Kod szkolenia: **BIG BATA/ANA**

Tytuł szkolenia: **Big Data dla analityków**

Dni: 4

Opis:

Adresaci szkolenia:

Analitycy danych oraz programiści, którzy chcą rozpocząć swoją przygodę z analizą dużych zbiorów danych.

Cel szkolenia:

Przekrojowe szkolenie mające na celu zapoznanie się z narzędziami przeznaczonymi dla analityka big data. Szkolenie skupia się na płynnym wejściu w podstawy każdego narzędzia, tak aby analityk danych mógł w przyszłości bez problemu poruszać się po ekosystemie Hadoop.

Mocne strony szkolenia:

Zapoznanie z wieloma narzędziami i językami programowania, szkolenie ma na celu pokazanie jak łatwo można analizować dane bez użycia konsoli i narzędzi IDE.

Wymagania:

Podstawy SQL, podstawowa umiejętność programowania, najlepiej w: Python, R, Java lub Scala

Parametry szkolenia:

Szkolenie trwa cztery dni, każdy dzień to 8h wykładów i warsztatów, z wyraźną przewagą warsztatów, w tym dwie przerwy kawowe i jedna obiadowa (7h netto). Wielkość grupy: maks. 8-10 osób.

Program szkolenia:

1. Wprowadzenie do Big Data
    I. Czym jest Big Data?
        i. Definicja
        ii. Geneza i historia Big Data
        iii. Problemy Big Data
        iv. Zastosowania i przypadki użycia
        v. Umiejętności w projektach Big Data
        vi. Big Data a Business Intelligence (Hurtownie danych)
        vii. Data Science i sztuczna inteligencja w Big Data
        viii. Bazy NoSQL
    II. Architektura systemów Big Data
        i. Przetwarzanie wsadowe
        ii. Architektura Lambda
        iii. Architektura Kappa
        iv. Data Lake
    III. Dystrybucje Big Data
        i. Geneza powstania
        ii. Zastosowania i przypadki użycia
        iii. Porównanie popularnych dystrybucji Big Data
        iv. Zalety i wady korzystania z dystrybucji Big Data
    IV. Przegląd ekosystemu Apache Hadoop
    V. Rozwiązania w chmurze
2. Apache Hadoop
    I. HDFS
        i. Wprowadzenie do rozproszonego systemu plików
        ii. Architektura
        iii. Zarządzanie za pomocą linii komend
        iv. Dostęp przez WWW
        v. Korzystanie za pomocą API
        vi. Importowanie i eksportowanie danych
        vii. Formaty plików popularne w Big Data
        viii. Wykorzystanie kompresji danych
    II. YARN & MapReduce
        i. Wprowadzenie do platformy obliczeniowej YARN
        ii. Zasada działania i podstawowa konfiguracja YARN
        iii. Podstawowe operacje YARN
        iv. Przetwarzanie zadań za pomocą MapReduce
        v. Uruchamianie i zarządzanie zadaniami uruchomionymi w oparciu o

architekturę YARN

3. Apache Hive
    I. Czym jest Hive
    II. Architektura
    III. Unikalne cechy Hive
    IV. HiveQL
    V. Model danych w Hive
    VI. Uruchamianie zadań
    VII. Różne źródła danych
    VIII. Korzystanie w konsoli
    IX. Interfejsy użytkownika
    X. Funkcje wbudowane
    XI. Funkcje użytkownika (UDF)
    XII. Wykorzystanie Apache Tez i optymalizacja zadań
4. Apache Pig
    I. Wstęp
    II. PigLatin w szczegółach
    III. PigLatin vs HiveQL
    IV. Uruchamianie zadań
    V. Różne źródła danych
    VI. Funkcje wbudowane
    VII. Funkcje użytkownika (UDF)
5. 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 YARN
            C. Apache Mesos
            D. Goooogle Kubernetes
        vii. Konfiguracja i zarządzanie
    II. Spark Core
        i. Wstęp
        ii. Języki programowania (Scala vs Python vs Java vs R)
        iii. RDD vs Dataset vs DataFrame
        iv. Łączenie z klastrem
        v. Rozproszone dane
        vi. Operacje RDD
            A. Transformacje
            B. Akcje
        vii. Współdzielone zmienne
        viii. Uruchomienie i testowanie

ix. Dostrajanie zadań
    A. Serializacja
    B. 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 Structured Streaming
i. Wstęp
ii. Zasada działania
iii. Strumienie
    A. Wejście
    B. Transformacja
    C. Wyjście
iv. 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

6. Środowisko pracy i wizualizacja danych
I. Apache Zeppelin
II. Jupyter
III. HUE
7. Przegląd innych narzędzi Sztucznej Inteligencji i Data Science

Sages sp. z o.o., ul. Nowogrodzka 62c, 02-002 Warszawa
tel. +48 22 203 56 00 fax: +48 22 203 56 01
e-mail: biuro@sages.com.pl web: www.sages.com.pl