dostępnej literaturze, a wiedza na ten temat jest rozproszona. Program jest ciągle uaktualniany ze względu na szybki rozwój omawianych rozwiązań. Prezentowana wiedza jest wynikiem kilku lat praktyki trenerów w budowaniu systemów oparty o platformę Apache Hadoop.

Wymagania

Od uczestników wymagana jest podstawowa umiejętność programowania w języku Java oraz podstawy baz danych i języka SQL.

Parametry szkolenia

3*8 godzin (3*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. 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 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. Apache Hive w Spark
    i. Czym jest Hive
    ii. Architektura
    iii. Unikalne cechy Hive
    iv. HiveQL
    v. Tabele w Hive
        1.Wykorzystanie apache Hive w Spark

V. Spark Streaming
    i. Wstęp
    ii. Zasada działania
    iii. Strumienie
    iv. Wejście
    v. Transformacja
    vi. Wyjście
    vii. Uruchomienie i testowanie

VI. Spark MLlib
    i. Wstęp
    ii. Dostępne algorytmy
    iii. Transformery i estymatory
    iv. Dostępne transformacje

v. Budowa pipeline'u
vi. Uczenie modeli

4. 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
5. Przegląd Apache Hadoop & Family