THE UNIVERSITY OF MELBOURNE

Workshop 9 - Big Data Analytics

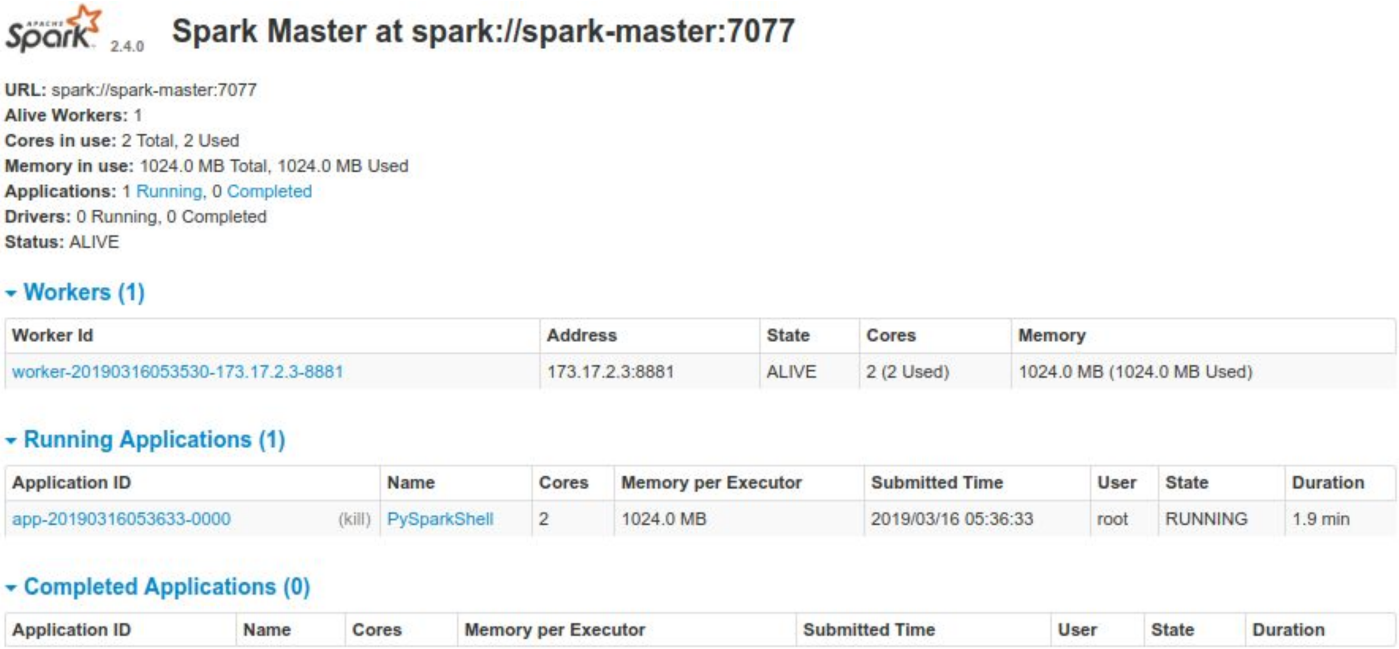
Luca Morandini
Data Architect – AURIN Project
University of Melbourne
luca.morandini@unimelb.edu.au

Starting Spark With Docker

- The instructions on how to build a mini-cluster (a Spark master and one Spark worker Docker containers), and an example of how to run a word-count MapReduce job is at the usual `https://github.com/AURIN/comp90024` repository, go under the `spark` directory and follow the README.

Visualizing a Spark Cluster

- Spark *Web-UI* allows to graphically see nodes (port 8080)



2.4.0 **Spark Master at spark://spark-master:7077**

URL: spark://spark-master:7077
Alive Workers: 1
Cores in use: 2 Total, 2 Used
Memory in use: 1024.0 MB Total, 1024.0 MB Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20190316053530-173.17.2.3-8881	173.17.2.3:8881	ALIVE	2 (2 Used)	1024.0 MB (1024.0 MB Used)

▾ Running Applications (1)

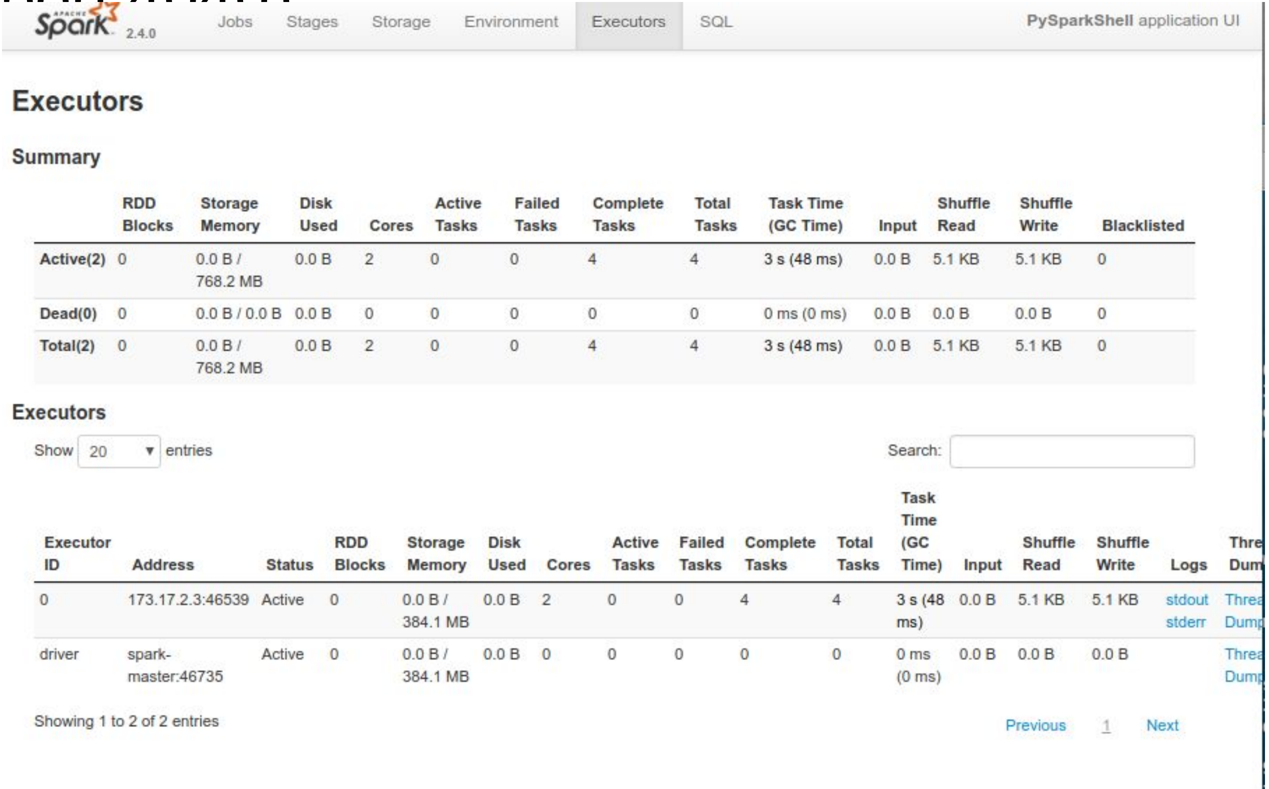
Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20190316053633-0000 (kill)	PySparkShell	2	1024.0 MB	2019/03/16 05:36:33	root	RUNNING	1.9 min

▾ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Visualizing Spark Jobs: Executors

- Spark *Web-UI* allows as well to see the division of jobs in stages and tasks, and the allocation of executors to nodes (port 4040)



Executors

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	Blacklisted
Active(2)	0	0.0 B / 768.2 MB	0.0 B	2	0	0	4	4	3 s (48 ms)	0.0 B	5.1 KB	5.1 KB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B	0
Total(2)	0	0.0 B / 768.2 MB	0.0 B	2	0	0	4	4	3 s (48 ms)	0.0 B	5.1 KB	5.1 KB	0

Executors

Show 20 entries

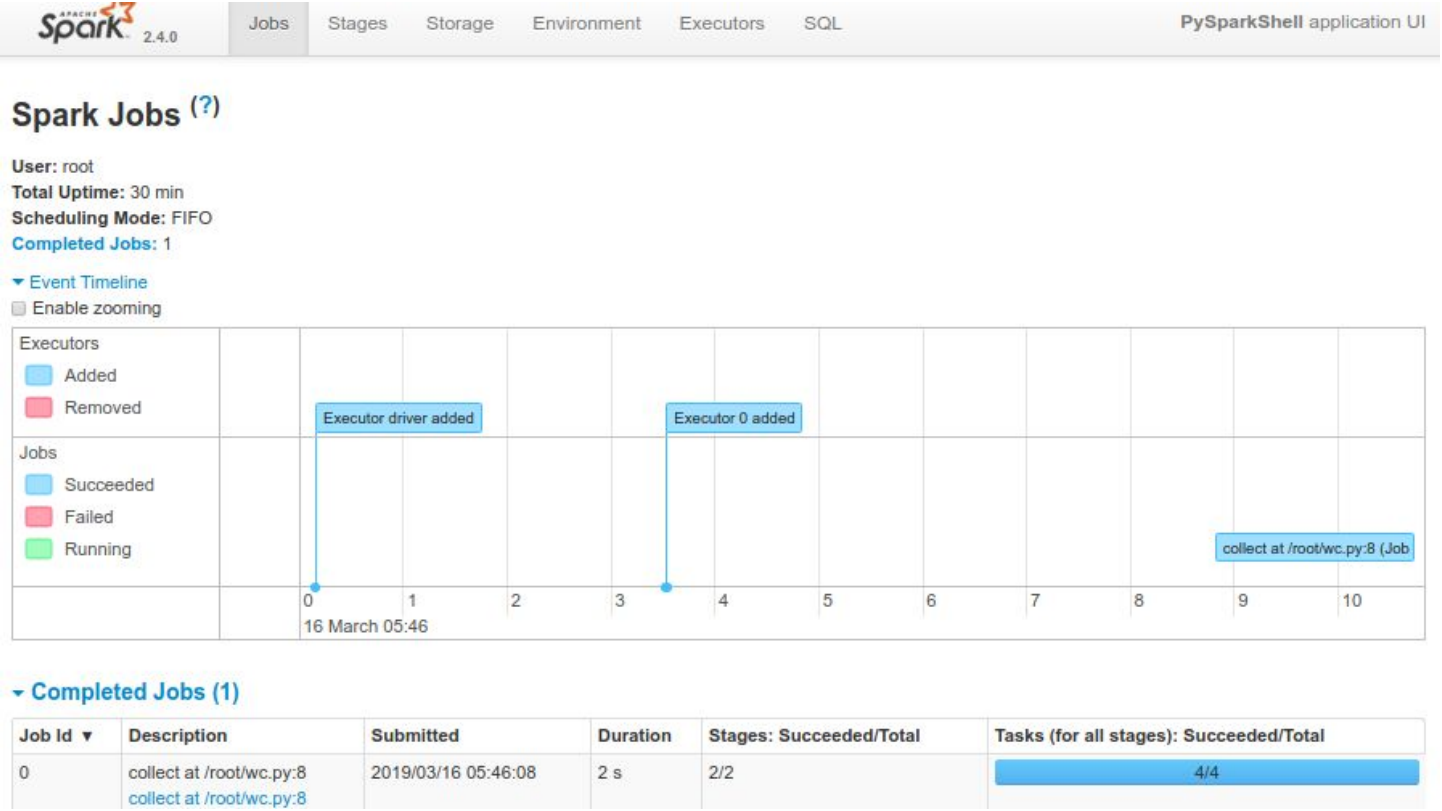
Search:

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	Logs	Thre Dum
0	173.17.2.3:46539	Active	0	0.0 B / 384.1 MB	0.0 B	2	0	0	4	4	3 s (48 ms)	0.0 B	5.1 KB	5.1 KB	stdout stderr	Threa Dump
driver	spark-master:46735	Active	0	0.0 B / 384.1 MB	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B		Threa Dump

Showing 1 to 2 of 2 entries

Previous 1 Next

Visualizing Spark Jobs: Jobs



Spark Jobs (?)

User: root
Total Uptime: 30 min
Scheduling Mode: FIFO
Completed Jobs: 1

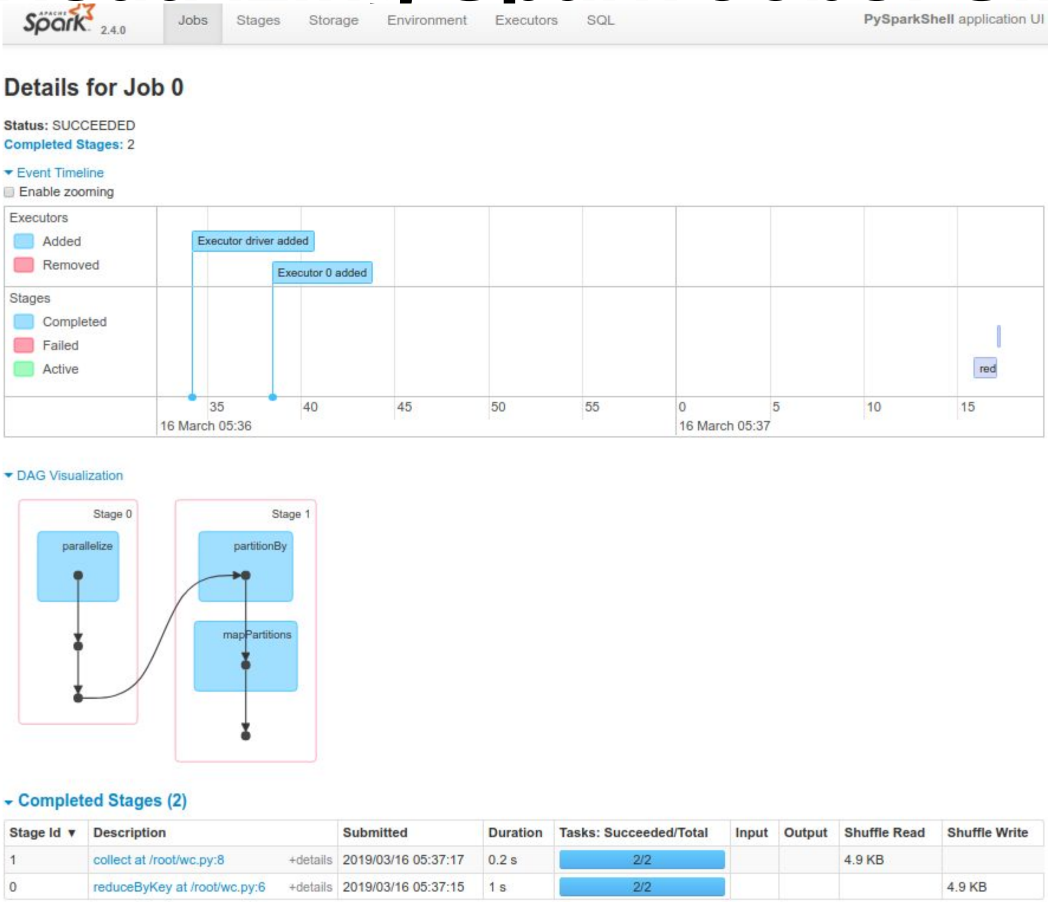
▾ Event Timeline

□ Enable zooming

▾ Completed Jobs (1)

Job Id ▼	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	collect at /root/wc.py:8 collect at /root/wc.py:8	2019/03/16 05:46:08	2 s	2/2	4/4

Visualizing Spark Jobs: Single Job



Details for Job 0

Status: SUCCEEDED
Completed Stages: 2

▾ Event Timeline

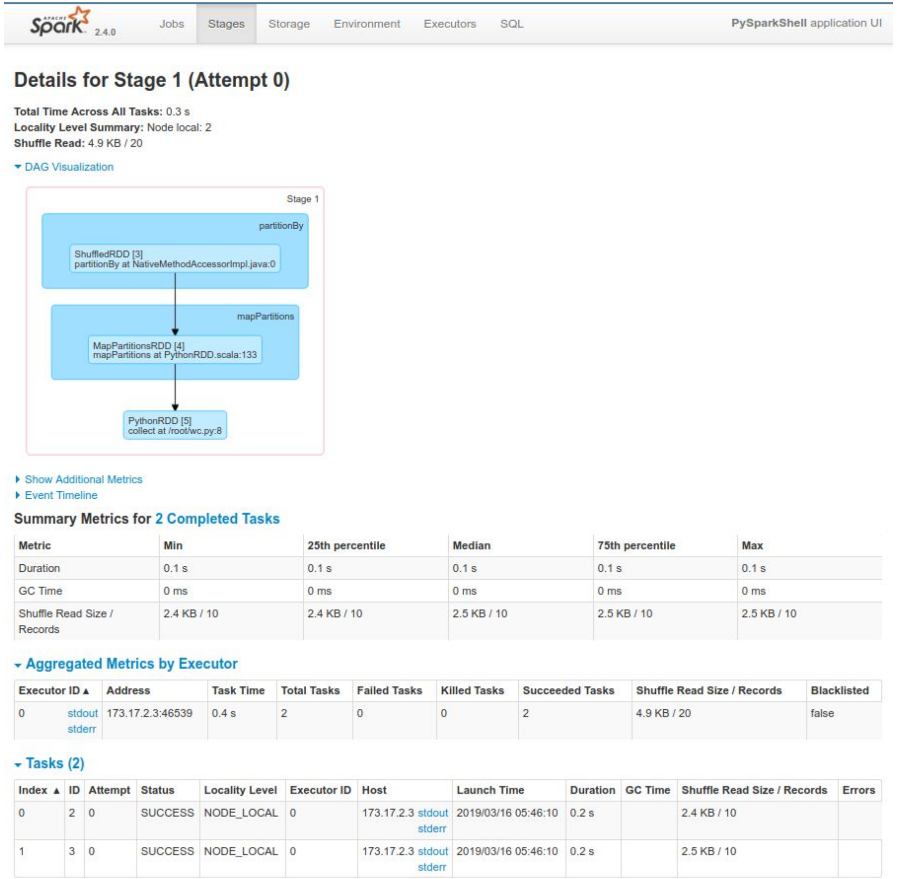
□ Enable zooming

▾ DAG Visualization

▾ **Completed Stages (2)**

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	collect at /root/wc.py:8	+details	2019/03/16 05:37:17	0.2 s	2/2			4.9 KB	
0	reduceByKey at /root/wc.py:6	+details	2019/03/16 05:37:15	1 s	2/2				4.9 KB

Visualizing Spark Jobs: Tasks of a Stage



Details for Stage 1 (Attempt 0)

Total Time Across All Tasks: 0.3 s
Locality Level Summary: Node local: 2
Shuffle Read: 4.9 KB / 20

▾ DAG Visualization

▸ Show Additional Metrics
▸ Event Timeline

Summary Metrics for 2 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.1 s	0.1 s	0.1 s	0.1 s	0.1 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Shuffle Read Size / Records	2.4 KB / 10	2.4 KB / 10	2.5 KB / 10	2.5 KB / 10	2.5 KB / 10

▾ Aggregated Metrics by Executor

Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Shuffle Read Size / Records	Blacklisted
0 stdout stderr	173.17.2.3:46539	0.4 s	2	0	0	2	4.9 KB / 20	false

▾ Tasks (2)

Index ▲	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Shuffle Read Size / Records	Errors
0	2	0	SUCCESS	NODE_LOCAL	0	173.17.2.3 stdout stderr	2019/03/16 05:46:10	0.2 s		2.4 KB / 10	
1	3	0	SUCCESS	NODE_LOCAL	0	173.17.2.3 stdout stderr	2019/03/16 05:46:10	0.2 s		2.5 KB / 10	