A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Додавання nuek\_processed.collect() створює 3 додаткові Job, бо collect() запускає всі попередні ледачі перетворення, які ще не були виконані. Без кешування Spark перераховує весь DAG, виконуючи репартицію, фільтрацію та групування заново.

A screenshot of a computer

Description automatically generated

При використанні cache() зменшилась кількість Job, оскільки Spark зберігає проміжний результат у пам’яті. Це дозволяє повторно використовувати вже обчислені дані для наступних дій, уникаючи повторного виконання попередніх перетворень (репартиції, фільтрації, групування). Таким чином, виклик collect() після кешування не запускає всі попередні операції, зменшуючи кількість Job.