Spark under the hood

02807 Computational Tools for Data Science

Today

- The anatomy of a Spark job
- Advanced Spark functions
    - Window functions
    - Sketches in Spark
    - The physical plan
    - Collecting/writing data
    - Broadcast joins
- Project 3

The anatomy of a Spark job

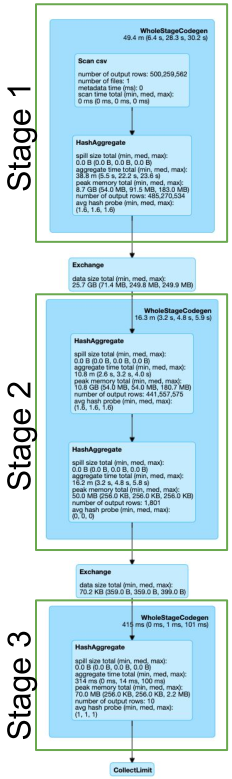
When you run an action in a cell in Jupyter:

- A **job** is submitted to Spark
- Data is split into logical **partitions**
- An execution **plan** is computed dividing your program into **stages**
- **Tasks** are executed on workers

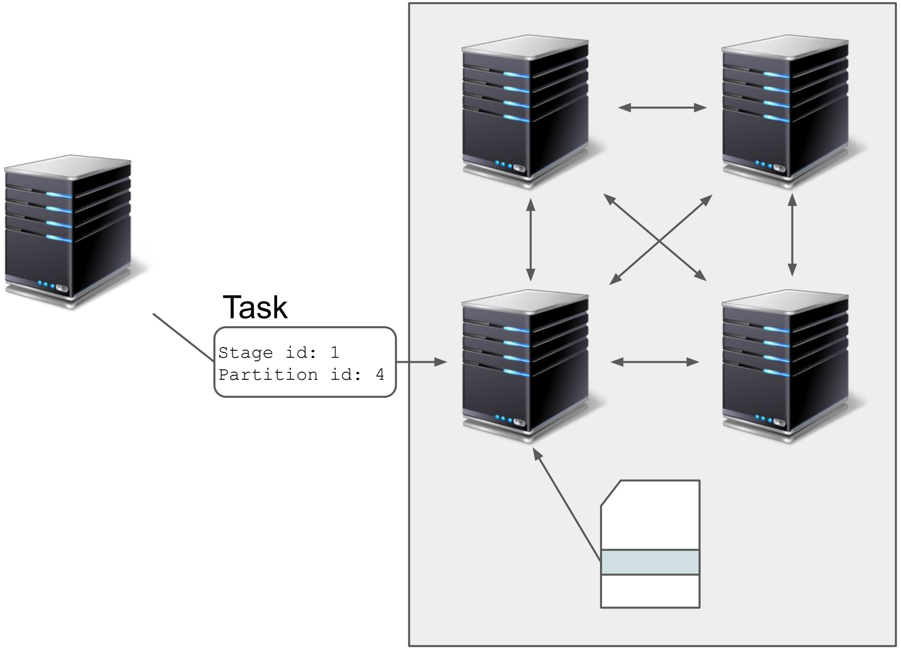
A task is executing one stage on one partition of the data.

A job consists of more tasks than there are partitions, and there should be more partitions than cores in your system.

The anatomy of a Spark job



```
WholeStageCodegen
49.4 m (6.4 s, 28.3 s, 30.2 s)

Scan csv

number of output rows: 500,259,562
number of files: 1
metadata time (ms): 0
scan time total (min, med, max):
0 ms (0 ms, 0 ms, 0 ms)

HashAggregate

spill size total (min, med, max):
0.0 B (0.0 B, 0.0 B, 0.0 B)
aggregate time total (min, med, max):
38.8 m (5.5 s, 22.2 s, 23.6 s)
peak memory total (min, med, max):
8.7 GB (54.0 MB, 91.5 MB, 183.0 MB)
number of output rows: 485,270,534
avg hash probe (min, med, max):
(1.6, 1.6, 1.6)

Exchange

data size total (min, med, max):
25.7 GB (71.4 MB, 249.8 MB, 249.9 MB)

WholeStageCodegen
16.3 m (3.2 s, 4.8 s, 5.9 s)

HashAggregate

spill size total (min, med, max):
0.0 B (0.0 B, 0.0 B, 0.0 B)
aggregate time total (min, med, max):
10.8 m (2.6 s, 3.2 s, 4.0 s)
peak memory total (min, med, max):
10.8 GB (54.0 MB, 54.0 MB, 180.7 MB)
number of output rows: 441,557,575
avg hash probe (min, med, max):
(1.6, 1.6, 1.6)

HashAggregate

spill size total (min, med, max):
0.0 B (0.0 B, 0.0 B, 0.0 B)
aggregate time total (min, med, max):
16.2 m (3.2 s, 4.8 s, 5.8 s)
peak memory total (min, med, max):
50.0 MB (256.0 KB, 256.0 KB, 256.0 KB)
number of output rows: 1,801
avg hash probe (min, med, max):
(0, 0, 0)

Exchange

data size total (min, med, max):
70.2 KB (359.0 B, 359.0 B, 399.0 B)

WholeStageCodegen
415 ms (0 ms, 1 ms, 101 ms)

HashAggregate

spill size total (min, med, max):
0.0 B (0.0 B, 0.0 B, 0.0 B)
aggregate time total (min, med, max):
314 ms (0 ms, 14 ms, 100 ms)
peak memory total (min, med, max):
70.0 MB (256.0 KB, 256.0 KB, 2.2 MB)
number of output rows: 10
avg hash probe (min, med, max):
(1, 1, 1)

CollectLimit
```



```
Stage id: 1
Partition id: 4
```