Zadanie 1 – zapoznanie się z Spark UI

1. Jobs

* Pokazuje listę wszystkich jobs uruchomionych w Spark.
* Każde zadanie może składać się z wielu stages

Informacje:

* + Status zadania (Running, Succeeded, Failed)
  + Ilość wykonanych zadań, czas trwania
  + Link do szczegółów (etapy, rozkład tasków, logi)

2. Stages

* Pokazuje wszystkie etapy (stages) wykonane przez Sparka.
* Etapy są jednostkami przetwarzania między operacjami typu shuffle.

Informacje:

* Input/Output danych
* Rozkład tasków, czas ich wykonania
* Statystyki shuffle

3. Storage

* Pokazuje informacje o tym, które RDD lub DataFrame’y są zcache’owane lub zapisywane w pamięci.

Informacje:

* Rozmiar danych w pamięci
* Liczba partycji, które są zcache’owane
* Status i lokalizacja danych w klastrze

4. Executors

* Pokazuje szczegóły dotyczące executorów Spark.

Informacje:

* + Ilość pamięci użytej przez każdy executor
  + Ilość tasków wykonanych przez executor
  + Czas wykonania, ilość błędów
  + Shuffle Read/Write

5. SQL/Dataframe

* Pokazuje plan wykonania zapytań SQL i operacji na DataFrame’ach.

Informacje:

* + Fizyczny i logiczny plan wykonania
  + Czas wykonania poszczególnych operacji
  + Umożliwia zrozumienie jak Spark optymalizuje zapytania

Informacje o dystrybucji danych – gdzie znaleźć

* Stages – w kolumnach dotyczących Shuffle Read/Write, gdzie można sprawdzić czy dane były przemieszczane między partycjami.
* Executors – w statystykach Shuffle Read/Write, można zobaczyć, ile danych zostało przeniesionych przez konkretnego executora.
* SQL/DataFrame – analiza planu wykonania może pokazać, które operacje powodują przemieszczanie danych.