Kod szkolenia: **SPARK**

Tytuł szkolenia: **Przetwarzanie Big Data z użyciem Apache Spark**

Dni: 2

Partner merytoryczny SigDelta

Opis:

Adresaci szkolenia

Szkolenie jest przeznaczone głównie dla programistów i analityków danych, którzy chcą się zapoznać z podstawami przetwarzania Big Data, bardzo dużych zbiorów danych przekraczającej możliwości tradycyjnego przetwarzania, z użyciem narzędzi z rodziny Apache Spark. Szkolenie stanowi zarówno dobrą podstawę dla osób pragnących zacząć pracę z Big Data, jak i osób z uprzednim doświadczeniem w tego typu systemach, np. rodziny Apache Hadoop, pragnących nauczyć się nowej technologii.

Cel szkolenia

Uczestnicy szkolenia zapoznają się z nowym problemem jakim jest analiza bardzo dużych zbiorów danych (Big Data) z różnych źródeł. Na szkoleniu przedstawiony zostanie podstawowy zbór problemów Big Data i ich rozwiązania z pomocą narzędzi rodziny Apache Spark. Ponadto, uczestnicy będą świadomi zalet i wad Apache Spark w podejściu do ich rozwiązania ich problemów biznesowych. Dodatkowo, kurs pozwala uczestnikom na zapoznanie się z szybko zmieniającą się dziedziną jaką jest Big Data i nowym podejściem do rozwiązywania problemów jaki prezentuje Apache Spark.

Mocne strony szkolenia

Szkolenie jest prowadzone przez osoby na co dzień pracujące z problemami Big Data i mającymi praktyczne doświadczenie w tej dziedzinie. Z tego powodu szkolenie często wykracza poza dostępne choć często rozproszone materiały. Ponadto, program jest ciągle uaktualniany ze względu na szybki rozwój rozwiązań, których dotyczy szkolenie.

Wymagania

Szkolenie wymaga podstawowej umiejętności programowania w Javie (zakres szkolenia: J/JP), Scali (zakres szkolenia: J/SCL) lub Pythonie (zakres szkolenia: PT/PP); preferowanym językiem szkolenia jest Python. Przydatne umiejętności: znajomość zagadnień związanych z

przetwarzaniem danych, programowanie funkcjonalne, przetwarzanie rozproszone, systemy *nix.

2 dni robocze, 2*7 godz roboczych, grupa 8-10 osób. Szkolenie w formie prezentacji i warsztatów programistycznych.

Program szkolenia:

1. Wstęp do BigData
    I. Definicja
    II. Czym jest BigData?
    III. Geneza i historia BigData
    IV. Strony w projektach BigData
    V. Problemy BigData
    VI. Typy przetwarzania BigData
        - Wsadowe
        - Strumieniowe
2. Apache Spark
    I. Wstęp
    II. Historia
    III. Spark a Hadoop
    IV. Paradygmat programowania MapReduce
    V. Rozproszone kolekcje obiektów Resilient Distributed Datasets (RDDs)
    VI. Przetwarzanie w pamięci a z dysku
    VII. Architektura
    VIII. Warianty uruchomienia klastra
        - Własny klaster Spark
        - Apache Mesos
        - Apache YARN
    IX. Administracja
3. Spark Core
    I. Wstęp
    II. Java vs Scala vs Python
    III. Łączenie z klastrem
    IV. Rozproszone dane
    V. Operacje RDD
        - Transformacje
        - Akcje
    VI. Współdzielone zmienne
    VII. Uruchomienie i testowanie
    VIII. Dostrajanie zadań
        - Serializacja
        - Pamięć
4. Spark SQL
    I. Wstęp
    II. Spark SQL a Hive