

# COMPSCI 516: Homework 2

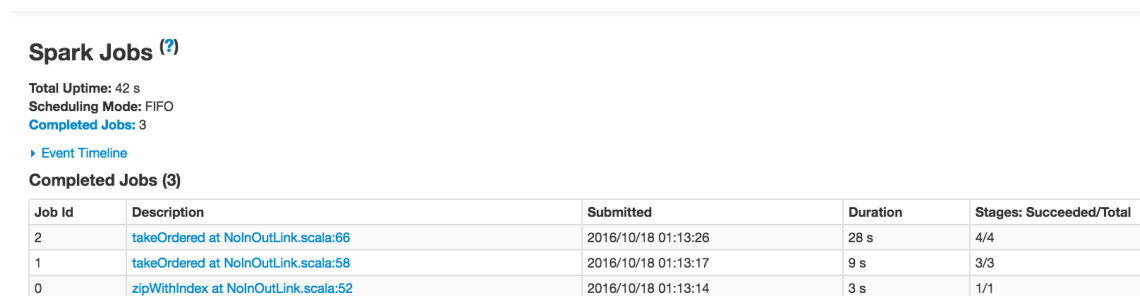
Akbota Anuarbek

## 0.1 NoInOutLink Application:

Runtime: 42s

There are 3 jobs corresponding to 3 different actions: zipWithIndex, takeOrdered (corresponds to no outlinks) and takeOrdered (corresponds to no inlinks).

Duration of each job, number of completed stages in each job and other general information about the application is given in the table 1.



The screenshot shows the 'Spark Jobs' section of a web interface. It includes summary statistics: 'Total Uptime: 42 s', 'Scheduling Mode: FIFO', and 'Completed Jobs: 3'. A link for 'Event Timeline' is also present. Below this is a table titled 'Completed Jobs (3)' with the following data:

| Job Id | Description  | Submitted           | Duration | Stages: Succeeded/Total |
|--------|--|---------------------|----------|-------------------------|
| 2      | <a href="#">takeOrdered at NoInOutLink.scala:66</a>  | 2016/10/18 01:13:26 | 28 s     | 4/4                     |
| 1      | <a href="#">takeOrdered at NoInOutLink.scala:58</a>  | 2016/10/18 01:13:17 | 9 s      | 3/3                     |
| 0      | <a href="#">zipWithIndex at NoInOutLink.scala:52</a> | 2016/10/18 01:13:14 | 3 s      | 1/1                     |

Figure 1: NoInOutLink output

We can see that Job 0 has only one stage (stage 0) and it took only 3s to run. We can also see that only three executors were employed in this stage, and the table 2 below shows the number of tasks implemented by each executor.

Table 3 shows stages of the job 1. Interestingly, in some of the stages only some number of executors are given tasks, whereas in some stages all 10 executors are given tasks. We can see how stage 2 is implemented by all 10 executors from table 4.

Job 2 was the one that took the longest time, and it consists of 4 stages, details of each stage can be seen from table 5. As we can see from the table the stage 5 take the longest to complete. I think

#### Aggregated Metrics by Executor

| Executor ID ▲ | Address   | Task Time | Total Tasks |
|---------------|---|-----------|-------------|
| 4             | ip-172-31-6-93.us-west-2.compute.internal:43101 | 3 s       | 4           |
| 5             | ip-172-31-6-87.us-west-2.compute.internal:55746 | 3 s       | 2           |
| 6             | ip-172-31-6-85.us-west-2.compute.internal:60567 | 3 s       | 3           |

Figure 2: NoInOutLink output

#### Details for Job 1

Status: SUCCEEDED

Completed Stages: 3

► Event Timeline

► DAG Visualization

##### Completed Stages (3)

| Stage Id | Description   |                          | Submitted           | Duration |
|----------|---|--------------------------|---------------------|----------|
| 3        | <a href="#">takeOrdered at NoInOutLink.scala:58</a> | <a href="#">+details</a> | 2016/10/18 01:13:23 | 3 s      |
| 2        | <a href="#">map at NoInOutLink.scala:41</a>         | <a href="#">+details</a> | 2016/10/18 01:13:17 | 6 s      |
| 1        | <a href="#">map at NoInOutLink.scala:52</a>         | <a href="#">+details</a> | 2016/10/18 01:13:17 | 3 s      |

Figure 3: NoInOutLink output

the reason is because in order to find distinct entries in the RDD task needs shuffle, which might be quite expensive.

## 0.2 PageRank Application:

Runtime: 35min

There are 4 jobs corresponding to 4 different actions: zipWithIndex, count, sum and takeOrdered. Summeries of the application are given in the table 6.

As illustrated in the table 6, first two jobs (job 0 and 1) took no more than 4 seconds each. Each of them had only one stage. In each of the stages 0 and 1 tasks were assigned only to the executors with ids 4, 5, 6. As an example table 7 shows how tasks were assigned in stage 1.

The longest job was the sum, it took 33min to run. Details and stages of the job are illustrated in the table 8. Job was divided into 22 stages, each of which corresponds to either map or mapValues and one corresponds to the sum.

Again at different stages different number of executors was used. Tables 9 and 10 illustrate task assignments for stages 3 and 7 respectively.

#### Aggregated Metrics by Executor

| Executor ID ▲ | Address   | Task Time | Total Tasks |
|---------------|---|-----------|-------------|
| 0             | ip-172-31-6-90.us-west-2.compute.internal:54223 | 3 s       | 1           |
| 1             | ip-172-31-6-84.us-west-2.compute.internal:44485 | 3 s       | 1           |
| 2             | ip-172-31-6-86.us-west-2.compute.internal:43649 | 2 s       | 1           |
| 4             | ip-172-31-6-93.us-west-2.compute.internal:43101 | 2 s       | 1           |
| 5             | ip-172-31-6-87.us-west-2.compute.internal:55746 | 1 s       | 1           |
| 6             | ip-172-31-6-85.us-west-2.compute.internal:60567 | 2 s       | 1           |
| 7             | ip-172-31-6-92.us-west-2.compute.internal:36673 | 3 s       | 1           |
| 8             | ip-172-31-6-89.us-west-2.compute.internal:60247 | 3 s       | 1           |
| 9             | ip-172-31-6-88.us-west-2.compute.internal:57411 | 4 s       | 2           |

Figure 4: NoInOutLink output

#### Details for Job 2

Status: SUCCEEDED

Completed Stages: 4

▶ Event Timeline

▶ DAG Visualization

##### Completed Stages (4)

| Stage Id | Description   |                          | Submitted           | Duration |
|----------|---|--------------------------|---------------------|----------|
| 7        | <a href="#">takeOrdered at NoInOutLink.scala:66</a> | <a href="#">+details</a> | 2016/10/18 01:13:52 | 2 s      |
| 6        | <a href="#">map at NoInOutLink.scala:48</a>         | <a href="#">+details</a> | 2016/10/18 01:13:50 | 2 s      |
| 5        | <a href="#">distinct at NoInOutLink.scala:47</a>    | <a href="#">+details</a> | 2016/10/18 01:13:26 | 24 s     |
| 4        | <a href="#">map at NoInOutLink.scala:52</a>         | <a href="#">+details</a> | 2016/10/18 01:13:26 | 2 s      |

Figure 5: NoInOutLink output

And finally we can see that the final job corresponding to takeOrdered had also relatively long time of running 1.6min. This is again because ordering task needs shuffle, because we are selecting the 10 links with the highest Page Rank and it is not enough to simply do the task in one node.

## Spark Jobs (?)

Total Uptime: 35 min  
Scheduling Mode: FIFO  
Completed Jobs: 4

► Event Timeline

### Completed Jobs (4)

| Job Id | Description                                       | Submitted           | Duration | Stages: Succeeded/Total       |
|--------|---|---------------------|----------|-------------------------------|
| 3      | <a href="#">takeOrdered at PageRank.scala:116</a> | 2016/10/18 02:57:51 | 1.6 min  | 2/2 (22 skipped)              |
| 2      | <a href="#">sum at PageRank.scala:103</a>         | 2016/10/18 02:24:53 | 33 min   | 22/22 (10 failed) (1 skipped) |
| 1      | <a href="#">count at PageRank.scala:46</a>        | 2016/10/18 02:24:51 | 0.9 s    | 1/1                           |
| 0      | <a href="#">zipWithIndex at PageRank.scala:41</a> | 2016/10/18 02:24:47 | 4 s      | 1/1                           |

Figure 6: PageRank output

## Tasks

| Index ▲ | ID | Attempt | Status  | Locality Level | Executor ID / Host                            |
|---------|----|---------|---------|----------------|---|
| 0       | 9  | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal |
| 1       | 10 | 0       | SUCCESS | NODE_LOCAL     | 4 / ip-172-31-6-93.us-west-2.compute.internal |
| 2       | 11 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal |
| 3       | 12 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal |
| 4       | 13 | 0       | SUCCESS | NODE_LOCAL     | 4 / ip-172-31-6-93.us-west-2.compute.internal |
| 5       | 14 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal |
| 6       | 15 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal |
| 7       | 16 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal |
| 8       | 17 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal |
| 9       | 18 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal |

Figure 7: PageRank output

### Completed Stages (22)

| Stage Id     | Description                                    | Submitted                     | Duration | Tasks: Succeeded/Total      | Input     | Output | Shuffle Read | Shuffle Write |
|--------------|--|-------------------------------|----------|-----------------------------|-----------|--------|--------------|---------------|
| 24           | <a href="#">sum at PageRank.scala:103</a>      | <details> 2016/10/18 02:56:23 | 1.5 min  | <div><div></div>10/10</div> |           |        | 5.3 GB       |               |
| 23 (retry 1) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:54:28 | 1.9 min  | <div><div></div>10/10</div> |           |        | 5.5 GB       | 270.4 MB      |
| 21 (retry 4) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:53:17 | 1.2 min  | <div><div></div>1/1</div>   |           |        | 508.6 MB     | 27.0 MB       |
| 20 (retry 5) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:48:29 | 7 s      | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 80.8 MB       |
| 19 (retry 3) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:46:14 | 1.2 min  | <div><div></div>1/1</div>   |           |        | 455.9 MB     | 27.1 MB       |
| 18 (retry 4) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:45:36 | 10 s     | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 81.0 MB       |
| 17 (retry 1) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:42:36 | 1.0 min  | <div><div></div>1/1</div>   |           |        | 402.2 MB     | 27.1 MB       |
| 15 (retry 3) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:42:49 | 49 s     | <div><div></div>1/1</div>   |           |        | 348.1 MB     | 27.1 MB       |
| 16 (retry 4) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:38:28 | 7 s      | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 81.0 MB       |
| 13 (retry 2) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:38:28 | 53 s     | <div><div></div>1/1</div>   |           |        | 293.8 MB     | 27.0 MB       |
| 8 (retry 3)  | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:35:25 | 6 s      | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 81.0 MB       |
| 11 (retry 2) | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:32:16 | 1.2 min  | <div><div></div>4/4</div>   |           |        | 958.1 MB     | 108.2 MB      |
| 14 (retry 3) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:30:55 | 19 s     | <div><div></div>2/2</div>   | 201.9 MB  |        |              | 157.3 MB      |
| 12 (retry 1) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:30:55 | 9 s      | <div><div></div>2/2</div>   | 201.9 MB  |        |              | 157.3 MB      |
| 2 (retry 1)  | <a href="#">mapValues at PageRank.scala:87</a> | <details> 2016/10/18 02:30:55 | 2 s      | <div><div></div>2/2</div>   | 21.3 MB   |        |              | 5.9 MB        |
| 22 (retry 2) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:29:29 | 10 s     | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 79.5 MB       |
| 10 (retry 1) | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:27:22 | 9 s      | <div><div></div>1/1</div>   | 100.9 MB  |        |              | 78.6 MB       |
| 7            | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:25:42 | 1.2 min  | <div><div></div>10/10</div> |           |        | 1311.6 MB    | 270.2 MB      |
| 5            | <a href="#">map at PageRank.scala:77</a>       | <details> 2016/10/18 02:25:11 | 29 s     | <div><div></div>10/10</div> |           |        | 829.7 MB     | 241.2 MB      |
| 6            | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:24:53 | 18 s     | <div><div></div>10/10</div> | 1009.4 MB |        |              | 800.4 MB      |
| 4            | <a href="#">mapValues at PageRank.scala:36</a> | <details> 2016/10/18 02:24:53 | 19 s     | <div><div></div>10/10</div> | 1009.4 MB |        |              | 800.4 MB      |
| 3            | <a href="#">mapValues at PageRank.scala:54</a> | <details> 2016/10/18 02:24:53 | 3 s      | <div><div></div>10/10</div> | 106.4 MB  |        |              | 29.4 MB       |

Figure 8: PageRank output

| 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     | 39 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:24:55 | 1.0 s    |         | 10.6 MB (hadoop) / 566288 | 9 ms       | 2.9 MB / 566288              |        |
| 1     | 40 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:24:55 | 1.0 s    |         | 10.6 MB (hadoop) / 578787 | 8 ms       | 3.0 MB / 578787              |        |
| 2     | 41 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal | 2016/10/18 02:24:55 | 1.0 s    | 19 ms   | 10.6 MB (hadoop) / 578783 | 9 ms       | 3.0 MB / 578783              |        |
| 3     | 42 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal | 2016/10/18 02:24:55 | 1.0 s    | 19 ms   | 10.6 MB (hadoop) / 577314 | 9 ms       | 3.0 MB / 577314              |        |
| 4     | 45 | 0       | SUCCESS | NODE_LOCAL     | 4 / ip-172-31-6-93.us-west-2.compute.internal | 2016/10/18 02:24:56 | 2 s      | 81 ms   | 10.6 MB (hadoop) / 615961 | 10 ms      | 3.2 MB / 615961              |        |
| 5     | 46 | 0       | SUCCESS | NODE_LOCAL     | 4 / ip-172-31-6-93.us-west-2.compute.internal | 2016/10/18 02:24:56 | 2 s      | 81 ms   | 10.6 MB (hadoop) / 507837 | 0.1 s      | 2.6 MB / 507837              |        |
| 6     | 47 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:24:56 | 1 s      | 18 ms   | 10.6 MB (hadoop) / 578879 | 51 ms      | 3.0 MB / 578879              |        |
| 7     | 48 | 0       | SUCCESS | NODE_LOCAL     | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:24:56 | 1 s      | 18 ms   | 10.6 MB (hadoop) / 585616 | 47 ms      | 3.0 MB / 585616              |        |
| 8     | 51 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal | 2016/10/18 02:24:56 | 0.9 s    |         | 10.6 MB (hadoop) / 538518 | 9 ms       | 2.8 MB / 538518              |        |
| 9     | 52 | 0       | SUCCESS | NODE_LOCAL     | 5 / ip-172-31-6-87.us-west-2.compute.internal | 2016/10/18 02:24:56 | 1.0 s    |         | 10.6 MB (hadoop) / 590885 | 18 ms      | 3.0 MB / 590885              |        |

Figure 9: PageRank output

| Index | ID  | Attempt | Status  | Locality Level | Executor ID / Host                            | Launch Time         | Duration | GC Time | Shuffle Read Size / Records | Write Time | Shuffle Write Size / Records | Errors |
|-------|-----|---------|---------|----------------|---|---------------------|----------|---------|-----------------------------|------------|------------------------------|--------|
| 0     | 138 | 0       | SUCCESS | PROCESS_LOCAL  | 3 / ip-172-31-6-91.us-west-2.compute.internal | 2016/10/18 02:25:42 | 1.1 min  | 2 s     | 130.9 MB / 4963483          | 30 ms      | 27.1 MB / 191226             |        |
| 1     | 140 | 0       | SUCCESS | ANY            | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:25:46 | 1.0 min  | 9 s     | 131.0 MB / 4968203          | 27 ms      | 27.0 MB / 1911309            |        |
| 2     | 150 | 0       | SUCCESS | NODE_LOCAL     | 2 / ip-172-31-6-86.us-west-2.compute.internal | 2016/10/18 02:25:50 | 1.1 min  | 15 s    | 131.4 MB / 4961961          | 36 ms      | 27.0 MB / 1913295            |        |
| 3     | 143 | 0       | SUCCESS | NODE_LOCAL     | 7 / ip-172-31-6-92.us-west-2.compute.internal | 2016/10/18 02:25:47 | 47 s     | 7 s     | 131.4 MB / 4972372          | 0.1 s      | 27.1 MB / 1914736            |        |
| 4     | 141 | 0       | SUCCESS | NODE_LOCAL     | 0 / ip-172-31-6-90.us-west-2.compute.internal | 2016/10/18 02:25:46 | 47 s     | 5 s     | 131.3 MB / 4969146          | 26 ms      | 27.0 MB / 1909861            |        |
| 5     | 142 | 0       | SUCCESS | NODE_LOCAL     | 1 / ip-172-31-6-84.us-west-2.compute.internal | 2016/10/18 02:25:46 | 50 s     | 17 s    | 131.2 MB / 4967482          | 27 ms      | 27.0 MB / 1909525            |        |
| 6     | 146 | 0       | SUCCESS | NODE_LOCAL     | 1 / ip-172-31-6-84.us-west-2.compute.internal | 2016/10/18 02:25:48 | 57 s     | 17 s    | 130.9 MB / 4969219          | 27 ms      | 27.0 MB / 1907377            |        |
| 7     | 154 | 0       | SUCCESS | ANY            | 5 / ip-172-31-6-87.us-west-2.compute.internal | 2016/10/18 02:25:54 | 45 s     | 5 s     | 131.6 MB / 4971695          | 28 ms      | 27.1 MB / 1914715            |        |
| 8     | 151 | 0       | SUCCESS | NODE_LOCAL     | 2 / ip-172-31-6-86.us-west-2.compute.internal | 2016/10/18 02:25:50 | 60 s     | 15 s    | 130.6 MB / 4967096          | 57 ms      | 26.9 MB / 1906301            |        |
| 9     | 139 | 0       | SUCCESS | PROCESS_LOCAL  | 6 / ip-172-31-6-85.us-west-2.compute.internal | 2016/10/18 02:25:46 | 1.0 min  | 9 s     | 131.3 MB / 4968543          | 27 ms      | 27.0 MB / 1913816            |        |

Figure 10: PageRank output