|  |  |
| --- | --- |
| **Laboratorium:** | **2** |
| **Temat:** | **Przetwarzanie równoległe z wykorzystaniem Apache Spark. Analiza wydajności przetwarzania równoległego w trybie lokalnym.** |
| **Imię i nazwisko:** |  |

**Zadania do wykonania:**

**1. Konfiguracja środowiska:**

* Wczytanie niezbędnych bibliotek
* Instalacja PySpark
* Inicjalizacja środowiska Apache Spark w trybie lokalnym

**2. Badanie wpływu liczby rdzeni na wydajność przetwarzania:**

* Utworzenie sesji Spark z różnymi konfiguracjami rdzeni (1, 2, 4 i wszystkie dostępne)
* Generowanie testowego zbioru danych (1 milion rekordów)
* Wykonanie operacji grupowania i zliczania
* Pomiar i porównanie czasów wykonania

**Wyniki:**

**Komentarz:**

**3. Porównanie wydajności Pandas vs Spark:**

* Implementacja monitorowania CPU w czasie rzeczywistym
* Wykonanie operacji grupowania na zbiorze 1 miliona rekordów
* Porównanie wydajności między Pandas a różnymi konfiguracjami Spark
* Analiza wykorzystania CPU dla różnych konfiguracji

**Wyniki:**

**Komentarz:**

**4. Kompleksowa analiza wydajności i zasobów:**

* Testowanie dla różnych wielkości danych (100k, 1M, 5M rekordów)
* Monitorowanie CPU i RAM w czasie rzeczywistym
* Wykonanie złożonych operacji na danych (grupowanie, agregacje, filtrowanie, sortowanie)
* Generowanie i analiza wykresów porównawczych

**Wyniki:**

**Komentarz:**