

Kod szkolenia: **BIGDATA/STR**

Tytuł szkolenia: **Strumieniowe przetwarzanie Big Data**

Dni: 2



Partner merytoryczny

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, w czasie (niemal) rzeczywistym z użyciem narzędzi z rodziny Apache Storm i Spark. Szkolenie stanowi zarówno dobrą podstawę dla osób pragnących zacząć pracę ze strumieniowym przetwarzaniem 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 w czasie (niemal) rzeczywistym bardzo dużych zbiorów danych (Big Data) z różnych źródeł. Na szkoleniu przedstawiony zostanie podstawowy zbiór problemów strumieniowego przetwarzania Big Data i ich rozwiązania z pomocą narzędzi rodziny Apache Storm i Spark. Ponadto, uczestnicy będą świadomi zalet i wad przedstawionych technologii w podejściu do 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 z użyciem Apache Storm i 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 Java. Przydatne umiejętności: znajomość zagadnień związanych z przetwarzaniem danych, programowanie funkcjonalne, przetwarzanie rozproszone, systemy *nix.

Parametry szkolenia

2*7h (netto). Szkolenie w formie prezentacji i warsztatów programistycznych.

Wielkość grupy: maks. 8-10 osób

Program szkolenia:

- Wstęp do Big Data
 - Definicja
 - Czym jest Big Data?
 - Geneza i historia Big Data
 - Strony w projektach Big Data
 - Problemy Big Data
 - Typy przetwarzania Big Data
 - Wsadowe
 - Strumieniowe
 - Kiedy Hadoop nie wystarcza
 - Przetwarzanie danych (niemal) w czasie rzeczywistym
 - Definicja
 - Zalety i wady
 - Przykłady
 - Typy gwarancji dostarczenia wiadomości
 - Najwyżej raz (at-most-once)



- Przynajmniej raz (at-least-once)
- Dokładnie raz (exactly-once)
- Apache Storm
 - Wstęp
 - Historia
 - Architektura
 - Warianty uruchomienia
 - Własny klaster
 - Apache Mesos
 - Apache YARN
 - Administracja
 - Zasada działania
 - Topologie
 - Strumienie
 - Spouts
 - Bolts
 - Model danych
 - Grupowanie strumieni
 - Mieszanie języków programowania
 - Gwarancja przetwarzania wiadomości
 - Uruchamianie i testowanie
 - Serwer RPC
 - Kolejki wejściowe



- Kestrel
- Apache Kafka
- Trident
 - Zasada działania
 - Model danych
 - Stan
 - Uruchomienie i testowanie
- Apache Spark
 - Wstęp
 - Historia
 - Rozproszone kolekcje obiektów Resilient Distributed Datasets (RDDs)
 - Przetwarzanie w pamięci a z dysku
 - Architektura
 - Warianty uruchomienia klastra
 - Własny klaster Spark
 - Apache Mesos
 - Apache YARN
 - Administracja
 - Spark Core
 - Wstęp
 - Java vs Scala vs Python
 - Łączenie z klastrem
 - Rozproszone dane



- Operacje RDD
 - Transformacje
 - Akcje
- Współdzielone zmienne
- Uruchomienie i testowanie
- Spark Streaming
 - Wstęp
 - Zasada działania
 - Strumienie
 - Wejście
 - Transformacja
 - Wyjście
 - Uruchomienie i testowanie
- Inne związane technologie strumieniowe
 - Apache Flume
 - Amazon Kinesis
 - Akka
 - Apache Samza

