Zadania projektowe – Big Data (Google Colab Edition)

Każdy student wybiera **jedno** z poniższych zagadnień projektowych.
Celem jest przygotowanie **notebooka w Google Colab**, który demonstruje praktyczne użycie wybranej technologii Big Data.
Do każdego zadania należy dołączyć **krótki opis kroków**, **komentarze w kodzie** oraz **podsumowanie wyników**.

---

1. ETL + Analiza danych publicznych (Spark w Colab)

Cel: pobranie danych z publicznego źródła, przetworzenie ich w Spark i wykonanie prostych analiz.

Wymagania:

- Pobierz dane z otwartego źródła (np. dane COVID-19, dane lotnicze, dane pogodowe, dane o populacji).
- Załaduj dane do Spark (PySpark w Colab).
- Wykonaj podstawowy proces **ETL**:
    - czyszczenie (usunięcie braków, konwersje typów, filtrowanie),
    - wybór lub agregacja istotnych kolumn.
- Oblicz kilka wskaźników (np. średnie, maksima, sumy, trendy).
- Zaprezentuj wyniki na prostym wykresie (np. matplotlib).

Wynik: notebook z kompletnym procesem ETL i analizą wyników.

---

2. Streaming demo (Spark Structured Streaming)

Cel: zademonstrowanie działania strumieniowego przetwarzania danych w Spark.

Wymagania:

- Skonfiguruj Structured Streaming w PySpark (Google Colab).
- Wybierz jedno ze źródeł:
    - syntetyczne źródło rate (generuje dane w czasie rzeczywistym),
    - lub katalog z plikami CSV, do którego będą dopisywane kolejne dane.
- Zdefiniuj **okno czasowe** i zliczaj zdarzenia per okno (np. co 10 sekund).
- Prezentuj wyniki z bufora (memory sink) w postaci tabeli lub wykresu trendu.

Wynik: działający notebook pokazujący przetwarzanie strumieniowe i wizualizację zmian w czasie.

---

3. NoSQL + BigQuery – porównanie środowisk

Cel: porównać dwa podejścia do analizy danych: **MongoDB (NoSQL)** i **Google BigQuery (SQL w chmurze)**.

Wymagania:

- Utwórz lub pobierz dane w formacie JSON (np. dane o transakcjach, użytkownikach, produktach).
- Załaduj dane do bazy **MongoDB Atlas** i wykonaj kilka agregacji (aggregate pipeline).
- Te same dane załaduj do **Google BigQuery** i wykonaj analogiczne zapytania SQL.
- Porównaj:
    - łatwość zapytań,
    - czas wykonania,
    - wygodę pracy,
    - ograniczenia i zalety obu rozwiązań.

Wynik: notebook z kodem dla obu środowisk oraz krótkim podsumowaniem porównawczym.

---

4. Uczenie maszynowe na dużych danych (Spark MLlib)

Cel: wytrenowanie modelu klasyfikacyjnego na dużym zbiorze danych z użyciem Spark MLlib.

Wymagania:

- Przygotuj duży zbiór danych (wygenerowany lub pobrany z zewnętrznego źródła).
- Użyj Spark MLlib do trenowania modelu klasyfikacyjnego (np. **regresja logistyczna**, **drzewo decyzyjne**).
- Zastosuj pipeline:
  VectorAssembler → (opcjonalnie) StandardScaler → Model.
- Oceń jakość modelu (metryki: **AUC**, **Accuracy**).
- Dodaj **wizualizację krzywej ROC** lub innego wskaźnika jakości.

Wynik: notebook z kompletnym procesem treningu, ewaluacji i prezentacji wyników.

Zasady ogólne

- Wszystkie projekty należy przygotować w **Google Colab** (można korzystać z Dysku Google).
- Notebook powinien być **samowystarczalny** — możliwy do uruchomienia od początku do końca.
- Kod powinien zawierać **komentarze i krótkie opisy kroków**.
- Ostateczny plik z notebookiem w formacie .pdf należy przesłać na adres email prowadzącego
- W opisie notebooka dodaj:
    - tytuł projektu,
    - autora,
    - źródło danych,
    - użyte biblioteki,
    - krótkie podsumowanie wyników.