Zadanie 2

Przejdź przez Spark UI i opisz w kilku zdaniach co można znaleźć w każdym z elementów Spark UI.

* Jobs

Pokazuje podsumowanie wszystkich prac w aplikacji, stronę ze szczegółami dla każdej pracy, aktualnego użytkownika, czas od kiedy aplikacja zaczęła pracę, ilość prac w statusach aktywnych, zakończonym i nieudanym, oraz timeline wydarzeń.

* Stages

Pokazuje podsumowanie wszystkich etapów wszystkich prac aplikacji podzielone na statusy (aktywne, zakończone, pominięte, nieudane, trwające). Aktywne etapy można zatrzymać.

* Storage

Pokazuje trwające RDDsy i DataFramey, jeśli istnieją w aplikacji, podsumowanie poziomy, wielkości i partycje pamięci wszystkich RDDsów.

* Environment

Pokazuje wartości dla różnych środowisk i konfigurację zmiennych, włącznie z JVM, Spark i własnościami systemu.

* Executors

Pokazuje podsumowanie informacji o wykonawcach stworzonych dla tej aplikacji włącznie z użyciem pamięci i dysku oraz informację o wydajności.

* SQL/DataFrame

Pokazuje informacje o długościach czasu, pracach oraz fizycznych i logicznych planach dla zapytań, które pokazują jak Spark analizuje, optymalizuje i wykonuje zapytanie.

* JDBC/ODBC Server

Dostępny, gdy Spark używa rozproszony engine SQL. Pokazuje informacje ogólne o serwerze, jak czas rozpoczęcia i trwania, informacje o aktywnych i zakończonych sesjach oraz statystyki SQL wysłanych operacji.

* Structured Streaming

Dostępny, gdy używamy prac Structured Streaming w trybie mikro-serii. Pokazuje podstawowe statystyki działających i zakończonych zapytań, najnowsze wyjątki i nieudane zapytania.

Zadanie 3

Do jednej z Dataframe dołóż transformacje groupBy i porównaj, jak wygląda plan wykonania.

Do plany fizycznego zostają dodane 3 kroki: dwa razy HashAggregate oraz Exchange hashpartitioning