

Домашнє завдання до теми «Apache Spark. Оптимізація та SparkUI»

Частина 1

Spark виконує один action → одне обчислення всієї лінійки (pipeline)
Проте через оптимізацію Spark ділить DAG на 5 Jobs

http://10.0.1.28:18080/history/app-20250824104129-0005/jobs/

Spark 3.5.1

Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

ProductCategoryAnalysis application UI

Spark Jobs (?)

User: vdubyna

Total Uptime: 2 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 /tmp/pykernel_67050/3656264403.py:26 collect at /tmp/pykernel_67050/3656264403.py:26	2025/08/24 10:41:31	39 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at /tmp/pykernel_67050/3656264403.py:26 collect at /tmp/pykernel_67050/3656264403.py:26	2025/08/24 10:41:31	0.1 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at /tmp/pykernel_67050/3656264403.py:26 collect at /tmp/pykernel_67050/3656264403.py:26	2025/08/24 10:41:31	0.1 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:30	0.6 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:29	1 s	1/1	1/1

Page: 1

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

Частина 2

Spark перераховує pipeline двічі: один раз для першого collect(), другий — для другого.
Тобто Spark повторно виконує всі обчислення з нуля, бо дані не зберігались

http://10.0.1.28:18080/history/app-20250824104133-0006/jobs/

Spark 3.5.1

Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

ProductCategoryAnalysis application UI

Spark Jobs (?)

User: vdubyna

Total Uptime: 2 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 /tmp/pykernel_67050/3571944695.py:28 collect at /tmp/pykernel_67050/3571944695.py:28	2025/08/24 10:41:35	33 ms	1/1 (2 skipped)	1/1 (3 skipped)
6	collect at /tmp/pykernel_67050/3571944695.py:28 collect at /tmp/pykernel_67050/3571944695.py:28	2025/08/24 10:41:35	25 ms	1/1 (1 skipped)	2/2 (1 skipped)
5	collect at /tmp/pykernel_67050/3571944695.py:28 collect at /tmp/pykernel_67050/3571944695.py:28	2025/08/24 10:41:35	30 ms	1/1	1/1
4	collect at /tmp/pykernel_67050/3571944695.py:25 collect at /tmp/pykernel_67050/3571944695.py:25	2025/08/24 10:41:35	34 ms	1/1 (2 skipped)	1/1 (3 skipped)
3	collect at /tmp/pykernel_67050/3571944695.py:25 collect at /tmp/pykernel_67050/3571944695.py:25	2025/08/24 10:41:35	0.1 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at /tmp/pykernel_67050/3571944695.py:25 collect at /tmp/pykernel_67050/3571944695.py:25	2025/08/24 10:41:35	0.1 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:34	0.6 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:33	1 s	1/1	1/1

Page: 1

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

Частина 3

cache() каже Spark: “збережи результат цієї трансформації в пам’яті”

Перший .collect() → виконує всі трансформації та кешує результат

Другий .collect() вже не перераховує DAG → просто бере з пам’яті

http://10.0.1.28:18080/history/app-20250824104136-0007/jobs/

Spark 3.5.1 Jobs Stages Storage Environment Executors SQL / DataFrame ProductCategoryAnalysis application UI

Spark Jobs (?)

User: vdubyna
Total Uptime: 2 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 /tmp/ipykernel_67050/3952951936.py:29 collect at /tmp/ipykernel_67050/3952951936.py:29	2025/08/24 10:41:39	25 ms	1/1 (2 skipped)	2/2 (3 skipped)
5	collect at /tmp/ipykernel_67050/3952951936.py:26 collect at /tmp/ipykernel_67050/3952951936.py:26	2025/08/24 10:41:39	37 ms	1/1 (2 skipped)	2/2 (3 skipped)
4	collect at /tmp/ipykernel_67050/3952951936.py:26 collect at /tmp/ipykernel_67050/3952951936.py:26	2025/08/24 10:41:39	64 ms	1/1 (2 skipped)	2/2 (3 skipped)
3	collect at /tmp/ipykernel_67050/3952951936.py:26 collect at /tmp/ipykernel_67050/3952951936.py:26	2025/08/24 10:41:39	0.1 s	1/1 (1 skipped)	2/2 (1 skipped)
2	collect at /tmp/ipykernel_67050/3952951936.py:26 collect at /tmp/ipykernel_67050/3952951936.py:26	2025/08/24 10:41:38	0.1 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:38	0.6 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2025/08/24 10:41:36	1 s	1/1	1/1

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