Yandex

Spark UI

$ pyspark

```
>>> from collections import namedtuple
>>> from datetime import datetime, timedelta
>>> Record = namedtuple("Record", ["date", "open", "high", "low", "close",
                                   "adj_close", "volume"])
>>> def parse_record(s):
...   fields = s.split(",")
...   return Record(fields[0], *map(float, fields[1:6]), int(fields[6]))
...
>>> def get_next_date(s):
...   fmt = "%Y-%m-%d"
...   return (datetime.strptime(s, fmt) + timedelta(days=1)).strftime(fmt)
...
>>> parsed_data = sc.textFile("nasdaq.csv").map(parse_record).cache()
>>> date_and_close_price = parsed_data.map(lambda r: (r.date, r.close))
>>> date_and_prev_close_price = parsed_data.map(lambda r: (get_next_date(r.date), r.close))
>>> joined = date_and_close_price.join(date_and_prev_close_price)
>>> returns = joined.mapValues(lambda p: (p[0] / p[1] - 1.0) * 100.0)
>>>
```

$ pyspark

```
>>> from collections import namedtuple
>>> from datetime import datetime, timedelta
>>> Record = namedtuple("Record", ["date", "open", "high", "low", "close",
                                   "adj_close", "volume"])
>>> def parse_record(s):
...   fields = s.split(",")
...   return Record(fields[0], *map(float, fields[1:6]), int(fields[6]))
...
>>> def get_next_date(s):
...   fmt = "%Y-%m-%d"
...   return (datetime.strptime(s, fmt) + timedelta(days=1)).strftime(fmt)
...
>>> parsed_data = sc.textFile("nasdaq.csv").map(parse_record).cache()
>>> date_and_close_price = parsed_data.map(lambda r: (r.date, r.close))
>>> date_and_prev_close_price = parsed_data.map(lambda r: (get_next_date(r.date), r.close))
>>> joined = date_and_close_price.join(date_and_prev_close_price)
>>> returns = joined.mapValues(lambda p: (p[0] / p[1] - 1.0) * 100.0)
>>>
>>> sc.uiWebUrl
'http://127.0.0.1:4040'
>>>
```

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | **PySparkShell** application UI

Spark Jobs (?)

User: sandello
Total Uptime: 10 min
Scheduling Mode: FIFO

▸ Event Timeline

$ pyspark

```
>>> from collections import namedtuple
>>> from datetime import datetime, timedelta
>>> Record = namedtuple("Record", ["date", "open", "high", "low", "close",
                                   "adj_close", "volume"])
>>> def parse_record(s):
...   fields = s.split(",")
...   return Record(fields[0], *map(float, fields[1:6]), int(fields[6]))
...
>>> def get_next_date(s):
...   fmt = "%Y-%m-%d"
...   return (datetime.strptime(s, fmt) + timedelta(days=1)).strftime(fmt)
...
>>> parsed_data = sc.textFile("nasdaq.csv").map(parse_record).cache()
>>> date_and_close_price = parsed_data.map(lambda r: (r.date, r.close))
>>> date_and_prev_close_price = parsed_data.map(lambda r: (get_next_date(r.date), r.close))
>>> joined = date_and_close_price.join(date_and_prev_close_price)
>>> returns = joined.mapValues(lambda p: (p[0] / p[1] - 1.0) * 100.0)
>>>
>>> returns.top(1, lambda x: x[1])
[('2017-06-28', 1.4282652470614554)]
>>>
```

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | **PySparkShell** application UI

Spark Jobs (?)

User: sandello
Total Uptime: 15 min
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	top at <ipython-input-15-0b58397eb290>:1	2017/07/23 22:16:13	2 s	2/2	8/8

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | **PySparkShell** application UI

Spark Jobs (?)

User: sandello
Total Uptime: 15 min
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	top at <ipython-input-15-0b58397eb290>:1	2017/07/23 22:16:13	2 s	2/2	8/8

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | **PySparkShell** application UI

Spark Jobs (?)

User: sandello
Total Uptime: 15 min
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	top at <ipython-input-15-0b58397eb290>:1	2017/07/23 22:16:13	2 s	2/2	8/8

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | PySparkShell application UI

Spark Jobs (?)

User: sandello
Total Uptime: 15 min
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	top at <ipython-input-15-0b58397eb290>:1	2017/07/23 22:16:13	2 s	2/2	8/8

Spark 2.2.0 | Jobs | Stages | Storage | Environment | Executors | SQL | **PySparkShell** application UI

Spark Jobs (?)

User: sandello
Total Uptime: 15 min
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

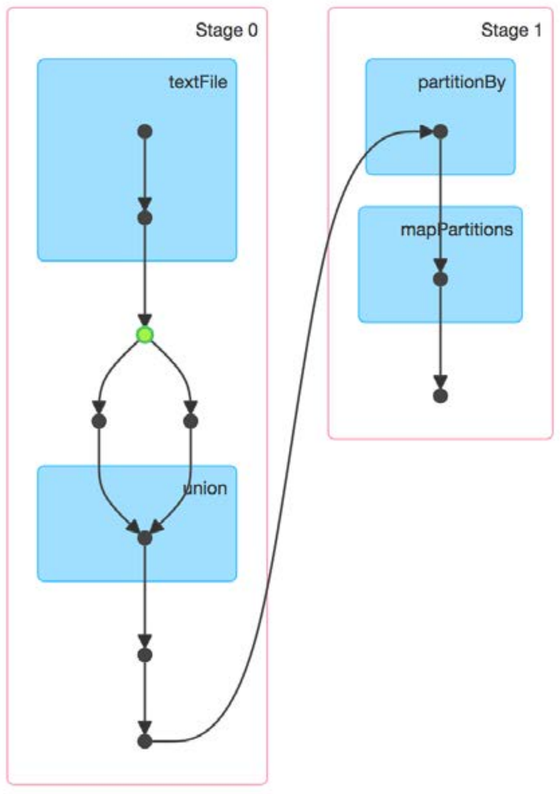
Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	top at <ipython-input-15-0b58397eb290>:1	2017/07/23 22:16:13	2 s	2/2	8/8

Details for Job 0

Status: SUCCEEDED
Completed Stages: 2

▸ Event Timeline
▾ DAG Visualization



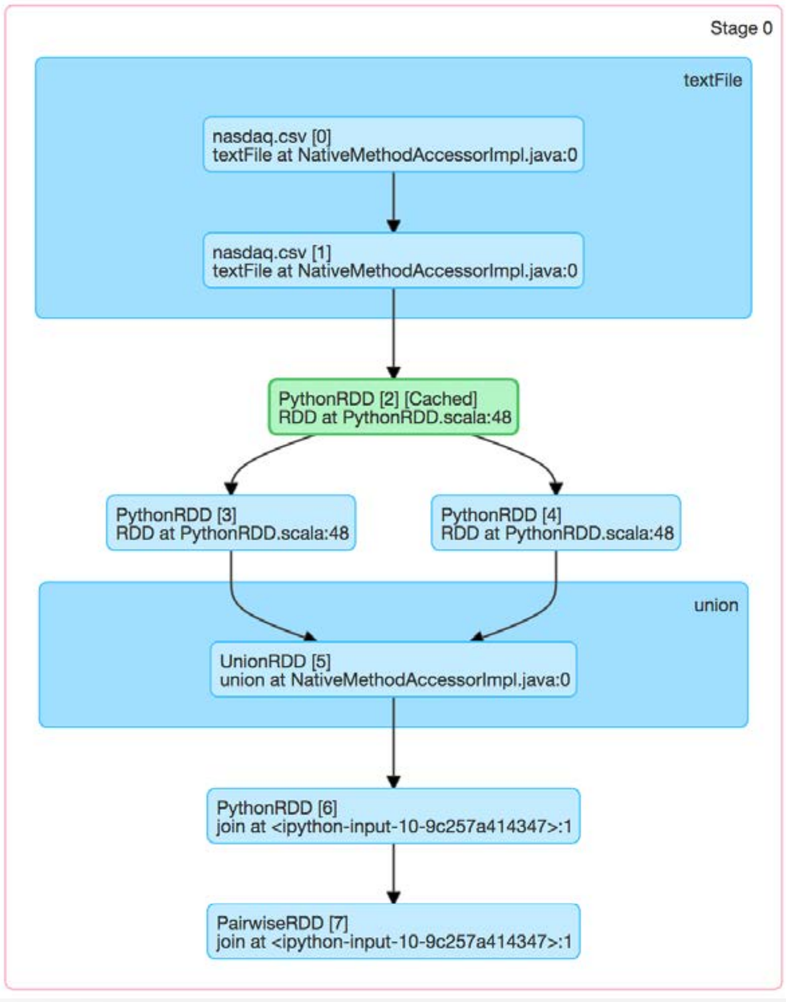
Completed Stages (2)

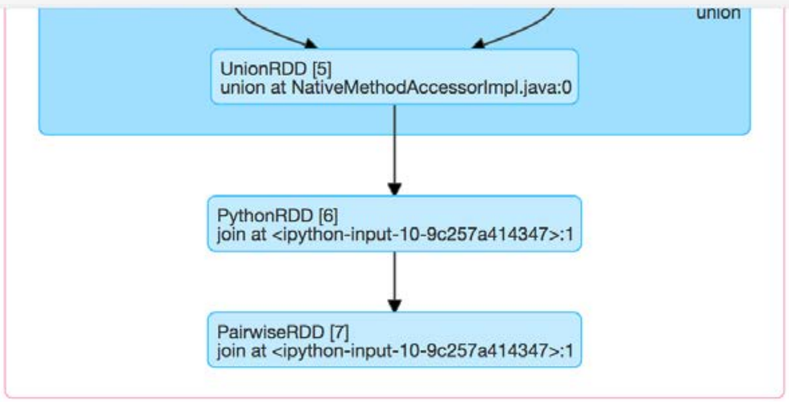
Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	top at <ipython-input-15-0b58397eb290>:1 +details	2017/07/23 22:16:15	0,1 s	4/4			7.3 KB	
0	join at <ipython-input-10-9c257a414347>:1 +details	2017/07/23 22:16:13	2 s	4/4	23.0 KB			7.3 KB

Details for Stage 0 (Attempt 0)

Total Time Across All Tasks: 7 s
Locality Level Summary: Process local: 4
Input Size / Records: 23.0 KB / 153
Shuffle Write: 7.3 KB / 32

▾ DAG Visualization





▸ Show Additional Metrics
▸ Event Timeline

Summary Metrics for 4 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	2 s	2 s	2 s	2 s	2 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Input Size / Records	3.1 KB / 7	3.1 KB / 7	5.6 KB / 69	11.1 KB / 70	11.1 KB / 70
Shuffle Write Size / Records	1849.0 B / 8	1856.0 B / 8	1865.0 B / 8	1867.0 B / 8	1867.0 B / 8

▾ Aggregated Metrics by Executor

Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Write Size / Records	Blacklisted
driver	77.88.19.2:53626	8 s	4	0	0	4	23.0 KB / 153	7.3 KB / 32	0

Tasks (4)

Index ▲	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	0	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2017/07/23 22:16:13	2 s		11.1 KB / 70	9 ms	1865.0 B / 8	
1	1	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2017/07/23 22:16:13	2 s		3.1 KB / 7	20 ms	1856.0 B / 8	
2	2	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2017/07/23 22:16:13	2 s		3.1 KB / 7	36 ms	1849.0 B / 8	
3	3	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2017/07/23 22:16:13	2 s		5.6 KB / 69	10 ms	1867.0 B / 8	

Storage

RDDs

RDD Name	Storage Level	Cached Partitions	Fraction Cached	Size in Memory	Size on Disk
PythonRDD	Memory Serialized 1x Replicated	2	100%	6.3 KB	0.0 B

RDD Storage Info for PythonRDD

Storage Level: Memory Serialized 1x Replicated
Cached Partitions: 2
Total Partitions: 2
Memory Size: 6.3 KB
Disk Size: 0.0 B

Data Distribution on 1 Executors

Host	On Heap Memory Usage	Off Heap Memory Usage	Disk Usage
77.88.19.2:53626	6.3 KB (366.3 MB Remaining)	0.0 B (0.0 B Remaining)	0.0 B

2 Partitions

Block Name ▲	Storage Level	Size in Memory	Size on Disk	Executors
rdd_2_0	Memory Serialized 1x Replicated	3.1 KB	0.0 B	77.88.19.2:53626
rdd_2_1	Memory Serialized 1x Replicated	3.1 KB	0.0 B	77.88.19.2:53626

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	Blacklisted
Active(1)	3	29.9 KB / 384.1 MB	0.0 B	4	0	0	8	8	**8 s (0 ms)**	23.5 KB	0.0 B	7.4 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(1)	3	29.9 KB / 384.1 MB	0.0 B	4	0	0	8	8	**8 s (0 ms)**	23.5 KB	0.0 B	7.4 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	Thread Dump
driver	77.88.19.2:53626	Active	3	29.9 KB / 384.1 MB	0.0 B	4	0	0	8	8	**8 s (0 ms)**	23.5 KB	0.0 B	7.4 KB	Thread Dump

Showing 1 to 1 of 1 entries

Previous 1 Next

Summary

- You have learned how to open and use Spark UI
- In the next course of the specialization you will learn how to use the interface to optimize your application performance

BigDATAteam