Szkolenie Apache Spark w implementacji Pythona (PySpark)

Opis

Szkolenie PySpark wprowadza do narzędzia Apache Spark zaimplementowanego dla języka programowania Python (PySpark). Uczestnik w czasie kursu zapozna się z architekturą i sposobem działania narzędzia, a przez ćwiczenia praktyczna nauczy się przetwarzania danych w oparciu o funkcje wbudowane i moduł Spark SQL. Zostaną również poruszone kwestie z zakresu Delta Lake i formatu delta.

Grupa docelowa: Analitycy danych, Inżynierowie Danych, Data Scientist, Developerzy

Środowisko pracy: maszyna wirtualna z systemem operacyjnym typu Linux, kontener Jupyter Notebook, Azure Databricks (dodatkowy koszt)

Wymagana wiedza: znajomość składni SQL, dobra znajomość języka Python

Czego Cię nauczymy

Architektura Apache Spark i Data Flow

Po przejściu tego szkolenia uczestnicy zyskają pełne zrozumienie architektury Apache Spark i sposobu przetwarzania danych w tym narzędziu. Obejmiemy kluczowe elementy, takie jak RDDs (Resilient Distributed Datasets) i operacje transformacyjne, pozwalając uczestnikom zobaczyć, jak dane przemieszczają się w obrębie klastra Spark. Praktyczne ćwiczenia pomogą w zrozumieniu, jak efektywnie przygotować środowisko pracy do pracy z PySpark.

Pierwszy program z użyciem PySpark

Uczestnicy nauczą się zakładania sesji PySpark, tworzenia DataFrame'ów i manipulowania nimi za pomocą wbudowanych funkcji. Będą w stanie wyświetlać dane, definiować schematy, a także pracować z kolekcjami danych. Poprzez stworzenie pierwszego programu, zyskają praktyczne umiejętności niezbędne do rozpoczęcia pracy z Apache Spark w języku Python.

Efektywne przetwarzanie danych

Uczestnicy nauczą się, jak odczytywać dane z plików CSV, dokonywać konwersji typów danych, sortować oraz filtrować dane. Przećwiczą techniki manipulacji danymi, co pozwoli im na efektywne selekcjonowanie informacji. Ponadto, będą w stanie pracować z różnymi formatami danych, co zwiększy ich elastyczność w obszarze przetwarzania danych.

Grupowanie, funkcje agregujące i łączenie zbiorów

W tym punkcie szkolenia uczestnicy zdobędą umiejętności grupowania danych, wykorzystywania funkcji agregujących i łączenia różnych zbiorów danych. Nauczą się, jak efektywnie analizować dane w kontekście bardziej zaawansowanych scenariuszy, co obejmuje również przegląd technik pracy z danymi w kontekście relacyjnym.

Program szkolenia

1. Wprowadzenie

- Architektura Apache Spark
- Data flow
- Przygotowanie środowiska pracy

2. Pierwszy program

- Utworzenie sesji
- Stworzenie DataFrame
- Wyświetlanie danych i schematu
- Tworzenie schematu danych
- Kolekcje danych

3. Selekcja danych i pliki csv

- Odczyt danych z pliku CSV
- Konwersja typu danych
- Sortowanie
- Filtrowanie danych

4. Grupowanie i zbiory

- Funkcje agregujące
- Grupowanie danych
- Łączenie zbiorów

5. UDF, SQL, Map

- Tworzenie funkcji użytkownika (UDF)
- Wykorzystywanie modułu Spark SQL
- Mapowanie danych/zbiorów danych

6. Format danych i optymalizacja

- Praca z plikami JSON
- Praca z plikami XML
- Praca z formatami: parquet, avro
- Cache i persistent w DataFrame

7. Delta Table

- Architektura Delta Lake
- Wprowadzenie do formatu delta
- Transakcje i operacje CRUD
- Optymalizacja i zarządzanie plikami

Czas trwania

3 dni | 24 godziny zajęć

Certyfikat

Uczestnicy szkolenia otrzymują imienne certyfikaty sygnowane przez Expose Sp. z o.o.

Cena szkolenia

2 990 PLN netto (VAT 23%) za osobę (szkolenie grupowe)

Cena szkolenia zawiera

- zapewnienie autorskich materiałów szkoleniowych dla uczestników szkolenia
- wystawienie certyfikatów po zakończonym szkoleniu
- rekomendacje dla uczestników szkolenia w zakresie dalszej pracy w obszarze szkolenia
- pakiet konsultacji z wykładowcą po zakończonym szkoleniu w razie jakichkolwiek niejasności przez okres 3 miesięcy
- całodzienny serwis kawowy oraz lunch