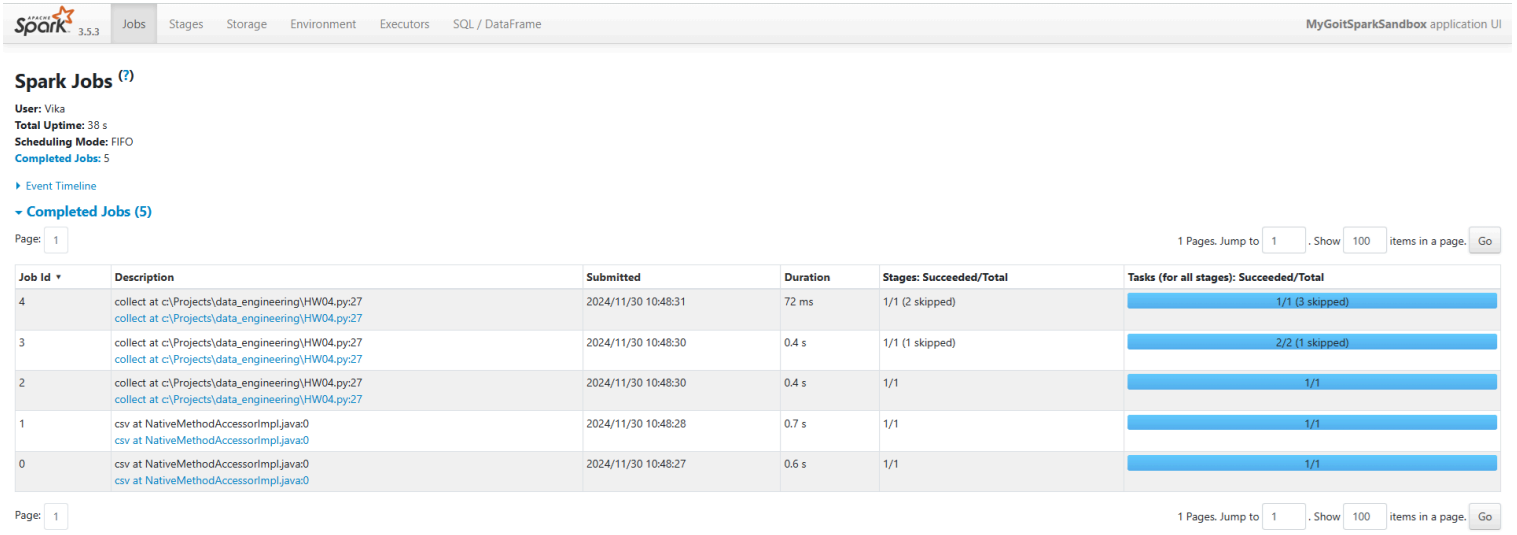
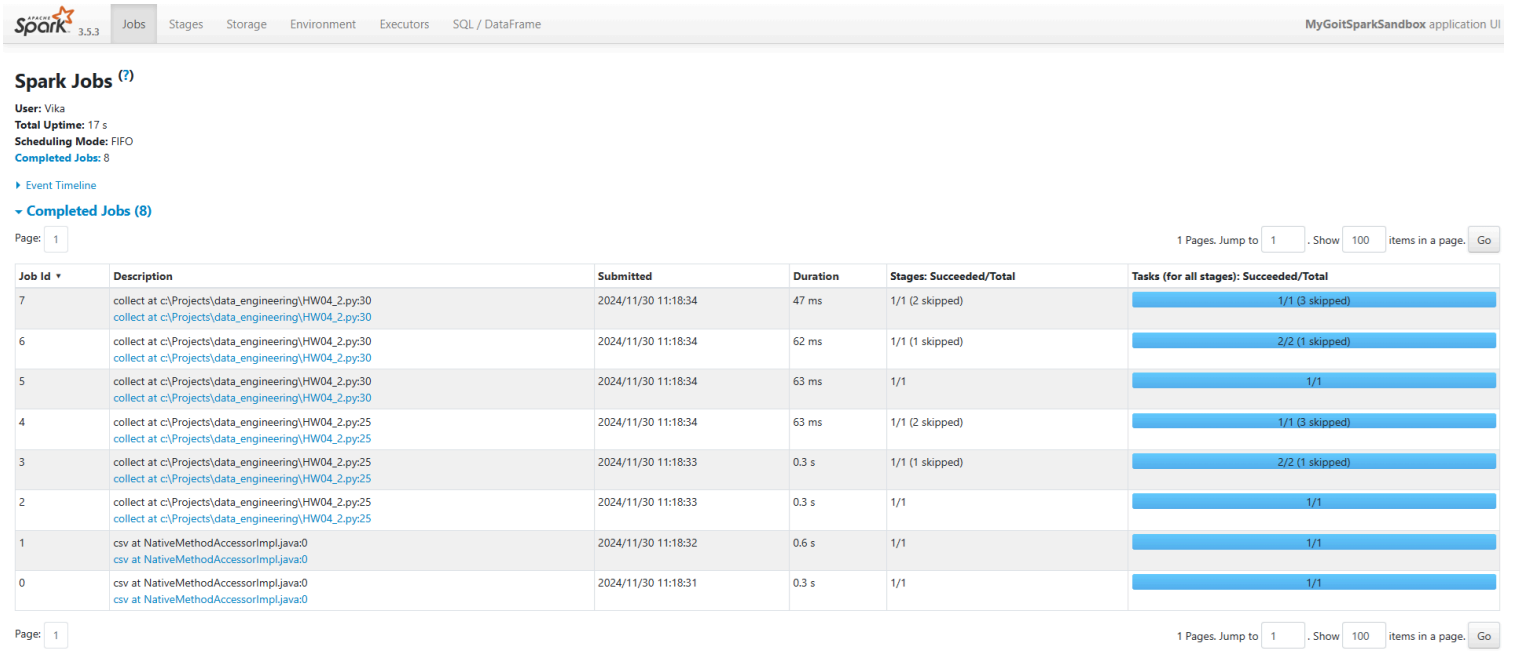
1.



Перші дві Jobs стосуються зчитування та серіалізації даних. Інші три для виконання функції активації collect перед якою має відбутися shuffle, а він в свою чергу складається з запису та зчитування.

2.



З додаванням ще однієї функції активації collect додалось ще три Jobs тому що для обох функцій повинен відбутися shuffle, тому і виходить по 3 Jobs на кожну.

3.

Spark Jobs (?)

User: Vika
Total Uptime: 20 s
Scheduling Mode: FIFO
Completed Jobs: 7

▸ Event Timeline

▾ **Completed Jobs (7)**

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	collect at c:\Projects\data_engineering\HW04_3.py:31 collect at c:\Projects\data_engineering\HW04_3.py:31	2024/11/30 11:37:31	65 ms	1/1 (2 skipped)	2/2 (3 skipped)
5	collect at c:\Projects\data_engineering\HW04_3.py:26 collect at c:\Projects\data_engineering\HW04_3.py:26	2024/11/30 11:37:31	0.1 s	1/1 (2 skipped)	2/2 (3 skipped)
4	collect at c:\Projects\data_engineering\HW04_3.py:26 collect at c:\Projects\data_engineering\HW04_3.py:26	2024/11/30 11:37:30	0.3 s	1/1 (2 skipped)	2/2 (3 skipped)
3	collect at c:\Projects\data_engineering\HW04_3.py:26 collect at c:\Projects\data_engineering\HW04_3.py:26	2024/11/30 11:37:30	0.3 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Projects\data_engineering\HW04_3.py:26 collect at c:\Projects\data_engineering\HW04_3.py:26	2024/11/30 11:37:29	0.3 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 11:37:28	0.6 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 11:37:28	0.3 s	1/1	1/1

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

З додаванням кешування замість 8 Jobs стало 7, бо після того як ми закешували проміжний результат Spark виконує подальші операції з ним in memory, тобто в оперативній пам’яті без shuffle.