Kod szkolenia: **BIGDATA/F**

Tytuł szkolenia: **Wprowadzenie do technologii Big Data**

Dni: 1

Opis:

Adresaci szkolenia

Szkolenie jest adresowane do analityków i programistów, którzy chcą zrobić swój pierwszy krok w kierunku poznania Big Data - technologii, gdzie wolumen przetwarzanych danych ma najwyższy priorytet i przekracza możliwości tradycyjnej architektury i systemów takich jak relacyjne bazy danych czy nawet hurtownie danych.

Cel szkolenia

Uczestnicy szkolenia zdobędą podstawową wiedzę dotyczącą problemów skali Big Data, zrozumieją algorytm MapReduce, poznają BigTable, bazy NoSQL na przykładzie HBase oraz rozproszone systemy plikowe HDFS, poznają narzędzia przetwarzania danych Spark i Hive. Uczestnicy będą wiedzieli jakie są zalety i wady danych technologii, będą wiedzieli kiedy użyć danej technologii.

Mocne strony szkolenia

Program oferuje szybki przegląd podstawowych technologii z ekosystemu Apache Hadoop. Oprócz prezentacji dla uczestników jest przygotowany warsztat, gdzie w praktyce będą mieli okazje samodzielnie eksplorować zbiory danych.

Wymagania

Od uczestników szkolenia wymagana jest podstawowa wiedza z SQL, bash'a, Python (lub innego języka skryptowego), Java.

Parametry szkolenia

8 godzin (7 godzin netto) wykładów i warsztatów (z wyraźną przewagą warsztatów).

Program szkolenia:

1. Wstęp do BigData

sages

I. Definicja
II. Czym jest BigData?
    i. Geneza i historia BigData
    ii. Strony w projektach BigData
    iii. Big Data a Hurtownie danych
    iv. Bazy NoSQL
III. Problemy BigData
IV. Typy przetwarzania BigData
    i. Wsadowe
    ii. Strumieniowe
V. Przegląd ekosystemu Apache Hadoop
VI. Dystrybucje Big Data
VII. Rozwiązania w chmurze

2. Wprowadzenie do Apache Hadoop
    I. Architektura
    II. Przechowywanie danych w HDFS
    III. Przetwarzanie danych w opraciu o YARN
    IV. Wprowadzenie do MapReduce
3. Wprowadzenie do analizy danych na przykładzie Hive
    I. Architektura
    II. Tryby pracy
    III. Typy danych
    IV. Składnia
    V. Formaty danych
    VI. Porównanie z Pig
    VII. Warsztat Hive
4. Przetwarzanie danych w oparciu o 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
    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
    III. Spark SQL
        i. Wstęp
        ii. Spark SQL a Hive

sages

iii. Zasada działania
iv. Dane i schematy
v. Zapytania
vi. Integracja z Hive

5. Wprowadzenie do NoSQL na podstawie HBase
   I. Czym jest NoSQL, NoSQL vs bazy relacyjne
   II. Przegląd baz nierelacyjnych, CAP theorem
   III. Projektowanie baz nierelacyjnych
   IV. Architektura
   V. Model danych
   VI. Korzystanie z HBase
6. Monitorowanie i zarządzanie klastrem na przykładzie Ambari

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