

Частина 1:

APACHE
Spark
3.5.3

JobsStagesStorageEnvironmentExecutorsSQL / DataFrame

MyGoItSparkSandbox application UI

Spark Jobs (?)

User: vdubr
Total Uptime: 60 s
Scheduling Mode: FIFO
Completed Jobs: 5

▼ Event Timeline
☐ Enable zooming

The event timeline chart displays the execution progress of Spark jobs. The x-axis represents time from 19:56:42 to 19:56:48. The y-axis lists events. A blue bar at 19:56:44 indicates 'Executor driver added'. Subsequent events include 'csv at NativeMethodAccessorImpl.java:0' and 'collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27'.

▼ Completed Jobs (5)

Page: 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:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27	2024/11/17 19:56:48	45 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27	2024/11/17 19:56:47	0,2 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part1.py:27	2024/11/17 19:56:47	0,2 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 19:56:46	0,3 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 19:56:46	0,2 s	1/1	1/1

Page: 1 . Show 100 items in a page. Go

Частина 2

Spark

3.5.3

JobsStagesStorageEnvironmentExecutorsSQL / DataFrame

MyGoitSparkSandbox application UI

Spark Jobs ^(?)

User: vdubr
Total Uptime: 12 s
Scheduling Mode: FIFO
Completed Jobs: 8

Event Timeline

Enable zooming

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30	2024/11/17 20:02:18	28 ms	1/1 (2 skipped)	1/1 (3 skipped)
6	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30	2024/11/17 20:02:18	30 ms	1/1 (1 skipped)	2/2 (1 skipped)
5	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:30	2024/11/17 20:02:18	34 ms	1/1	1/1
4	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25	2024/11/17 20:02:18	33 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25	2024/11/17 20:02:17	0,2 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part2.py:25	2024/11/17 20:02:17	0,2 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 20:02:16	0,3 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 20:02:16	0,2 s	1/1	1/1

Додавання `nuek_processed.collect()` на 25му рядку запускає обчислення попередніх дій у Spark (3 джоба). А другий виклик `nuek_processed.collect()` на 30 рядку знову обчислює всі дані (5 джобів, як у частині 1). $3 + 5 = 8$.

Частина 3

3.5.3

[Jobs](#)
[Stages](#)
[Storage](#)
[Environment](#)
[Executors](#)
[SQL / DataFrame](#)

MyGoitSparkSandbox application

Spark Jobs (?)

User: vdubr
Total Uptime: 9 s
Scheduling Mode: FIFO
Completed Jobs: 7

[Event Timeline](#)

☐ Enable zooming

The event timeline shows a single event at approximately 20:06:51:00 labeled "Executor driver added". The x-axis represents time from 20:06:49 to 20:06:55. The y-axis lists categories: Executors (Added, Removed) and Jobs (Succeeded, Failed, Running). Other job events are visible as small blue bars further along the timeline.

Completed Jobs (7)

Page: 1

Job Id *	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:31 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:31	2024/11/17 20:06:55	34 ms	1/1 (2 skipped)	2/2 (3 skipped)
5	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26	2024/11/17 20:06:55	97 ms	1/1 (2 skipped)	2/2 (3 skipped)
4	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26	2024/11/17 20:06:55	0,2 s	1/1 (2 skipped)	2/2 (3 skipped)
3	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26	2024/11/17 20:06:54	0,2 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26 collect at c:\Users\vdubr\OneDrive\Документы\GoIT Обучение\Data Engineering\goit-de-hw-04\part3.py:26	2024/11/17 20:06:54	0,2 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 20:06:53	0,3 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/11/17 20:06:53	0,2 s	1/1	1/1

Page: 1

До першого `newk_processed_cached.collect()` на 26 рядку ми виконали 3 джоба. Плюс 1 джоб на кешування.

Другий виклик `nuek_processed.collect()` виконує 1 джоб на отримання даних з кешу, виконує джобом код: `nuek_processed = nuek_processed_cached.where("count>2")`, та збирає дані на драйвері ще 1 джобом.

Тобто: $(3 + 1) + (1 + 1 + 1) = 7$ джобів.