A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Перший collect() змушує Spark виконати всю ланцюжок трансформацій, включаючи groupBy(), що потребує значного обчислення та шифрування (shuffle).
* Другий collect() викликає додаткову операцію фільтрації (where("count > 2")), що також потребує виконання.
* Таким чином, кожен виклик collect() генерує окрему Job, що збільшує загальну кількість виконуваних Jobs.

A screenshot of a computer

Description automatically generated

При використанні cache(), Spark зберігає результати обчислень у пам'яті після першого виконання, щоб не обчислювати їх знову при наступних операціях.

* Без cache(): Кожен виклик collect() або інших action призводить до виконання Jobs для кожної трансформації.
* З cache(): Після першого обчислення дані зберігаються в пам'яті, і наступні операції працюють з кешованими даними, не запускаючи нові Jobs.

Тому, використовуючи cache(), зменшується кількість Jobs, оскільки Spark повторно не виконує обчислення для вже кешованих даних.