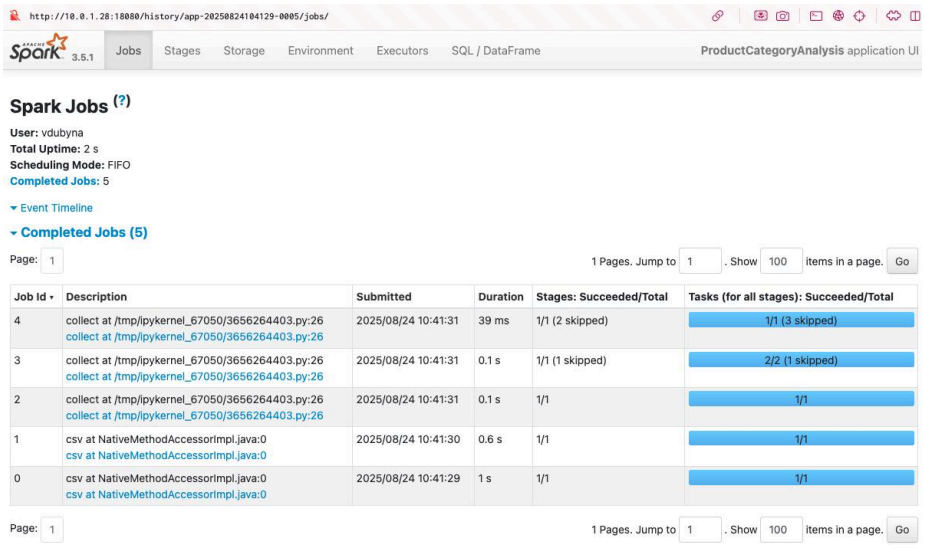
Домашнє завдання до теми «Apache Spark. Оптимізація та SparkUI»

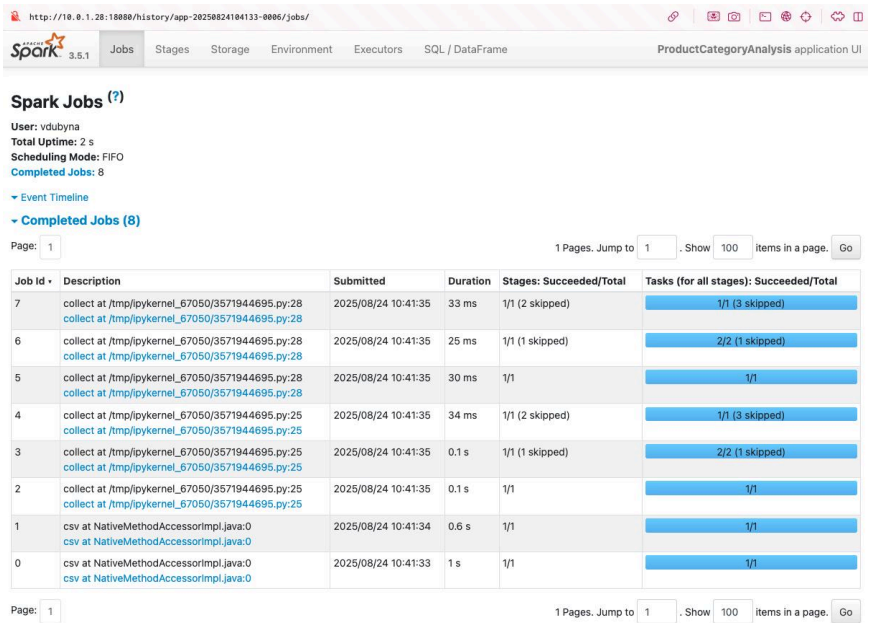
Частина 1

Spark виконує один action → одне обчислення всієї лінійки (pipeline)
Проте через оптимізацію Spark ділить DAG на 5 Jobs



Частина 2

Spark перераховує pipeline двічі: один раз для першого collect(), другий — для другого.
Тобто Spark повторно виконує всі обчислення з нуля, бо дані не зберігались



Частина 3

cache() каже Spark: “збережи результат цієї трансформації в пам’яті”
Перший .collect() → виконує всі трансформації та кешує результат
Другий .collect() вже не перераховує DAG → просто бере з пам’яті