Konspekt projektu

Przedmiot w ramach którego realizowany jest przedmiot: **Szkolenie techniczne 4**

Temat projektu*:* DataLab - system do przetwarzania dużego zbioru danych NYC Yellow Taxi

Czas trwania: Od 6 maja 2025 do 27 czerwca 2025

Data rozpoczęcia projektu: 06.05.2025

Data zakończenia projektu: 27.06.2025

Harmonogram:

Etap	Zakres prac	Termin realizacji
1. Analiza tematu	Analiza wymagań, danych NYC Taxi, zaplanowanie architektury	6–8 maja 2025
2. Projekt struktury systemu	Zaprojektowanie modułów: loader, walidacja, analiza, UI	9–10 maja 2025
3. Implementacja walidatorów	Utworzenie klas OOP walidujących dane	11–14 maja 2025
4. Przetwarzanie równoległe	Implementacja multiprocessing i pipeline	15–20 maja 2025
5. Wizualizacja wyników	Generowanie wykresów i raportów (matplotlib, seaborn)	21–23 maja 2025
6. Aplikacja webowa	Tworzenie interfejsu użytkownika w Streamlit	24–26 maja 2025
7. Testowanie	Pisanie testów jednostkowych, testy działania całego systemu	27–30 maja 2025
8. Profilowanie i logowanie	Pomiar wydajności CPU/RAM, obsługa logów	31 maja – 2 czerwca 2025
9. Dokumentacja projektu	Tworzenie dokumentacji technicznej i użytkowej	3–7 czerwca 2025
10. Prezentacja końcowa	Ostateczne testy, uruchomienie demo, przygotowanie prezentacji	8–10 czerwca 2025
11. Złożenie projektu	Finalne dopracowanie, wersjonowanie, oddanie projektu	11–27 czerwca 2025

Wymagania wstępne (czyli jaka wiedza, oprogramowanie i sprzęt):

- **Wiedza:** Znajomość języka Python, OOP, przetwarzania danych z Pandas, podstawy multiprocessing i systemów walidacji danych, podstawy testów jednostkowych, umiejętność tworzenia aplikacji Streamlit.
- **Oprogramowanie:** Python 3.11+, PyCharm, biblioteki: Pandas, PyArrow, Streamlit, matplotlib, seaborn, pytest, memory_profiler.

- **Sprzęt**: Komputer z min. 4-rdzeniowym CPU i 8 GB RAM, system Windows/Linux/macOS, przeglądarka internetowa.

Cele:

Celem projektu jest stworzenie modularnej aplikacji analizującej dane NYC Taxi, wykorzystującej równoległość obliczeniową i zapewniającej interaktywny interfejs użytkownika. Projekt umożliwia efektywne przetwarzanie dużych zbiorów danych, identyfikację anomalii oraz generowanie raportów i wizualizacji pomocnych przy analizie biznesowej. W szczególności, aplikacja pozwala użytkownikowi uruchomić pipeline analityczny bezpośrednio z poziomu przeglądarki.

Zakres projektu:

Projekt obejmuje: przetwarzanie danych z plików .parquet (chunkami), walidację danych przez system klas dziedziczących po BaseValidator, przetwarzanie równoległe z użyciem multiprocessing.Pool, generowanie wykresów (m.in. histogramy, boxploty, heatmapy), przygotowanie raportów tekstowych i logów, implementację UI w Streamlit, testy jednostkowe oraz profilowanie zasobów systemowych (RAM/CPU).

Rodzaj i wykonawca projektu:

Projekt indywidualny, wykonawca: Kacper Kulig (w69199)

Sprawozdanie:

Projekt został zrealizowany zgodnie z harmonogramem, a opracowane oprogramowanie działa w pełni lokalnie, spełniając wszystkie założone funkcjonalności. System umożliwia ładowanie danych, ich równoległe przetwarzanie, walidację, wizualizację oraz prezentację wyników w interaktywnym interfejsie użytkownika. Kod źródłowy został zaprojektowany w sposób modularny i testowalny, co ułatwia jego dalszy rozwój i utrzymanie. Szczegółowa dokumentacja techniczna oraz kompletna implementacja systemu zostały dołączone w formie załączników.

Ocena:

Podpis nauczyciela

dr inż. Leszek Puzio

Podpis studenta

Kacper Kulig w69199