Analiza dużych zbiorów danych

November 20, 2023

1 Konspekt Projektu: Wprowadzenie do Biblioteki Dask

1.1 Wstęp

- Cel projektu: Zapoznanie się z biblioteką Dask, narzędziem do wykonywania obliczeń równoległych, podobnym do Apache Spark.
- Główny cel badawczy: Ocena i porównanie potencjalnego zysku w wydajności obliczeniowej po wykorzystaniu biblioteki Dask.
- Finalny produkt: Raport podsumowujący wyniki porównania.

1.2 Plan Projektu

1. **Przygotowanie Proof of Concept (PoC) na Mniejszej Skali**
   - Wykorzystanie lokalnego środowiska do testowania i eksperymentowania.
   - Zbiór danych: 8 milionów wierszy recenzji z serwisu Amazon.
   - Zadanie: Analiza sentymentalna recenzji.
2. **Przeniesienie Projektu na Środowisko Docelowe**
   - Implementacja projektu w środowisku chmurowym.
   - Skalowanie projektu do obsługi większych zbiorów danych.
   - Analiza wydajności i skalowalności.

1.3 Proof of Concept (PoC)

- **Przygotowanie Lokalnego Notebooka**
  - Tworzenie środowiska programistycznego.
  - Implementacja wstępnej analizy sentymentalnej z wykorzystaniem Dask.
  - Analiza wyników i porównanie z tradycyjnymi metodami obliczeniowymi.
- **Analiza i Wyciąganie Wniosków**
  - Ocena wydajności biblioteki Dask w kontekście lokalnego środowiska.
  - Zbieranie danych dotyczących czasu przetwarzania i zużycia zasobów.

1.4 Przeniesienie i Skalowanie

- **Migracja do Środowiska Chmurowego**
  - Wybór odpowiedniej platformy chmurowej.
  - Konfiguracja środowiska do pracy z większymi zbiorami danych.
- **Analiza Skalowalności i Wydajności**
  - Testy wydajnościowe w środowisku chmurowym.
  - Porównanie wyników z lokalnym środowiskiem.

1.5 Podsumowanie i Raport

- **Opracowanie Raportu**
  - Zestawienie wyników z obu etapów projektu.
  - Analiza potencjalnych zastosowań biblioteki Dask w przyszłości.
  - Wnioski końcowe dotyczące skuteczności i wydajności Dask.
- **Rekomendacje i Sugestie**
  - Propozycje dalszych badań i eksperymentów z wykorzystaniem Dask.
  - Możliwe scenariusze użycia w różnych dziedzinach danych.

Maciej Wieloch, Jakub Gałecki