Tytuł szkolenia: Big Data with Apache Family (Hadoop, Spark)

Kod szkolenia: BIG-DATA

Wprowadzenie

Adresaci szkolenia

Adresatami szkolenia są programiści i analitycy biznesowi, których celem jest poznanie narzędzi Big Data. Zalecana jest znajomość na poziomie podstawowym języka Java lub Scala.

Dla grup zamkniętych: cena oraz program szkolenia są dostosowywane indywidualnie - w zależności od potrzeb.

Cel szkolenia

Głównym celem szkolenia jest nabycie praktycznych umiejętności z technologii Big Data oraz języka Scala. Uczestnicy nauczą się efektywnego użycia technologii (Apache Spark, Apache Kafka, Apache NiFi, Apache Druid oraz Apache Hadoop) powszechnie wykorzystywanych w projektowaniu platform przetwarzających duże ilości danych. Istotną częścią szkolenia są warsztaty.

Czas i forma szkolenia

- 35 godzin (5 dni x 7 godzin), w tym wykłady i warsztaty praktyczne.

Plan szkolenia

1. Przegląd rozwiązań Big Data z wykorzystaniem technologii Apache
2. Scala dla Big Data:
• Scala vs Java
• var vs val
• Case Class, Trait, Abstract Class, Tuple
• Lazy evaluation
• Interpolacja ciągów
• Dopasowanie wzoru (z osłonami)
• Obiekt towarzyszący
• Kolekcje i przekształcenia
• For comprehension, mapowania
• Obsługa błędów (Try / Either / Option)
• Option
• Implicits
3. Warsztaty
4. Spark
• RDD, DataFrame, Dataset
• Lazy evaluation
• Spark jobs
• Transformacje i akcje
• Spark vs Hadoop
• DataFrame vs DataSet API
5. Warsztaty Spark - jak wzbogacić swoje dane?
6. Warsztaty Apache NiFi - jak wygodnie, niezawodnie i efektywnie przesyłać dane bez konieczności pisania i utrzymywania kodu
7. Spark - poziom zaawansowany, część 1
• architektura (driver, worker, executor...)
• optymalizacja jobów
• deployment
• typowe błędy - key-skew, serializacja, OOM
• Spark internals - joins
8. Warsztaty Spark dla zaawansowanych, część 1
9. Spark - poziom zaawansowany, część 2
• Streaming
• UDF - DF vs DS
• broadcast, repartition, caching, execution plans, optymalizacja
• Spark Catalyst, plany wykonania
• dobre praktyki
10. Warszaty Spark, część 2
11. Hadoop - Hive queries, konfiguracja Hive, tabele zewnętrzne i wewnętrzne, metastore (właściwości tabeli, partycjonowanie), Spark on Hive
12. Apache Kafka
• przewodnik dla programistów (architektura, dobre praktyki, porady)
• Warsztaty z Apache Kafka - wdrożenie producera i konsumenta Kafki w Scali
13. Apache Druid
• nowoczesne rozwiązanie w Big Data
• analityka w czasie rzeczywistym
• architektura
• porównanie z rozwiązaniami opartymi na Hadoop
14. Podsumowanie