

1.

spark

3.5.3

Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

MyGoItSparkSandbox application UI

Spark Jobs (?)

User: Vika

Total Uptime: 38 s

Scheduling Mode: FIFO

Completed Jobs: 5

Event Timeline

Completed Jobs (5)

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
4	collect at c:\Projects\data_engineering\HW04.py:27 collect at c:\Projects\data_engineering\HW04.py:27	2024/11/30 10:48:31	72 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at c:\Projects\data_engineering\HW04.py:27 collect at c:\Projects\data_engineering\HW04.py:27	2024/11/30 10:48:30	0.4 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Projects\data_engineering\HW04.py:27 collect at c:\Projects\data_engineering\HW04.py:27	2024/11/30 10:48:30	0.4 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 10:48:28	0.7 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 10:48:27	0.6 s	1/1	1/1

Page: 1

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

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.

spark

3.5.3

Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

MyGoItSparkSandbox application UI

Spark Jobs (?)

User: Vika

Total Uptime: 17 s

Scheduling Mode: FIFO

Completed Jobs: 8

Event Timeline

Completed Jobs (8)

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
7	collect at c:\Projects\data_engineering\HW04_2.py:30 collect at c:\Projects\data_engineering\HW04_2.py:30	2024/11/30 11:18:34	47 ms	1/1 (2 skipped)	1/1 (3 skipped)
6	collect at c:\Projects\data_engineering\HW04_2.py:30 collect at c:\Projects\data_engineering\HW04_2.py:30	2024/11/30 11:18:34	62 ms	1/1 (1 skipped)	2/2 (1 skipped)
5	collect at c:\Projects\data_engineering\HW04_2.py:30 collect at c:\Projects\data_engineering\HW04_2.py:30	2024/11/30 11:18:34	63 ms	1/1	1/1
4	collect at c:\Projects\data_engineering\HW04_2.py:25 collect at c:\Projects\data_engineering\HW04_2.py:25	2024/11/30 11:18:34	63 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at c:\Projects\data_engineering\HW04_2.py:25 collect at c:\Projects\data_engineering\HW04_2.py:25	2024/11/30 11:18:33	0.3 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Projects\data_engineering\HW04_2.py:25 collect at c:\Projects\data_engineering\HW04_2.py:25	2024/11/30 11:18:33	0.3 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 11:18:32	0.6 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/30 11:18:31	0.3 s	1/1	1/1

Page: 1

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

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 <sup>(?)</sup>

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.