

# Spark Day 2 Assignment

## 1. Customer Data

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL CustomerCityCount application UI

**Spark Jobs [1]**  
User: racit  
Total Uptime: 1.2 h  
Scheduling Mode: FIFO  
Completed Jobs: 10

- Event Timeline  
Enable zooming

Executors  
All Active Removed  
Jobs  
Sucessed Failed Running

Executor driver added at 2 December 18:52 (Job 0) take at customerData.scala:40 (Job 0) show at customerData.scala:50 (Job 2) csv at customerData.scala:57 (Job 4) json at customerData.scala:61 (Job 6) parquet at customerData.scala:65 (Job 8)

parquet at customerData.scala:65 parquet at customerData.scala:65

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

**Completed Jobs (10)**

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
9	parquet at customerData.scala:65 parquet at customerData.scala:65	2025/12/02 18:53:40	1 s	1/1 (1 skipped)	1/1 (50 skipped)
8	parquet at customerData.scala:65 parquet at customerData.scala:65	2025/12/02 18:53:33	8 s	1/1	50/50
7	json at customerData.scala:61 json at customerData.scala:61	2025/12/02 18:53:32	0.2 s	1/1 (1 skipped)	1/1 (50 skipped)
6	json at customerData.scala:61 json at customerData.scala:61	2025/12/02 18:53:26	7 s	1/1	50/50
5	csv at customerData.scala:57 csv at customerData.scala:57	2025/12/02 18:53:25	0.3 s	1/1 (1 skipped)	1/1 (50 skipped)
4	csv at customerData.scala:57 csv at customerData.scala:57	2025/12/02 18:53:18	7 s	1/1	50/50
3	show at customerData.scala:50 show at customerData.scala:50	2025/12/02 18:53:18	76 ms	1/1 (1 skipped)	1/1 (50 skipped)
2	show at customerData.scala:50 show at customerData.scala:50	2025/12/02 18:53:10	8 s	1/1	50/50
1	take at customerData.scala:40 take at customerData.scala:40	2025/12/02 18:53:09	24 ms	1/1 (1 skipped)	4/4 (50 skipped)
0	take at customerData.scala:40 take at customerData.scala:40	2025/12/02 18:52:57	12 s	2/2	51/51

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

### Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL CustomerCityCount application UI

**Stages for All Jobs**  
Completed Stages: 11  
Skipped Stages: 5

- Completed Stages (11)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
15	parquet at customerData.scala:65	+details 2025/12/02 18:53:40	1 s	1/1	1127.0 B	159.8 kB		
13	parquet at customerData.scala:65	+details 2025/12/02 18:53:33	8 s	50/50			159.8 kB	
12	json at customerData.scala:61	+details 2025/12/02 18:53:32	0.1 s	1/1	1668.0 B	159.8 kB		
10	json at customerData.scala:61	+details 2025/12/02 18:53:26	7 s	50/50			159.8 kB	
9	csv at customerData.scala:57	+details 2025/12/02 18:53:25	0.2 s	1/1	716.0 B	159.8 kB		
7	csv at customerData.scala:57	+details 2025/12/02 18:53:18	7 s	50/50			159.8 kB	
6	show at customerData.scala:50	+details 2025/12/02 18:53:18	58 ms	1/1		159.8 kB		
4	show at customerData.scala:50	+details 2025/12/02 18:53:10	8 s	50/50			159.8 kB	
3	take at customerData.scala:40	+details 2025/12/02 18:53:09	19 ms	4/4		13.3 kB		
1	take at customerData.scala:40	+details 2025/12/02 18:53:09	58 ms	1/1		2.6 kB		
0	map at customerData.scala:36	+details 2025/12/02 18:52:57	12 s	50/50			108.7 kB	

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

- Skipped Stages (5)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
14	parquet at customerData.scala:65	+details Unknown	Unknown	0/50				
11	json at customerData.scala:61	+details Unknown	Unknown	0/50				
8	csv at customerData.scala:57	+details Unknown	Unknown	0/50				
5	show at customerData.scala:50	+details Unknown	Unknown	0/50				
2	map at customerData.scala:36	+details Unknown	Unknown	0/50				

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

## Executors

CustomerCityCount application UI							
Stages for All Jobs							
Completed Stages: 11 Skipped Stages: 5 - Completed Stages (11)							
Page: 1							
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read
15	parquet at customerData.scala:65	+details 2025/12/02 18:53:40	1 s	1/1	1127.0 B	159.8 kB	
13	parquet at customerData.scala:65	+details 2025/12/02 18:53:33	8 s	50/50			159.8 kB
12	json at customerData.scala:61	+details 2025/12/02 18:53:32	0.1 s	1/1	1668.0 B	159.8 kB	
10	json at customerData.scala:61	+details 2025/12/02 18:53:26	7 s	50/50			159.8 kB
9	csv at customerData.scala:57	+details 2025/12/02 18:53:25	0.2 s	1/1	715.0 B	159.8 kB	
7	csv at customerData.scala:57	+details 2025/12/02 18:53:18	7 s	50/50			159.8 kB
6	show at customerData.scala:50	+details 2025/12/02 18:53:18	58 ms	1/1			159.8 kB
4	show at customerData.scala:50	+details 2025/12/02 18:53:10	8 s	50/50			159.8 kB
3	take at customerData.scala:40	+details 2025/12/02 18:53:09	19 ms	4/4		13.3 kB	
1	take at customerData.scala:40	+details 2025/12/02 18:53:09	58 ms	1/1		2.6 kB	
0	map at customerData.scala:36	+details 2025/12/02 18:52:57	12 s	50/50			108.7 kB

  

CustomerCityCount application UI							
Skipped Stages (5)							
Page: 1							
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Write
14	parquet at customerData.scala:65	+details Unknown	Unknown	0/50			
11	json at customerData.scala:61	+details Unknown	Unknown	0/50			
8	csv at customerData.scala:57	+details Unknown	Unknown	0/50			
5	show at customerData.scala:50	+details Unknown	Unknown	0/50			
2	map at customerData.scala:36	+details Unknown	Unknown	0/50			

Observation	CSV	JSON	Parquet
<b>Write Speed</b>	<b>Fastest (very low overhead)</b>	Medium	<b>Slowest (metadata + encoding overhead)</b>
<b>Output Size (in this pipeline)</b>	<b>Smallest (raw text, no metadata)</b>	Largest (verbose text, repeats field names)	Medium (metadata overhead bigger than actual data since only ~50 rows)
<b>Operations Used</b>	-	-	-
<b>Narrow Ops</b>	map (customer generation), toDF, final part of take/show		
<b>Broad Ops</b>	reduceByKey, groupBy("city").count()(shuffle)		

## 2. Sales Data

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SalesPipeline10M application UI

**Spark Jobs (1)**

User: racit  
Total Uptime: 1.0 min  
Scheduling Mode: FIFO  
Completed Jobs: 8

- Event Timeline  Enable zooming

Executors	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49																				
All	Executor driver added																																							
Jobs	Succeeded	take at SalesPipeline.scala:41 (Job 0) [t]																			take at SalesPipeline.scala:51 (Job 2)	show at SalesPipeline.scala:61 (Job 4) [t] parquet at SalesPipeline.scala: parquet at SalesPipeline.scala:68 (Job 7)																		
	Failed																																							
	Running																																							
		2 December 20:21																																						

- 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	parquet at SalesPipeline.scala:68 parquet at SalesPipeline.scala:68	2025/12/02 20:21:45	4 s	1/1 (1 skipped)	1/1 (60 skipped)
6	parquet at SalesPipeline.scala:68 parquet at SalesPipeline.scala:68	2025/12/02 20:21:43	2 s	1/1	50/50
5	show at SalesPipeline.scala:61 show at SalesPipeline.scala:61	2025/12/02 20:21:43	87 ms	1/1 (1 skipped)	1/1 (60 skipped)
4	show at SalesPipeline.scala:61 show at SalesPipeline.scala:61	2025/12/02 20:21:40	3 s	1/1	50/50
3	take at SalesPipeline.scala:51 take at SalesPipeline.scala:51	2025/12/02 20:21:40	16 ms	1/1 (1 skipped)	3/2 (60 skipped)
2	take at SalesPipeline.scala:51 take at SalesPipeline.scala:51	2025/12/02 20:21:37	3 s	2/2	51/51
1	take at SalesPipeline.scala:41 take at SalesPipeline.scala:41	2025/12/02 20:21:37	0.2 s	1/1 (1 skipped)	3/2 (60 skipped)
0	take at SalesPipeline.scala:41 take at SalesPipeline.scala:41	2025/12/02 20:21:32	5 s	2/2	51/51

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

### Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SalesPipeline10M application UI

**Stages for All Jobs**

Completed Stages: 10  
Skipped Stages: 4

- Completed Stages (10)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	parquet at SalesPipeline.scala:68	+details 2025/12/02 20:21:45	4 s	1/1	1970.0 B	310.0 Kib		
11	parquet at SalesPipeline.scala:68	+details 2025/12/02 20:21:43	2 s	50/50			310.0 Kib	
10	show at SalesPipeline.scala:61	+details 2025/12/02 20:21:43	67 ms	1/1		310.0 Kib		
8	show at SalesPipeline.scala:61	+details 2025/12/02 20:21:40	3 s	50/50			310.0 Kib	
7	take at SalesPipeline.scala:51	+details 2025/12/02 20:21:40	13 ms	3/3		9.7 Kib		
5	take at SalesPipeline.scala:51	+details 2025/12/02 20:21:40	18 ms	1/1		3.8 Kib		
4	map at SalesPipeline.scala:24	+details 2025/12/02 20:21:37	3 s	50/50			186.3 Kib	
3	take at SalesPipeline.scala:41	+details 2025/12/02 20:21:37	0.2 s	3/3		3.8 MiB		
1	take at SalesPipeline.scala:41	+details 2025/12/02 20:21:36	0.2 s	1/1		2.1 MiB		
0	map at SalesPipeline.scala:24	+details 2025/12/02 20:21:32	4 s	50/50			102.8 MiB	

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

- Skipped Stages (4)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
12	parquet at SalesPipeline.scala:68	+details Unknown	Unknown	0/50				
9	show at SalesPipeline.scala:61	+details Unknown	Unknown	0/50				
6	map at SalesPipeline.scala:24	+details Unknown	Unknown	0/50				
2	map at SalesPipeline.scala:24	+details Unknown	Unknown	0/50				

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

## Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SalesPipeline10M application UI

**Executors**

Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	34.7 kB / 2.2 GB	0.0 B	8	0	0	210	210	1.6 min (2 s)	0.0 B	6.5 MB	103.5 MB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	34.7 kB / 2.2 GB	0.0 B	8	0	0	210	210	1.6 min (2 s)	0.0 B	6.5 MB	103.5 MB	0

**Executors**

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:63570	Active	0	34.7 kB / 2.2 GB	0.0 B	8	0	0	210	210	1.6 min (2 s)	0.0 B	6.5 MB	103.5 MB	Thread Dump

Showing 1 to 1 of 1 entries Previous  Next

Observation	Summary
<b>Slower Operation</b>	<b>groupByKey</b> is slower – shuffles all 10M values, very heavy.
<b>Shuffle Operations</b>	<b>groupByKey</b> , <b>reduceByKey</b> , and DF <b>groupBy</b> all trigger shuffles.
<b>Why reduceByKey is Narrow → Broad</b>	First does a narrow <b>map-side combine</b> , then a <b>broad shuffle</b> → more efficient than <b>groupByKey</b> .

## 3. Log files

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL LogPipeline5M application UI

**Spark Jobs [1]**

User: root  
Total Uptime: 1.6 min  
Scheduling Mode: FIFO  
Completed Jobs: 5

**Event Timeline**

Executor driver added

count at LogPipeline.scala:44 (Job 0) count at LogPipeline.scala:52 (Job 1) runJob at SparkHadoopWriter.scala:83 (Job 3)

json at LogPipeline.scala:66 (Job 4) json at LogPipeline.scala:66 (Job 5)

2 December 20:38 25 30 35 40 45 50 55 0 5 10 15 20 25 30 35

**Completed Jobs (5)**

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

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	json at LogPipeline.scala:66 json at LogPipeline.scala:66	2025/12/02 20:39:18	17 s	1/1	40/40
3	runJob at SparkHadoopWriter.scala:83 runJob at SparkHadoopWriter.scala:83	2025/12/02 20:39:02	15 s	1/1	40/40
2	count at LogPipeline.scala:52 count at LogPipeline.scala:52	2025/12/02 20:39:02	60 ms	1/1 (1 skipped)	1/1 (40 skipped)
1	count at LogPipeline.scala:52 count at LogPipeline.scala:52	2025/12/02 20:38:43	19 s	1/1	40/40
0	count at LogPipeline.scala:44 count at LogPipeline.scala:44	2025/12/02 20:38:22	20 s	1/1	40/40

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

## Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL LogPipeline5M application UI

**Stages for All Jobs**

Completed Stages: 5  
Skipped Stages: 1  
+ Completed Stages (5)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	json at LogPipeline.scala:66	+details 2025/12/02 20:39:18	17 s	40/40	425.4 MB			
4	runJob at SparkHadoopWriter.scala:83	+details 2025/12/02 20:39:02	15 s	40/40	65.5 MB			
3	count at LogPipeline.scala:52	+details 2025/12/02 20:39:02	45 ms	1/1			2.3 KIB	
1	count at LogPipeline.scala:52	+details 2025/12/02 20:38:43	19 s	40/40			2.3 KIB	
0	count at LogPipeline.scala:44	+details 2025/12/02 20:38:22	20 s	40/40				

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

**- Skipped Stages (1)**

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
2	count at LogPipeline.scala:52	+details Unknown	Unknown	0/40				

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

## Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL LogPipeline5M application UI

**Executors**

+ Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	118 KIB / 2.2 GB	0.0 B	8	0	0	161	161	9.4 min (4 s)	0.0 B	2.3 KIB	2.3 KIB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	118 KIB / 2.2 GB	0.0 B	8	0	0	161	161	9.4 min (4 s)	0.0 B	2.3 KIB	2.3 KIB	0

**Executors**

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:53206	Active	0	118 KIB / 2.2 GB	0.0 B	8	0	0	161	161	9.4 min (4 s)	0.0 B	2.3 KIB	2.3 KIB	Thread Dump

Showing 1 to 1 of 1 entries Previous  Next

Observation	Summary
Why is plain text slow?	Writes large raw strings → high I/O → many small files.
filter	Narrow (no shuffle).
map	Narrow (no shuffle).
sort	Broad (shuffle required).

## 4. Product Catalog

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL ProductPipeline2M application UI

**Spark Jobs (1)**  
User: racit  
Total Uptime: 1.7 min  
Scheduling Mode: FIFO  
Completed Jobs: 8

- Event Timeline  Enable zooming

Executors  All  Removed

Jobs  Succeeded  Failed  Running

45 50 55 0 5 10 15 20 25 30 35 40 45 50 55 0 5 10 15 20 25  
2 December 20:46 2 December 20:47 2 December 20:48

- Completed Jobs (8)

Page:	1	1 Pages. Jump to <input type="text"/> Show <input type="text"/> items in a page. Go			
Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	parquet at ProductPipeline.scala:61 parquet at ProductPipeline.scala:61	2025/12/02 20:48:23	2 s	1/1 (1 skipped)	9/9 (40 skipped)
6	parquet at ProductPipeline.scala:61 parquet at ProductPipeline.scala:61	2025/12/02 20:48:11	13 s	1/1	40/40
5	parquet at ProductPipeline.scala:61 parquet at ProductPipeline.scala:61	2025/12/02 20:47:53	18 s	1/1	40/40
4	csv at ProductPipeline.scala:54 csv at ProductPipeline.scala:54	2025/12/02 20:47:50	2 s	1/1 (1 skipped)	9/9 (40 skipped)
3	csv at ProductPipeline.scala:54 csv at ProductPipeline.scala:54	2025/12/02 20:47:36	14 s	1/1	40/40
2	csv at ProductPipeline.scala:54 csv at ProductPipeline.scala:54	2025/12/02 20:47:11	25 s	1/1	40/40
1	show at ProductPipeline.scala:47 show at ProductPipeline.scala:47	2025/12/02 20:46:45	25 s	1/1	40/40
0	show at ProductPipeline.scala:39 show at ProductPipeline.scala:39	2025/12/02 20:46:45	0.2 s	1/1	1/1

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

### Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL ProductPipeline2M application UI

**Stages for All Jobs**  
Completed Stages: 8  
Skipped Stages: 2

- Completed Stages (8)

Page:	1	1 Pages. Jump to <input type="text"/> Show <input type="text"/> items in a page. Go			
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input Output Shuffle Read Shuffle Write
9	parquet at ProductPipeline.scala:61	+details 2025/12/02 20:48:23	2 s	9/9	63.5 MiB 72.5 MiB
7	parquet at ProductPipeline.scala:61	+details 2025/12/02 20:48:11	13 s	40/40	
6	parquet at ProductPipeline.scala:61	+details 2025/12/02 20:47:53	18 s	40/40	
5	csv at ProductPipeline.scala:54	+details 2025/12/02 20:47:50	2 s	9/9	81.8 MiB 72.6 MiB
3	csv at ProductPipeline.scala:54	+details 2025/12/02 20:47:36	14 s	40/40	
2	csv at ProductPipeline.scala:54	+details 2025/12/02 20:47:11	25 s	40/40	
1	show at ProductPipeline.scala:47	+details 2025/12/02 20:46:45	25 s	40/40	
0	show at ProductPipeline.scala:39	+details 2025/12/02 20:46:45	85 ms	1/1	

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

- Skipped Stages (2)

Page:	1	1 Pages. Jump to <input type="text"/> Show <input type="text"/> items in a page. Go			
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input Output Shuffle Read Shuffle Write
8	parquet at ProductPipeline.scala:61	+details Unknown	Unknown	0/40	
4	csv at ProductPipeline.scala:54	+details Unknown	Unknown	0/40	

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

### Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL ProductPipeline2M application UI

**Executors**  
Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	179.1 KiB / 2.2 GiB	0.0 B	8	0	0	219	219	13 min (8 s)	0.0 B	145.1 MiB	145.1 MiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	179.1 KiB / 2.2 GiB	0.0 B	8	0	0	219	219	13 min (8 s)	0.0 B	145.1 MiB	145.1 MiB	0

Executors  
Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:56638	Active	0	179.1 KiB / 2.2 GiB	0.0 B	8	0	0	219	219	13 min (8 s)	0.0 B	145.1 MiB	145.1 MiB	Thread Dump

Show 1 to 1 of 1 entries Previous  Next

Observation	Summary
<b>CSV vs Parquet Write Speed</b>	CSV slower → writes huge plain text. Parquet faster → compressed columnar write.
<b>CSV vs Parquet Output Size</b>	CSV = largest (raw text). Parquet = smallest (compression + encoding).
<b>Why sort causes shuffle?</b>	Sorting needs global ordering, so Spark must shuffle all partitions → broad transformation.

## 5. IoT Sensor Data

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SensorPipeline3M application UI

**Spark Jobs (1)**  
User: racit  
Total Uptime: 1.2 min  
Scheduling Mode: FIFO  
Completed Jobs: 4

- Event Timeline  Enable zooming

Executors  Active  Removed

Jobs  Succeeded  Failed  Running

25 26 27 28 29 30 31 32 33 34  
2 December 20:59

show at SensorPipeline.scala:42 (Job 0) shd parquet at SensorPipeline.scala parquet at SensorPipeline.scala:50 (Job 3)

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
3	parquet at SensorPipeline.scala:50 parquet at SensorPipeline.scala:50	2025/12/02 20:59:31	3 s	1/1 (1 skipped)	1/1 (40 skipped)
2	parquet at SensorPipeline.scala:50 parquet at SensorPipeline.scala:50	2025/12/02 20:59:30	1.0 s	1/1	40/40
1	show at SensorPipeline.scala:42 show at SensorPipeline.scala:42	2025/12/02 20:59:29	0.1 s	1/1 (1 skipped)	1/1 (40 skipped)
0	show at SensorPipeline.scala:42 show at SensorPipeline.scala:42	2025/12/02 20:59:27	2 s	1/1	40/40

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

### Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SensorPipeline3M application UI

**Stages for All Jobs**  
Completed Stages: 4  
Skipped Stages: 2

- Completed Stages (4)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	parquet at SensorPipeline.scala:50	+details 2025/12/02 20:59:31	3 s	1/1	12.1 KiB	63.5 KiB		
3	parquet at SensorPipeline.scala:50	+details 2025/12/02 20:59:30	0.9 s	40/40			63.5 KiB	
2	show at SensorPipeline.scala:42	+details 2025/12/02 20:59:29	0.1 s	1/1		63.5 KiB		
0	show at SensorPipeline.scala:42	+details 2025/12/02 20:59:27	2 s	40/40			63.5 KiB	

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

- Skipped Stages (2)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	parquet at SensorPipeline.scala:50	+details Unknown	Unknown	0/40				
1	show at SensorPipeline.scala:42	+details Unknown	Unknown	0/40				

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

## Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL SensorPipeline3M application UI

**Executors**

Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	86 KIB / 2.2 GB	0.0 B	8	0	0	82	82	25 s (0.7 s)	0.0 B	127 KIB	127 KIB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	86 KIB / 2.2 GB	0.0 B	8	0	0	82	82	25 s (0.7 s)	0.0 B	127 KIB	127 KIB	0

**Executors**

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:61764	Active	0	86 KIB / 2.2 GB	0.0 B	8	0	0	82	82	25 s (0.7 s)	0.0 B	127 KIB	127 KIB	Thread Dump

Showing 1 to 1 of 1 entries

Search:

Previous  Next

Observation	Summary
Output folders created	24 folders (one per hour).
Why groupBy is broad	Needs a shuffle to group all same-hour records together.

## 6. Social media posts

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL UserPostPipeline application UI

**Spark Jobs [1]**

User:   
Total Uptime: 1 min  
Scheduling Mode: FIFO  
Completed Jobs: 11

Event Timeline  
 Enable zooming

**Executors**

- Added
- Removed

**Jobs**

- Succeeded
- Failed
- Running

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26  
2 December 21:37

**- Completed Jobs (11)**

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

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
10	json at UserPostPipeline.scala:70 json at UserPostPipeline.scala:70	2025/12/02 21:37:26	0.3 s	1/1 (3 skipped)	1/1 (78 skipped)
9	json at UserPostPipeline.scala:70 json at UserPostPipeline.scala:70	2025/12/02 21:37:25	1.0 s	1/1 (2 skipped)	8/8 (70 skipped)
8	json at UserPostPipeline.scala:70 json at UserPostPipeline.scala:70	2025/12/02 21:37:18	7 s	1/1	30/30
7	json at UserPostPipeline.scala:70 json at UserPostPipeline.scala:70	2025/12/02 21:37:18	6 s	1/1	40/40
6	show at UserPostPipeline.scala:63 show at UserPostPipeline.scala:63	2025/12/02 21:37:18	31 ms	1/1 (3 skipped)	1/1 (78 skipped)
5	show at UserPostPipeline.scala:63 show at UserPostPipeline.scala:63	2025/12/02 21:37:16	1 s	1/1 (2 skipped)	8/8 (70 skipped)
4	show at UserPostPipeline.scala:63 show at UserPostPipeline.scala:63	2025/12/02 21:37:10	7 s	1/1	30/30
3	show at UserPostPipeline.scala:63 show at UserPostPipeline.scala:63	2025/12/02 21:37:10	5 s	1/1	40/40
2	show at UserPostPipeline.scala:53 show at UserPostPipeline.scala:53	2025/12/02 21:37:09	0.4 s	1/1 (2 skipped)	1/1 (70 skipped)
1	show at UserPostPipeline.scala:53 show at UserPostPipeline.scala:53	2025/12/02 21:37:02	7 s	1/1	30/30
0	show at UserPostPipeline.scala:53 show at UserPostPipeline.scala:53	2025/12/02 21:37:02	6 s	1/1	40/40

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

## Stages

Stages for All Jobs								UserPostPipeline application UI			
Completed Stages: 11		Skipped Stages: 12									
Completed Stages (11)											
Page:	1									1 Pages. Jump to	1
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write		Show	100
22	json at UserPostPipeline.scala:70	+details 2025/12/02 21:37:26	0.3 s	1/1	1500.0 B	26.1 kB					
18	json at UserPostPipeline.scala:70	+details 2025/12/02 21:37:25	0.9 s	8/8		20.7 MB	26.1 kB				
15	json at UserPostPipeline.scala:70	+details 2025/12/02 21:37:18	2 s	30/30			10.4 MB				
14	json at UserPostPipeline.scala:70	+details 2025/12/02 21:37:18	6 s	40/40			10.3 MB				
13	show at UserPostPipeline.scala:63	+details 2025/12/02 21:37:18	21 ms	1/1		26.1 kB					
9	show at UserPostPipeline.scala:63	+details 2025/12/02 21:37:16	1 s	8/8		20.7 MB	26.1 kB				
6	show at UserPostPipeline.scala:63	+details 2025/12/02 21:37:10	2 s	30/30			10.4 MB				
5	show at UserPostPipeline.scala:63	+details 2025/12/02 21:37:10	5 s	40/40			10.3 MB				
4	show at UserPostPipeline.scala:53	+details 2025/12/02 21:37:09	0.4 s	1/1		10.2 MB					
1	show at UserPostPipeline.scala:53	+details 2025/12/02 21:37:02	2 s	30/30			19.2 MB				
0	show at UserPostPipeline.scala:53	+details 2025/12/02 21:37:02	6 s	40/40			65.5 MB				
Page:	1									1 Pages. Jump to	1
+ Skipped Stages (12)										Show	100
Page:	1									1 Pages. Jump to	1
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write		Show	100
21	json at UserPostPipeline.scala:70	+details Unknown	Unknown	0/8							
20	json at UserPostPipeline.scala:70	+details Unknown	Unknown	0/40							
19	json at UserPostPipeline.scala:70	+details Unknown	Unknown	0/30							
17	json at UserPostPipeline.scala:70	+details Unknown	Unknown	0/40							
16	json at UserPostPipeline.scala:70	+details Unknown	Unknown	0/30							
12	show at UserPostPipeline.scala:63	+details Unknown	Unknown	0/8							
11	show at UserPostPipeline.scala:63	+details Unknown	Unknown	0/30							
10	show at UserPostPipeline.scala:63	+details Unknown	Unknown	0/40							
8	show at UserPostPipeline.scala:63	+details Unknown	Unknown	0/30							
7	show at UserPostPipeline.scala:63	+details Unknown	Unknown	0/40							
3	show at UserPostPipeline.scala:53	+details Unknown	Unknown	0/30							
2	show at UserPostPipeline.scala:53	+details Unknown	Unknown	0/40							
Page:	1									1 Pages. Jump to	1

## Executors

Executors												UserPostPipeline application UI			
Show Additional Metrics															
Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	86.2 kB / 2.2 GiB	0.0 B	8	0	0	0	229	229	3.0 min (6 s)	0.0 B	51.8 MB	126.6 MB	0	
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0	
Total(1)	0	86.2 kB / 2.2 GiB	0.0 B	8	0	0	0	229	229	3.0 min (6 s)	0.0 B	51.8 MB	126.6 MB	0	
Showing 1 to 1 of 1 entries															Search: <input type="text"/>
Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24-53613	Active	0	86.2 kB / 2.2 GiB	0.0 B	8	0	0	229	229	3.0 min (6 s)	0.0 B	51.8 MB	126.6 MB	Thread Dump
Previous <span style="float: right;">Next</span>															

Observation	Summary
Why join is broad?	Needs a shuffle to move the same userId records to one partition.
Why DF join is easier?	Auto-optimized by Catalyst; less code than RDD key-value joins.

# 7. Financial Transactions

## Jobs

**Spark Jobs (1)**

User: racit  
Total Uptime: 28 s  
Scheduling Mode: FIFO  
Completed Jobs: 4

- Event Timeline  
Enable zooming

Executors  
All Active Removed

Jobs  
Succeeded Failed Running

63 54 55 56 57 58 59 60 1 2 3  
2 December 21:54 2 December 21:55

- Completed Jobs (4)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	show at TransactionPipeline.scala:52 show at TransactionPipeline.scala:52	2025/12/02 21:55:03	0.6 s	2/2 (1 skipped)	9/9 (40 skipped)
2	show at TransactionPipeline.scala:52 show at TransactionPipeline.scala:52	2025/12/02 21:55:00	2 s	1/1	40/40
1	take at TransactionPipeline.scala:37 take at TransactionPipeline.scala:37	2025/12/02 21:54:58	1 s	2/2 (1 skipped)	41/41 (40 skipped)
0	sortBy at TransactionPipeline.scala:36 sortBy at TransactionPipeline.scala:36	2025/12/02 21:54:55	3 s	2/2	80/80

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

Completed Stages (7)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
8	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:03	26 ms	1/1			2.1 kB	35.9 kB
7	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:03	0.5 s	8/8			2.1 kB	35.9 kB
5	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:00	2 s	40/40			60.4 kB	29.0 kB
4	take at TransactionPipeline.scala:37	+details 2025/12/02 21:54:59	66 ms	1/1			29.0 kB	2.4 kB
3	sortBy at TransactionPipeline.scala:36	+details 2025/12/02 21:54:58	1 s	40/40			29.0 kB	
1	sortBy at TransactionPipeline.scala:36	+details 2025/12/02 21:54:57	0.8 s	40/40			29.0 kB	
0	map at TransactionPipeline.scala:23	+details 2025/12/02 21:54:55	2 s	40/40			29.0 kB	

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

Skipped Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	show at TransactionPipeline.scala:52	+details Unknown	Unknown	0/40				
2	map at TransactionPipeline.scala:23	+details Unknown	Unknown	0/40				

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

## Stages

**Stages for All Jobs**

Completed Stages: 7  
Skipped Stages: 2

- Completed Stages (7)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
8	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:03	26 ms	1/1			2.1 kB	35.9 kB
7	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:03	0.5 s	8/8			2.1 kB	35.9 kB
5	show at TransactionPipeline.scala:52	+details 2025/12/02 21:55:00	2 s	40/40			60.4 kB	29.0 kB
4	take at TransactionPipeline.scala:37	+details 2025/12/02 21:54:59	66 ms	1/1			29.0 kB	2.4 kB
3	sortBy at TransactionPipeline.scala:36	+details 2025/12/02 21:54:58	1 s	40/40			29.0 kB	
1	sortBy at TransactionPipeline.scala:36	+details 2025/12/02 21:54:57	0.8 s	40/40			29.0 kB	
0	map at TransactionPipeline.scala:23	+details 2025/12/02 21:54:55	2 s	40/40			29.0 kB	

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

- Skipped Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	show at TransactionPipeline.scala:52	+details Unknown	Unknown	0/40				
2	map at TransactionPipeline.scala:23	+details Unknown	Unknown	0/40				

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

## Executors

**Executors**

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	54 kB / 2.2 GiB	0.0 B	8	0	0	170	170	56 s (3 s)	0.0 B	94.1 MiB	67.4 MiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	54 kB / 2.2 GiB	0.0 B	8	0	0	170	170	56 s (3 s)	0.0 B	94.1 MiB	67.4 MiB	0

Executors

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24-60962	Active	0	54 kB / 2.2 GiB	0.0 B	8	0	0	170	170	56 s (3 s)	0.0 B	94.1 MiB	67.4 MiB	Thread Dump

Showing 1 to 1 of 1 entries

Search:

Previous Next

Observation	Summary
Sorting	Broad – needs a full shuffle.
Grouping	Broad – keys must be shuffled together.
Why DF faster?	Catalyst + Tungsten give optimized execution and lower overhead.

## 8. Employee Data

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL EmployeePipeline1M application UI

**Spark Jobs (5)**

User: root  
Total Uptime: 25 s  
Scheduling Mode: FIFO  
Completed Jobs: 5

- Event Timeline  
Enable zooming

Executors  
Added  
Removed

Jobs  
Succeeded  
Failed  
Running

Executor driver added

show at EmployeePipeline.scala:40 (Job 0) sh csv at EmployeePipeline.scala:48 (Job 2) csv at c

2 December 22:29 10 11 12 13 14 15 16

**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	csv at EmployeePipeline.scala:55 csv at EmployeePipeline.scala:55	2025/12/02 22:29:16	41 ms	1/1	1/1
3	csv at EmployeePipeline.scala:48 csv at EmployeePipeline.scala:48	2025/12/02 22:29:15	0.2 s	1/1 (1 skipped)	1/1 (2 skipped)
2	csv at EmployeePipeline.scala:48 csv at EmployeePipeline.scala:48	2025/12/02 22:29:14	1 s	1/1	20/20
1	show at EmployeePipeline.scala:40 show at EmployeePipeline.scala:40	2025/12/02 22:29:14	71 ms	1/1 (1 skipped)	1/1 (20 skipped)
0	show at EmployeePipeline.scala:40 show at EmployeePipeline.scala:40	2025/12/02 22:29:12	2 s	1/1	20/20

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

### Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL EmployeePipeline1M application UI

**Stages for All Jobs**

Completed Stages: 5  
Skipped Stages: 2

- Completed Stages (5)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	csv at EmployeePipeline.scala:55	+details 2025/12/02 22:29:16	34 ms	1/1	123.0 B			
5	csv at EmployeePipeline.scala:48	+details 2025/12/02 22:29:15	0.2 s	1/1	111.0 B	6.1 KiB		
3	csv at EmployeePipeline.scala:48	+details 2025/12/02 22:29:14	1 s	20/20			6.1 KiB	
2	show at EmployeePipeline.scala:40	+details 2025/12/02 22:29:14	56 ms	1/1		6.1 KiB		
0	show at EmployeePipeline.scala:40	+details 2025/12/02 22:29:12	2 s	20/20			6.1 KiB	

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

**Skipped Stages (2)**

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	csv at EmployeePipeline.scala:48	+details Unknown	Unknown	0/20				
1	show at EmployeePipeline.scala:40	+details Unknown	Unknown	0/20				

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

## Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL EmployeePipeline1M application UI

**Executors**

Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	156.5 Kib / 2.2 GiB	0.0 B	8	0	0	43	43	26 s (0.6 s)	123 B	12.2 Kib	12.2 Kib	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	156.5 Kib / 2.2 GiB	0.0 B	8	0	0	43	43	26 s (0.6 s)	123 B	12.2 Kib	12.2 Kib	0

Executors

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:55209	Active	0	156.5 Kib / 2.2 GiB	0.0 B	8	0	0	43	43	26 s (0.6 s)	123 B	12.2 Kib	12.2 Kib	Thread Dump

Showing 1 to 1 of 1 entries

Search:

Previous  Next

Observation	Summary
CSV schema inference	CSV loads all columns as STRING, because CSV has no schema/metadata.
Parquet behavior	Parquet preserves original data types since it stores full schema internally.

## 9. Student Scores

### Jobs

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL StudentPipeline1\_5M application UI

**Spark Jobs [7]**

User: root  
Total Job Time: 21 s  
Scheduling Mode: FIFO  
Completed Jobs: 4

Event Timeline  Enable zooming

Event	Time	Description
Executor driver added	2 December 22:38:39	show at StudentPipeline.scala:37 (Job 0)
	40	[json at StudentPipeline.scala:44 (Job 1)]
	41	[json at StudentPipeline.scala:44 (Job 2)]
	42	[json at StudentPipeline.scala:44 (Job 3)]
	43	[json at StudentPipeline.scala:44 (Job 4)]
	44	[json at StudentPipeline.scala:44 (Job 5)]
	45	[json at StudentPipeline.scala:44 (Job 6)]
	46	[json at StudentPipeline.scala:44 (Job 7)]
	47	[json at StudentPipeline.scala:44 (Job 8)]
	48	[json at StudentPipeline.scala:44 (Job 9)]

2 December 22:38:39 40 41 42 43 44 45 46 47 48

- Completed Jobs (4)

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
3	json at StudentPipeline.scala:44 json at StudentPipeline.scala:44	2025/12/02 22:38:47	1 s	1/1 (1 skipped)	8/8 (20 skipped)
2	json at StudentPipeline.scala:44 json at StudentPipeline.scala:44	2025/12/02 22:38:45	2 s	1/1	20/20
1	json at StudentPipeline.scala:44 json at StudentPipeline.scala:44	2025/12/02 22:38:43	2 s	1/1	20/20
0	show at StudentPipeline.scala:37 show at StudentPipeline.scala:37	2025/12/02 22:38:40	2 s	1/1	20/20

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

## Stages

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL StudentPipeline1\_5M application UI

**Stages for All Jobs**

Completed Stages: 4  
Skipped Stages: 1  
+ Completed Stages (4)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	json at StudentPipeline.scala:44	+details 2025/12/02 22:38:47	1 s	8/8	68.9 MiB	19.5 MiB		
2	json at StudentPipeline.scala:44	+details 2025/12/02 22:38:45	1 s	20/20				19.5 MiB
1	json at StudentPipeline.scala:44	+details 2025/12/02 22:38:43	2 s	20/20				
0	show at StudentPipeline.scala:37	+details 2025/12/02 22:38:41	2 s	20/20				

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

+ Skipped Stages (1)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
3	json at StudentPipeline.scala:44	+details Unknown	Unknown	0/20				

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

## Executor

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL StudentPipeline1\_5M application UI

**Executors**

+ Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	80 KiB / 2.2 GiB	0.0 B	8	0	0	68	68	50 s (1.0 s)	0.0 B	19.5 MiB	19.5 MiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	80 KiB / 2.2 GiB	0.0 B	8	0	0	68	68	50 s (1.0 s)	0.0 B	19.5 MiB	19.5 MiB	0

**Executors**

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:59106	Active	0	80 KiB / 2.2 GiB	0.0 B	8	0	0	68	68	50 s (1.0 s)	0.0 B	19.5 MiB	19.5 MiB	Thread Dump

Showing 1 to 1 of 1 entries Previous  Next

Observation	Summary
Sorting	Broad – requires a full shuffle to globally order all students by score.
JSON Writing	Slowest format – verbose text, repeated field names, large output size.

# 10. Customer Transaction

## Jobs

**Spark Jobs (1)**

User: racit  
Total UpTime: 20 s  
Scheduling Policy: FIFO  
Active Jobs: 1  
Completed Jobs: 7

- Event Timeline  
Enable zooming

Executors (1 Active, 0 Removed)  
Jobs (1 Succeeded, 0 Failed, 0 Pending)

2 December 22:52 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 2 December 22:53

**Active Jobs (1)**

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	parquet at CustomerTransactionPipeline.scala:68 parquet at CustomerTransactionPipeline.scala:68	2025/12/02 22:53:05 (kill)	3 s	0/1	13/50 (9 running)

**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	parquet at CustomerTransactionPipeline.scala:68 parquet at CustomerTransactionPipeline.scala:68	2025/12/02 22:53:05	3 s	1/1	80/80
5	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:53:04	0.8 s	1/1 (2 skipped)	1/1 (10 skipped)
4	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:52:59	5 s	1/1	50/50
3	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:52:59	3 s	1/1	80/80
2	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:52:58	0.5 s	1/1 (2 skipped)	1/1 (10 skipped)
1	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:52:51	7 s	1/1	50/50
0	show at CustomerTransactionPipeline.scala:61 show at CustomerTransactionPipeline.scala:61	2025/12/02 22:52:51	4 s	1/1	80/80

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

## Stages

**Stages for All Jobs**

Completed Stages: 9  
Skipped Stages: 6  
- Completed Stages (9)

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
14	parquet at CustomerTransactionPipeline.scala:68	+details 2025/12/02 22:53:10	4 s	9/9	21.0 MB	79.6 MB		
11	parquet at CustomerTransactionPipeline.scala:68	+details 2025/12/02 22:53:05	3 s	50/50			10.4 MB	
10	parquet at CustomerTransactionPipeline.scala:68	+details 2025/12/02 22:53:05	3 s	80/80			69.2 MB	
9	show at CustomerTransactionPipeline.scala:61	+details 2025/12/02 22:53:04	0.8 s	1/1			9.6 MB	
6	show at CustomerTransactionPipeline.scala:61	+details 2025/12/02 22:52:59	2 s	50/50			10.4 MB	
5	show at CustomerTransactionPipeline.scala:61	+details 2025/12/02 22:52:59	3 s	80/80			69.2 MB	
4	show at CustomerTransactionPipeline.scala:51	+details 2025/12/02 22:52:58	0.4 s	1/1			14.4 MB	
1	show at CustomerTransactionPipeline.scala:51	+details 2025/12/02 22:52:51	3 s	50/50			28.9 MB	
0	show at CustomerTransactionPipeline.scala:51	+details 2025/12/02 22:52:51	4 s	80/80			91.1 MB	

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

**Skipped Stages (6)**

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

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
13	parquet at CustomerTransactionPipeline.scala:68	+details Unknown	Unknown	0/50				
12	parquet at CustomerTransactionPipeline.scala:68	+details Unknown	Unknown	0/50				
8	show at CustomerTransactionPipeline.scala:61	+details Unknown	Unknown	0/80				
7	show at CustomerTransactionPipeline.scala:61	+details Unknown	Unknown	0/50				
3	show at CustomerTransactionPipeline.scala:51	+details Unknown	Unknown	0/80				
2	show at CustomerTransactionPipeline.scala:51	+details Unknown	Unknown	0/50				

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

## Executors

Spark 3.2.1 Jobs Stages Storage Environment Executors SQL CustomerTransactionPipeline application UI

**Executors**

Show Additional Metrics

**Summary**

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	90.8 KIB / 2.2 GB	0.0 B	8	0	0	401	401	2.7 min (5 s)	0.0 B	103.6 MB	279.2 MB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	90.8 KIB / 2.2 GB	0.0 B	8	0	0	401	401	2.7 min (5 s)	0.0 B	103.6 MB	279.2 MB	0

Executors

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.1.24:64925	Active	0	90.8 KIB / 2.2 GB	0.0 B	8	0	0	401	401	2.7 min (5 s)	0.0 B	103.6 MB	279.2 MB	Thread Dump

Showing 1 to 1 of 1 entries

Search:

Previous  Next

Observation	Summary
Why joins & groupBy heavy?	Both need <b>full shuffle</b> so same keys go to the same partition → biggest data movement.
Why Parquet best?	<b>Columnar + compressed + schema-aware</b> → smallest size and fastest for analytics.