Spark Web UI

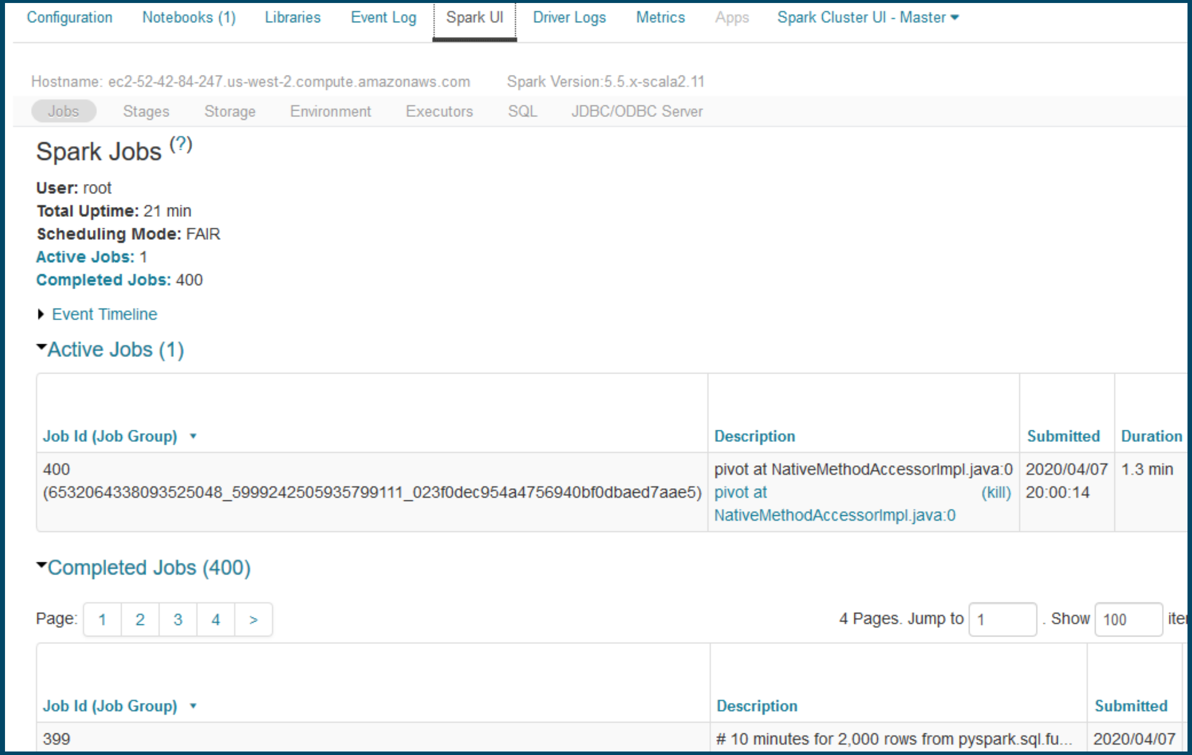
Interface to view Spark job progress

Spark Web UI

- Useful for
  - Visualizing Spark tasks
  - Diagnosing performance issues or failed tasks

Spark Web UI in Databricks

- Menu
  - Clusters -> your cluster -> Spark UI



Hostname: ec2-52-42-84-247.us-west-2.compute.amazonaws.com Spark Version:5.5.x-scala2.11

Spark Jobs (?)

User: root
Total Uptime: 21 min
Scheduling Mode: FAIR
Active Jobs: 1
Completed Jobs: 400

Event Timeline

Active Jobs (1)

Job Id (Job Group) ▾	Description	Submitted	Duration
400 (6532064338093525048_5999242505935799111_023f0dec954a4756940bf0dbaed7aae5)	pivot at NativeMethodAccessorImpl.java:0 pivot at NativeMethodAccessorImpl.java:0 (kill)	2020/04/07 20:00:14	1.3 min

Completed Jobs (400)

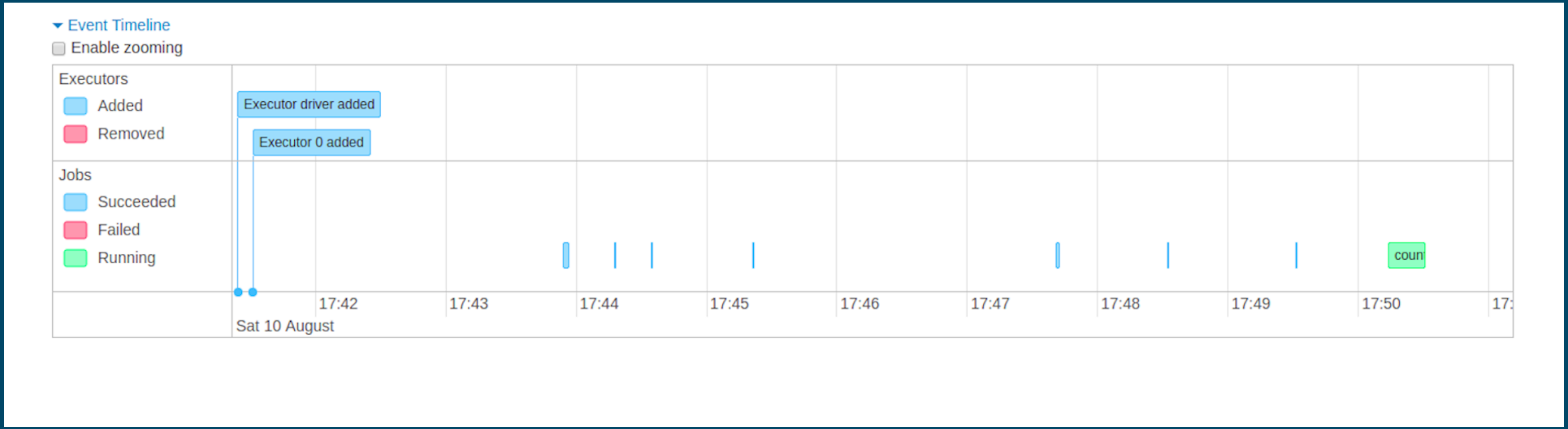
Page: 1 2 3 4 > 4 Pages. Jump to 1 . Show 100 ite

Job Id (Job Group) ▾	Description	Submitted
399	# 10 minutes for 2,000 rows from pyspark.sql.fu...	2020/04/07

https://spark.apache.org/docs/3.0.0-preview/web-ui.html

Event timeline

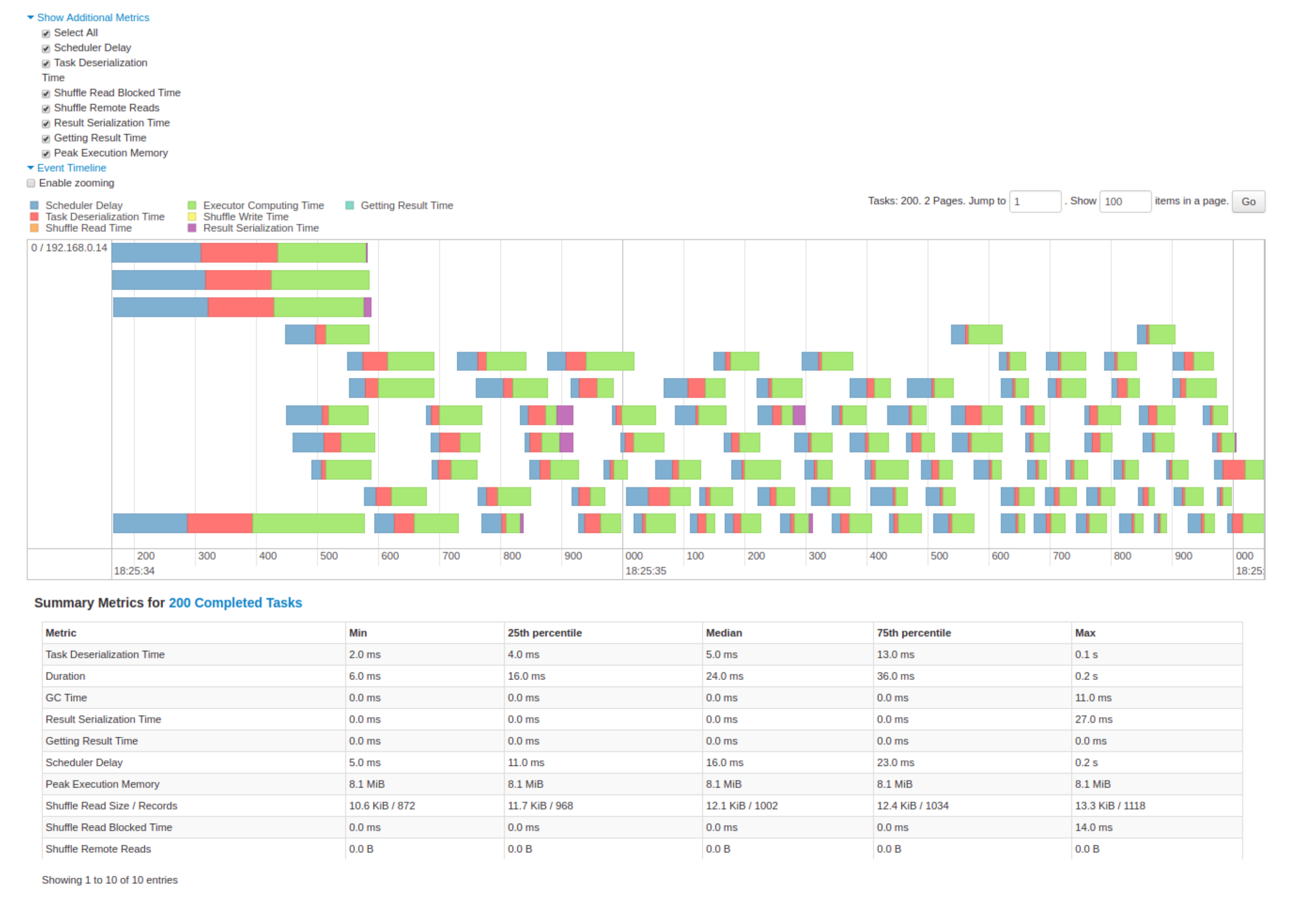
- Displays in chronological order the events related to the executors (added, removed) and the jobs



▾ Event Timeline

□ Enable zooming

Show detailed metrics (shuffle, sort)



▾ Show Additional Metrics

- ☑ Select All
- ☑ Scheduler Delay
- ☑ Task Deserialization Time
- ☑ Shuffle Read Blocked Time
- ☑ Shuffle Remote Reads
- ☑ Result Serialization Time
- ☑ Getting Result Time
- ☑ Peak Execution Memory

▾ Event Timeline

☐ Enable zooming

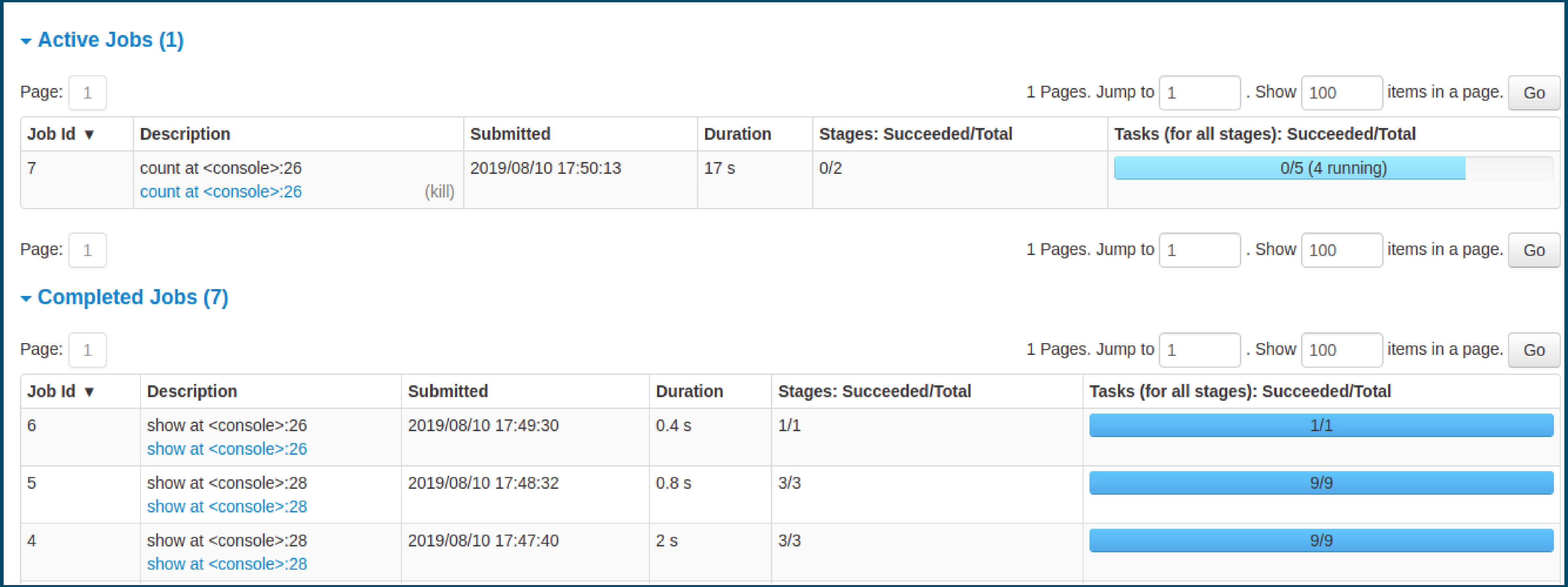
Tasks: 200. 2 Pages. Jump to 1 . Show 100 items in a page. Go

Summary Metrics for 200 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Task Deserialization Time	2.0 ms	4.0 ms	5.0 ms	13.0 ms	0.1 s
Duration	6.0 ms	16.0 ms	24.0 ms	36.0 ms	0.2 s
GC Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	11.0 ms
Result Serialization Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	27.0 ms
Getting Result Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	0.0 ms
Scheduler Delay	5.0 ms	11.0 ms	16.0 ms	23.0 ms	0.2 s
Peak Execution Memory	8.1 MiB	8.1 MiB	8.1 MiB	8.1 MiB	8.1 MiB
Shuffle Read Size / Records	10.6 KiB / 872	11.7 KiB / 968	12.1 KiB / 1002	12.4 KiB / 1034	13.3 KiB / 1118
Shuffle Read Blocked Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	14.0 ms
Shuffle Remote Reads	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B

Showing 1 to 10 of 10 entries

Details of jobs



▾ 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	count at <console>:26 count at <console>:26 (kill)	2019/08/10 17:50:13	17 s	0/2	0/5 (4 running)

Page: 1

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

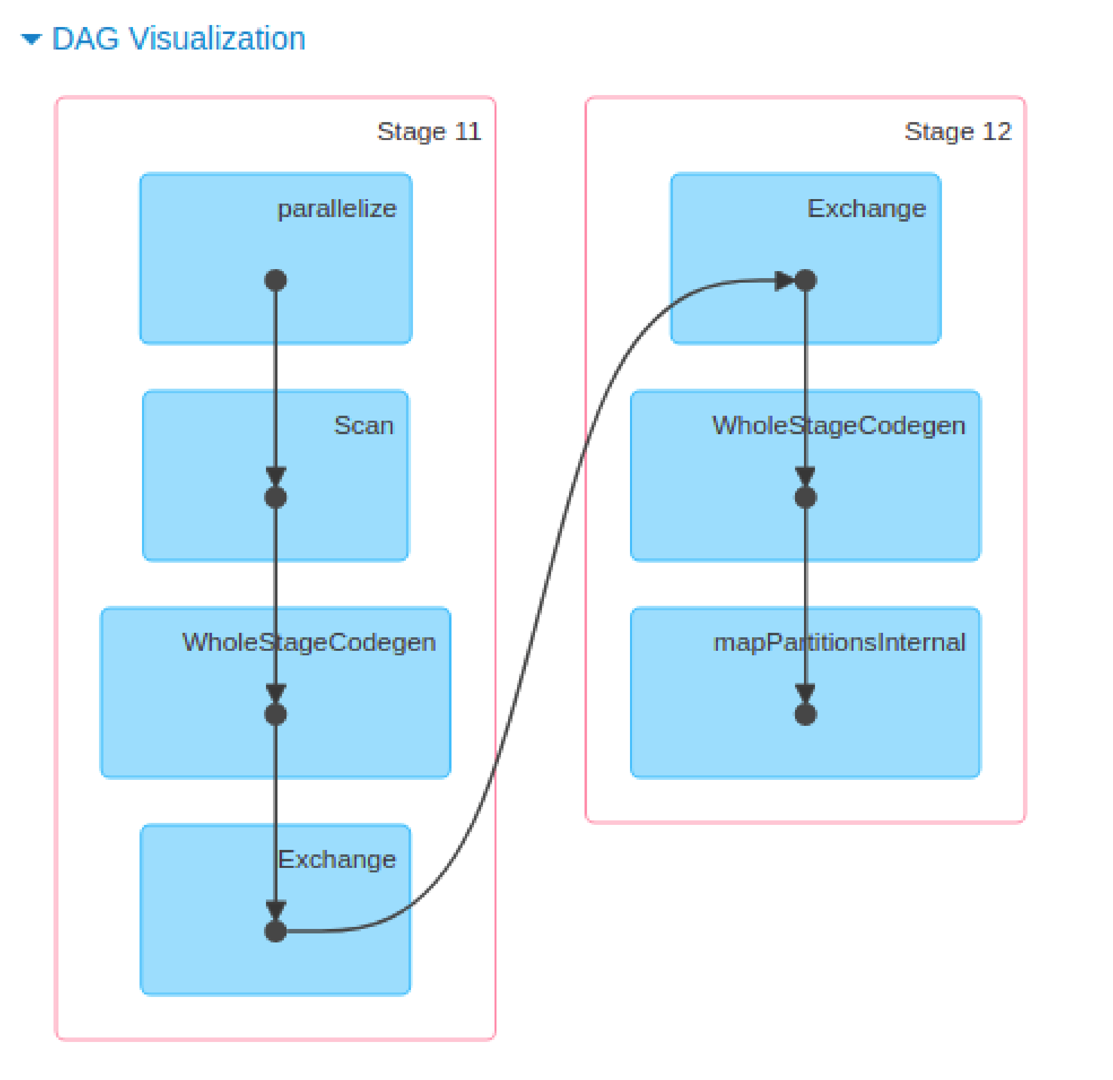
▾ Completed Jobs (7)

Page: 1

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

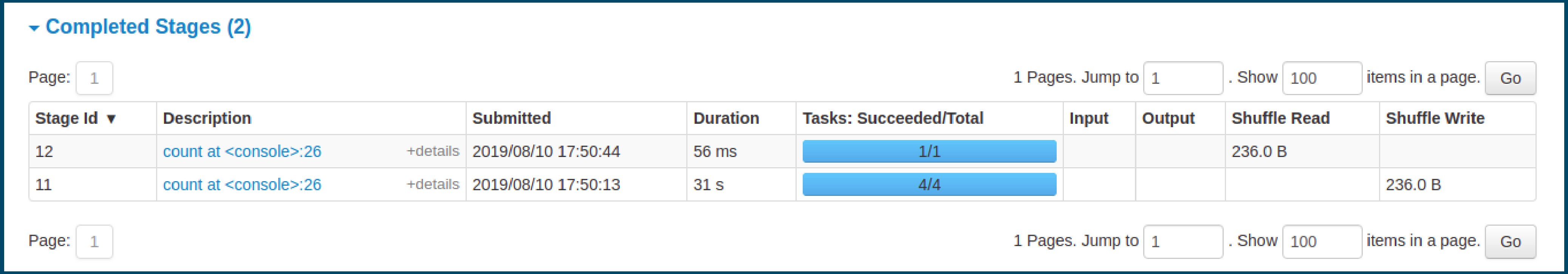
Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	show at <console>:26 show at <console>:26	2019/08/10 17:49:30	0.4 s	1/1	1/1
5	show at <console>:28 show at <console>:28	2019/08/10 17:48:32	0.8 s	3/3	9/9
4	show at <console>:28 show at <console>:28	2019/08/10 17:47:40	2 s	3/3	9/9

DAG visualization



Stages

- Helpful to diagnose long running or failed tasks



▾ Completed Stages (2)

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

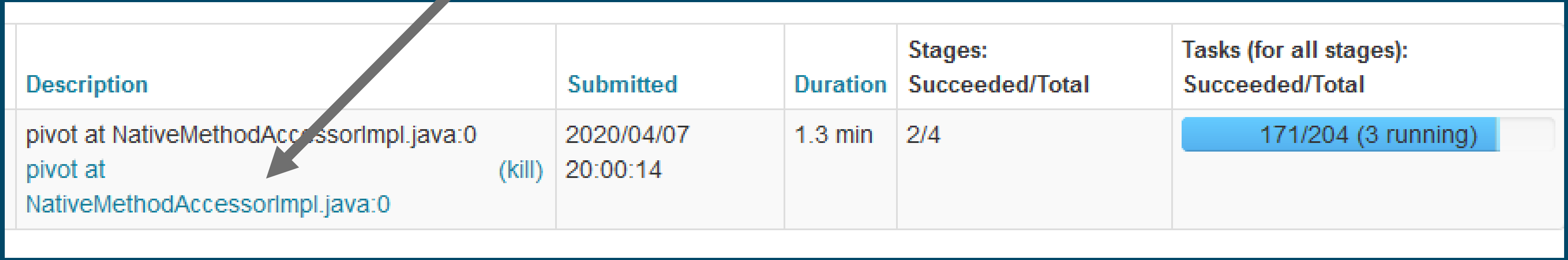
Stage Id ▼	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
12	count at <console>:26 +details	2019/08/10 17:50:44	56 ms	1/1			236.0 B	
11	count at <console>:26 +details	2019/08/10 17:50:13	31 s	4/4				236.0 B

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

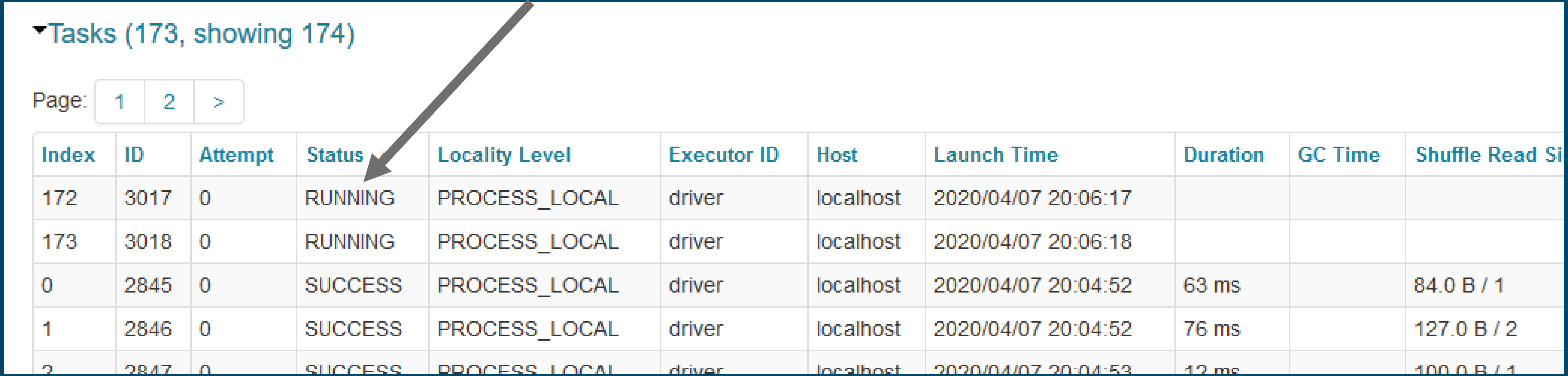
Diagnose long running or failed tasks

- Click through to see processes that are still running

Click for details



Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
pivot at NativeMethodAccessorImpl.java:0 pivot at NativeMethodAccessorImpl.java:0 (kill)	2020/04/07 20:00:14	1.3 min	2/4	171/204 (3 running)



▾Tasks (173, showing 174)

Page: 1 2 >

Index	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Shuffle Read Si
172	3017	0	RUNNING	PROCESS_LOCAL	driver	localhost	2020/04/07 20:06:17			
173	3018	0	RUNNING	PROCESS_LOCAL	driver	localhost	2020/04/07 20:06:18			
0	2845	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2020/04/07 20:04:52	63 ms		84.0 B / 1
1	2846	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2020/04/07 20:04:52	76 ms		127.0 B / 2
2	2847	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2020/04/07 20:04:53	12 ms		100.0 B / 1

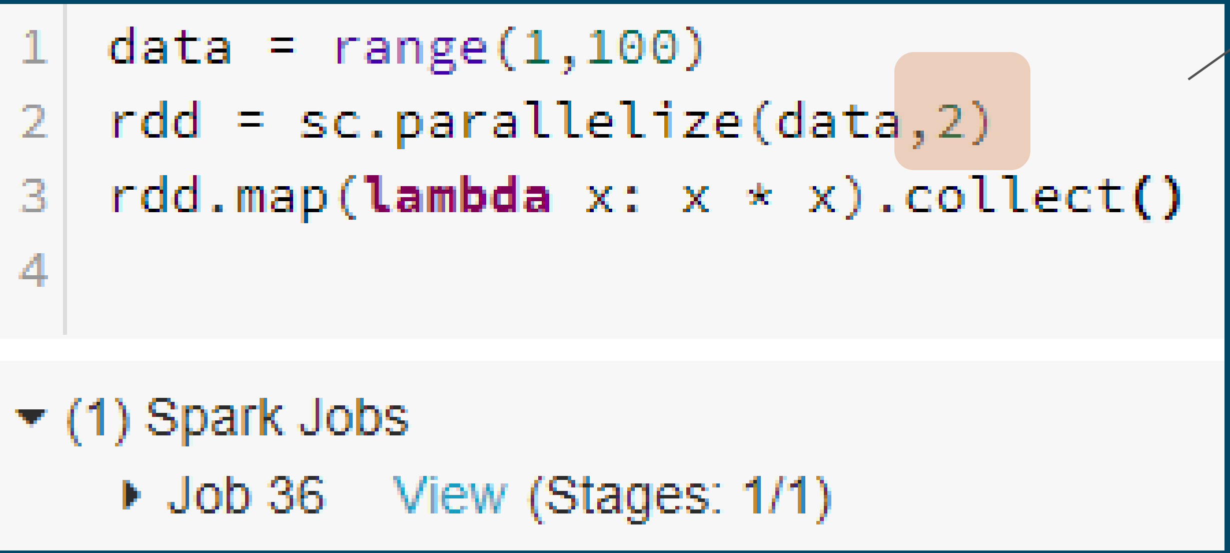
Each partition is processed by a task

1:1 One task for one partition

2 partitions -> 2 tasks

Notebook cell will be a **job** (a submit)

A **job** has many **tasks**

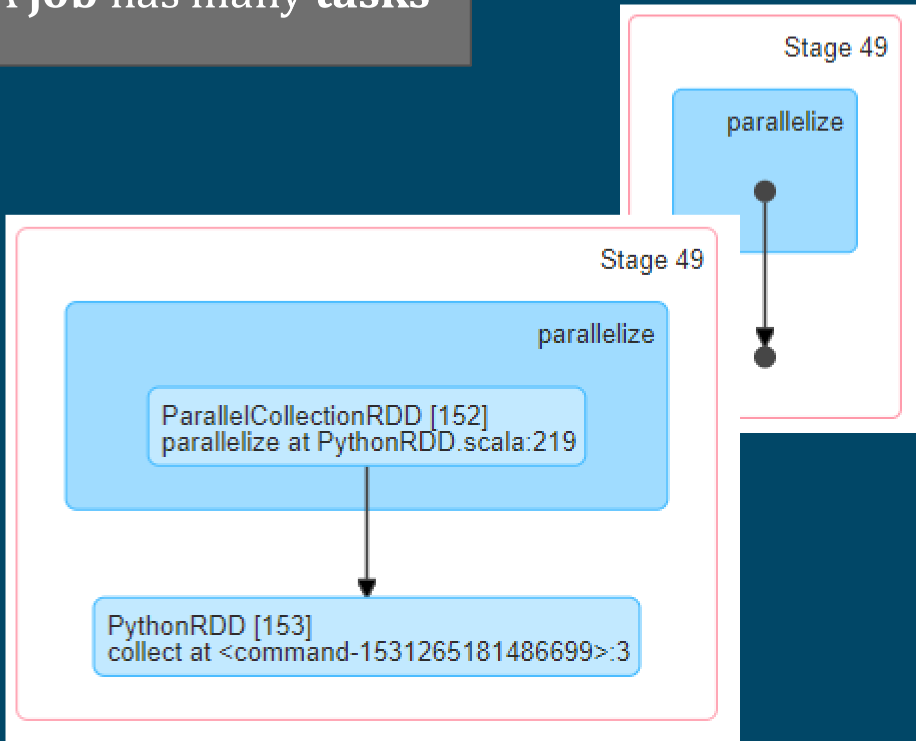


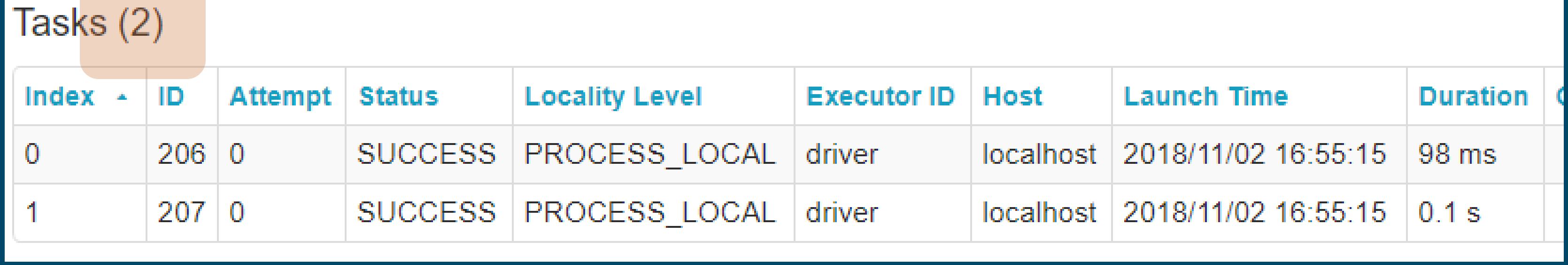
```
data = range(1,100)
rdd = sc.parallelize(data,2)
rdd.map(lambda x: x * x).collect()

```

▾ (1) Spark Jobs

▸ Job 36 View (Stages: 1/1)

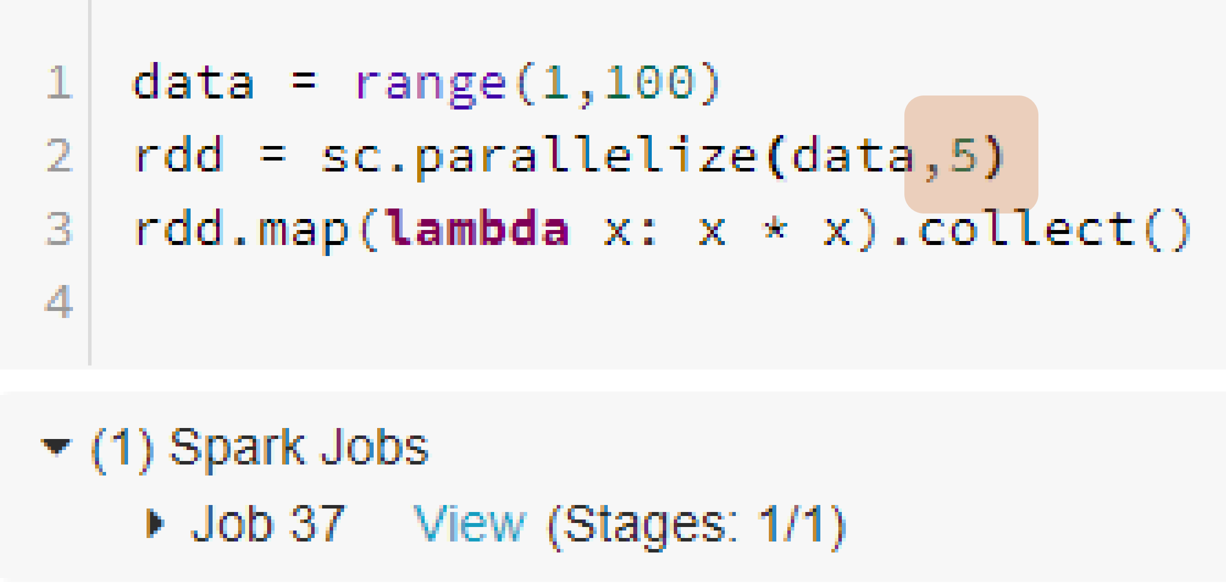

Stage 49
parallelize
Stage 49
parallelize
ParallelCollectionRDD [152]
parallelize at PythonRDD.scala:219
PythonRDD [153]
collect at <command-1531265181486699>:3



Tasks (2)

Index ▴	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration
0	206	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:55:15	98 ms
1	207	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:55:15	0.1 s

5 partitions -> 5 tasks

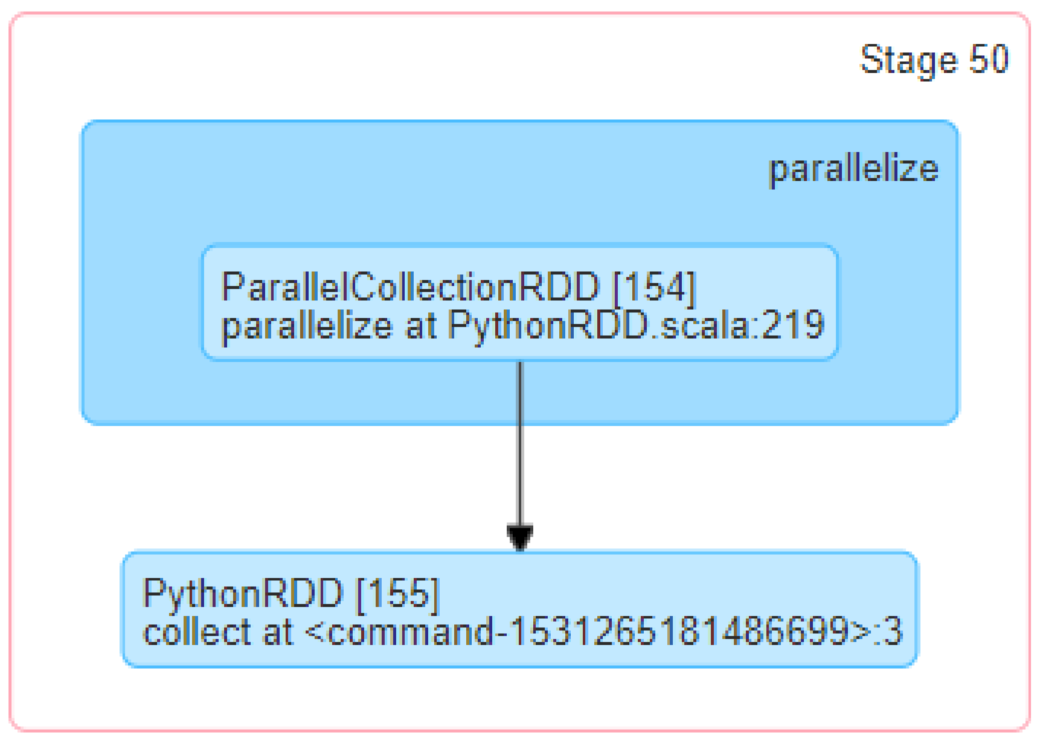


```
data = range(1,100)
rdd = sc.parallelize(data,5)
rdd.map(lambda x: x * x).collect()

```

▾ (1) Spark Jobs

▸ Job 37 View (Stages: 1/1)

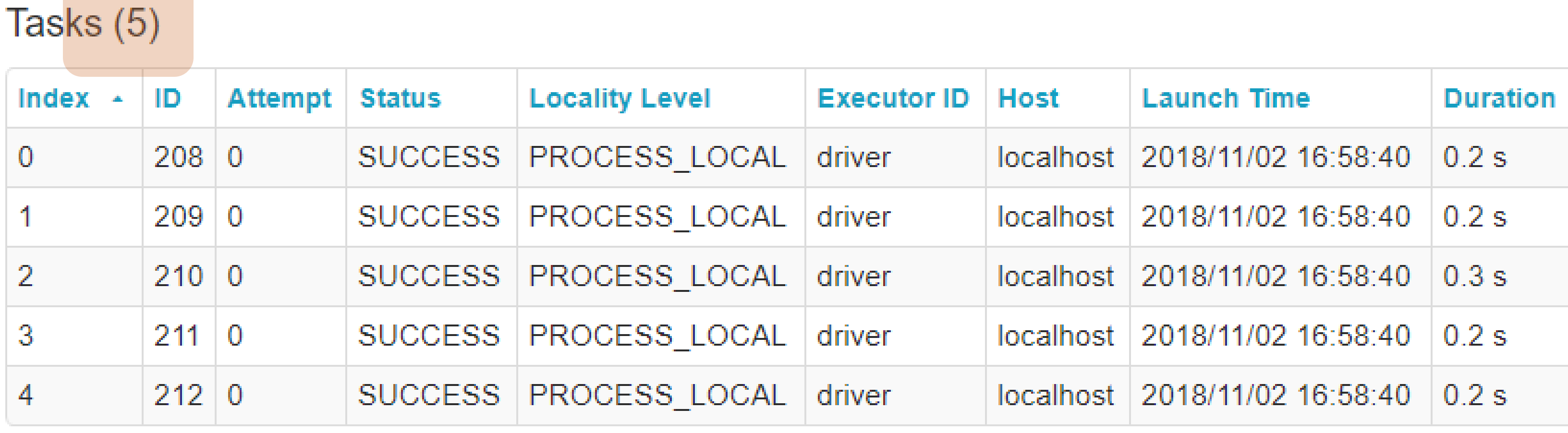


Stage 50

parallelize

ParallelCollectionRDD [154]
parallelize at PythonRDD.scala:219

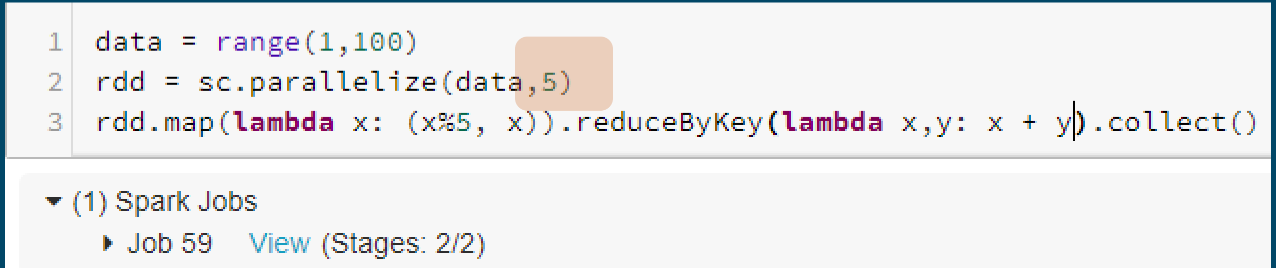
PythonRDD [155]
collect at <command-1531265181486699>:3



Tasks (5)

Index ▴	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration
0	208	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:58:40	0.2 s
1	209	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:58:40	0.2 s
2	210	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:58:40	0.3 s
3	211	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:58:40	0.2 s
4	212	0	SUCCESS	PROCESS_LOCAL	driver	localhost	2018/11/02 16:58:40	0.2 s

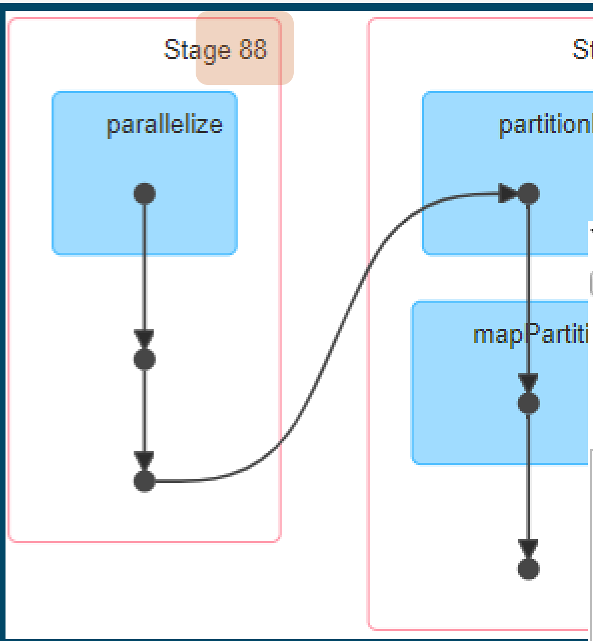
Map Reduce with 5 tasks

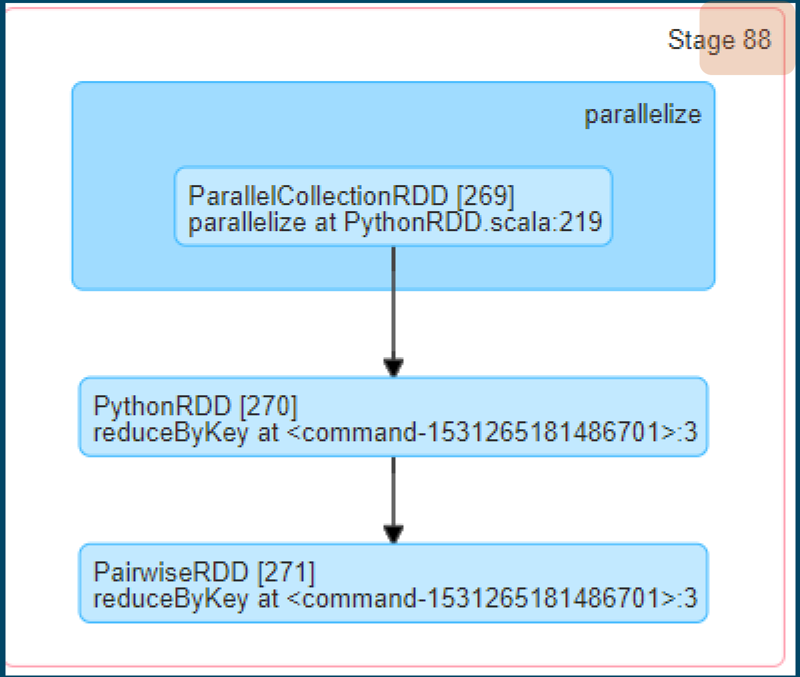


```
data = range(1,100)
rdd = sc.parallelize(data,5)
rdd.map(lambda x: (x%5, x)).reduceByKey(lambda x,y: x + y).collect()
```

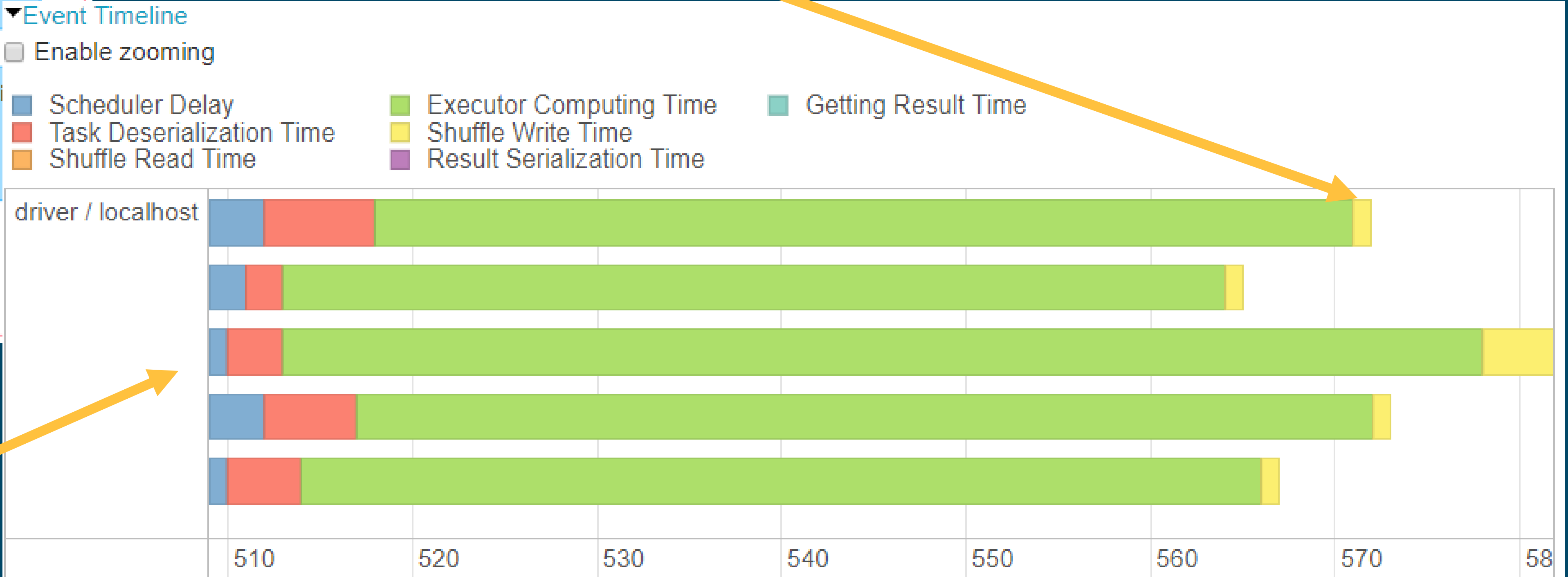
▾ (1) Spark Jobs

▸ Job 59 View (Stages: 2/2)



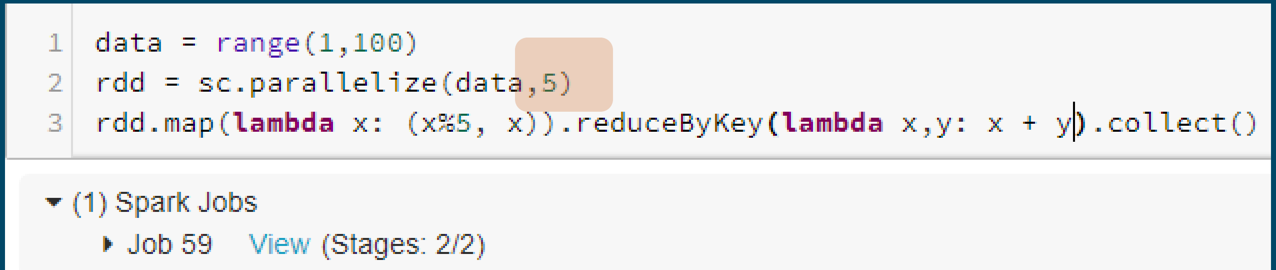

Stage 88
parallelize
ParallelCollectionRDD [269]
parallelize at PythonRDD.scala:219
PythonRDD [270]
reduceByKey at <command-1531265181486701>:3
PairwiseRDD [271]
reduceByKey at <command-1531265181486701>:3

Shuffle shown @ end



5 Tasks

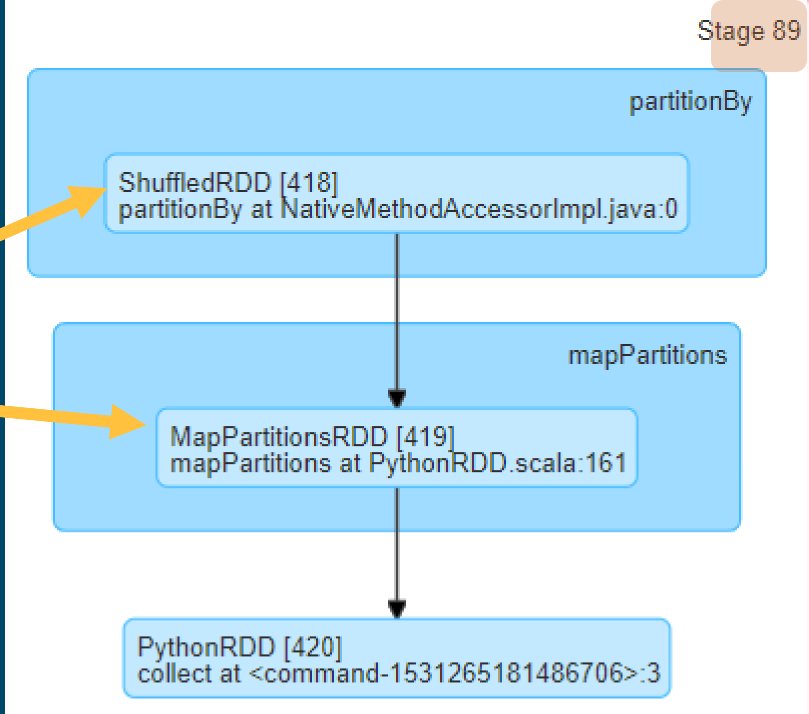
Map Reduce with 5 tasks



```
data = range(1,100)
rdd = sc.parallelize(data,5)
rdd.map(lambda x: (x%5, x)).reduceByKey(lambda x,y: x + y).collect()
```

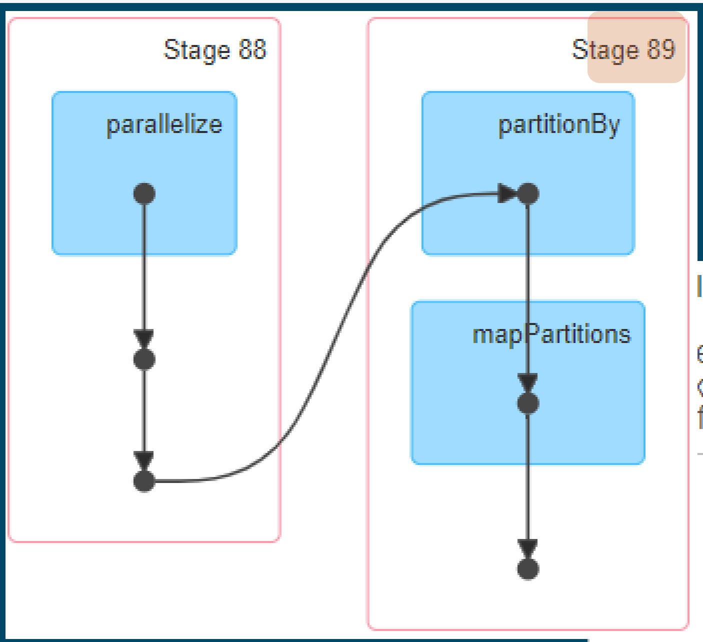
(1) Spark Jobs

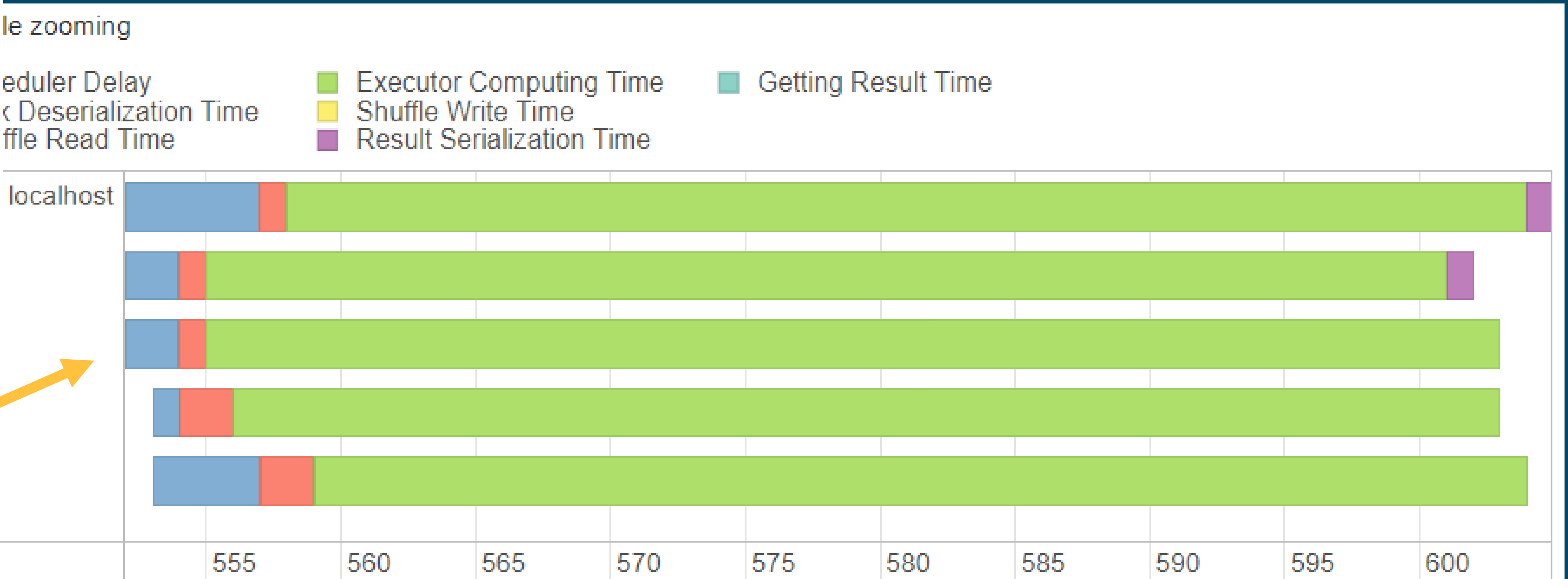
Job 59 View (Stages: 2/2)


Stage 89
partitionBy
ShuffledRDD [418]
partitionBy at NativeMethodAccessorImpl.java:0
mapPartitions
MapPartitionsRDD [419]
mapPartitions at PythonRDD.scala:161
PythonRDD [420]
collect at <command-1531265181486706>:3

Shuffled

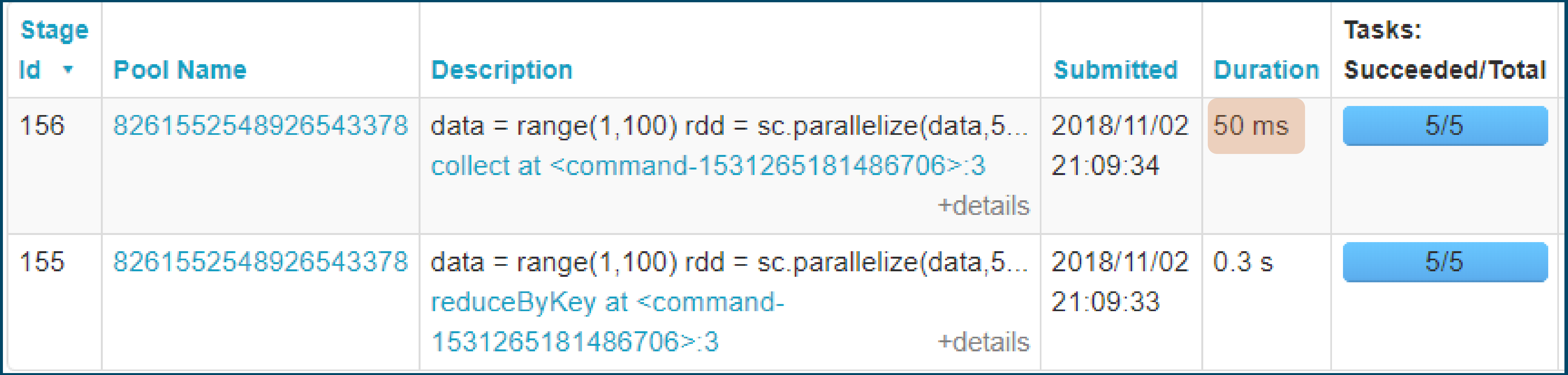
Reduce





5 Tasks

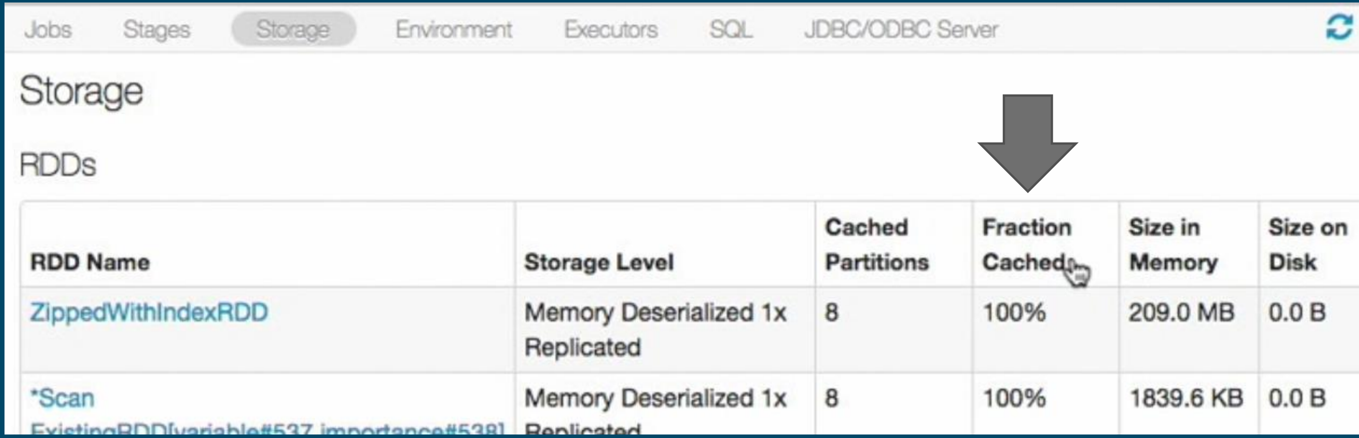
Map Reduce with 5 tasks



Stage Id ▾	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total
156	8261552548926543378	data = range(1,100) rdd = sc.parallelize(data,5… collect at <command-1531265181486706>:3 +details	2018/11/02 21:09:34	50 ms	5/5
155	8261552548926543378	data = range(1,100) rdd = sc.parallelize(data,5… reduceByKey at <command-1531265181486706>:3 +details	2018/11/02 21:09:33	0.3 s	5/5

Important to remember

- Spark Web UI
  - Useful for
    - Visualizing Spark tasks
    - Diagnosing performance issues or failed tasks
- Look for
  - Failed task
    - View the log
  - Slow tasks (high duration)
    - Lots of shuffle
    - In memory (Fraction cached is Storage tab of UI)



Storage

RDDs

RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
ZippedWithIndexRDD	Memory Deserialized 1x Replicated	8	100%	209.0 MB	0.0 B
*Scan ExistingRDD[variable#537,importance#538]	Memory Deserialized 1x Replicated	8	100%	1839.6 KB	0.0 B