# Пояснення до Д34

## Частина 1

| № | Крок | Тип | Кількість Job | Опис |
|---|---|---|---|---|
| 1 | nuek_df = spark.read \<br>.option("header", "true") \<br>.option("inferSchema", "true") \<br>.csv('./nuek-vuh3.csv') | - | 1 Job | Scan Text |
| 2 | nuek_df = spark.read \<br>.option("header", "true") \<br>.option("inferSchema", "true") \<br>.csv('./nuek-vuh3.csv') | - | 1 Job | Deserialization |
| 3 | nuek_repart = nuek_df.repartition(2) | трансформація | No jobs наразі, але потенційно 1 Job | Потребує Shuffle |
| 4 | nuek_processed = nuek_repart \<br>.where("final_priority < 3") \<br>.select("unit_id", "final_priority") \<br>.groupBy("unit_id") \<br>.count() | трансформація | No jobs наразі, але потенційно 1 Job | Потребує Shuffle |
| 5 | nuek_processed = nuek_processed.where("count>2") | трансформація | No jobs | Немає Shuffle |
| 6 | nuek_processed.collect() | дія | 1 Job | Виклик попередніх трансформацій, які потребували Shuffle |
| | | **Загально** | **5 Jobs** | |

## Частина 2

| № | Крок | Тип | Кількість Job | Опис |
|---|---|---|---|---|
| 1 | nuek_df = spark.read<br>\.option("header", "true")<br>\.option("inferSchema", "true")<br>\.csv('./nuek-vuh3.csv') | - | 1 Job | Scan Text |
| 2 | nuek_df = spark.read<br>\.option("header", "true")<br>\.option("inferSchema", "true")<br>\.csv('./nuek-vuh3.csv') | - | 1 Job | Deserialization |

| | | | | |
|---|---|---|---|---|
| 3 | nuek_repart = nuek_df.repartition(2) | трансформація | No jobs наразі, але потенційно 1 Job | Потребує Shuffle |
| 4 | nuek_processed = nuek_repart \.where("final_priority < 3") \.select("unit_id", "final_priority") \.groupBy("unit_id") \.count() | трансформація | No jobs наразі, але потенційно 1 Job | Потребує Shuffle |
| 5 | nuek_processed.collect() | дія | 1 Job + виклик попередніх трансформацій | Виклик попередніх трансформацій, які потребували Shuffle |
| 6 | nuek_processed = nuek_processed.where("count>2") | трансформація | No jobs | Немає Shuffle |
| 7 | nuek_processed.collect() | дія | 1 Job + виклик попередніх дій і трансформацій | Виклик попередніх дій і трансформацій, які потребували Shuffle |
| | | **Загально** | **8 Jobs** | |

**Пояснення**: Повторний виклик collect() змусив нас заново виконати 2 трансформації, які вона тригерила, тому загальна кількість Jobs зросла на 3 (1 додатковий collect() + 2 трансформації).

## Частина 3

| № | Крок | Тип | Кількість Job | Опис |
|---|---|---|---|---|
| 1 | nuek_df = spark.read \.option("header", "true") \.option("inferSchema", "true") \.csv('./nuek-vuh3.csv') | - | 1 Job | Scan Text |
| 2 | nuek_df = spark.read \.option("header", "true") \.option("inferSchema", "true") \.csv('./nuek-vuh3.csv') | - | 1 Job | Deserialization |
| 3 | nuek_repart = nuek_df.repartition(2) | трансформація | No jobs наразі, але потенційно 1 Job | Потребує Shuffle |
| 4 | nuek_processed = nuek_repart \.where("final_priority < 3") | трансформація | No jobs наразі, але потенційно | Потребує Shuffle |

| | | | | |
|---|---|---|---|---|
| | \.select("unit_id", "final_priority")<br>\.groupBy("unit_id") \.count()<br><br>**\.cache() # Додано функцію cache** | | 1 Job | |
| 5 | nuek_processed.collect() | дія | 1 Job + виклик попередніх трансформацій | Виклик попередніх трансформацій, які потребували Shuffle |
| 6 | nuek_processed = nuek_processed.where("count>2") | трансформація | No jobs | Немає Shuffle |
| 7 | nuek_processed.collect() | дія | 1 Job + виклик попередніх дій і трансформацій | Виклик попередніх дій і трансформацій, які потребували Shuffle |
| | | **Загально** | **7 Jobs** | |

**Пояснення**: Завдяки функції кешування у nuek_processed другий виклик collect() не спровокував повторний обрахунок nuek_processed, тому кількість Jobs зменшилась на 1.