cloudera

Petabytes of Data - How to Use R at Scale

Jeff Fletcher
jfletcher@cloudera.com
https://www.linkedin.com/in/jefffletcher/

AGENDA

- WHAT PROBLEM ARE WE SOLVING?
- HOW IS IT SOLVED OUTSIDE OF R?
- QUICK COMPONENTS OVERVIEW
- HOW DOES THIS WORK WITH R?
- USEFUL THINGS TO KNOW
- GETTING STARTED

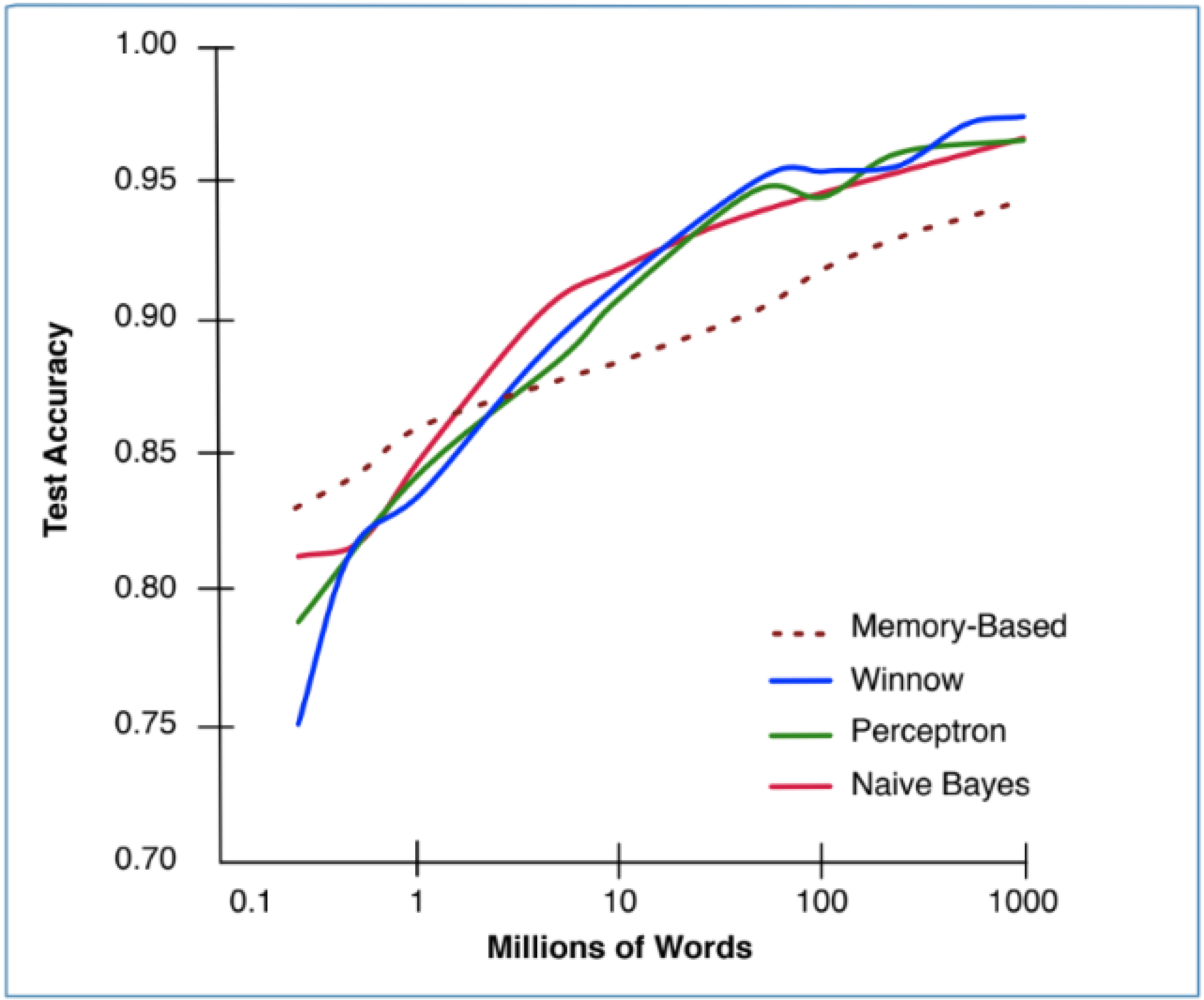
WHAT PROBLEM ARE WE SOLVING?

WHAT PROBLEM ARE WE SOLVING?

Your MacBook, while very cool and shiny is too small for the enormous amounts of data used by actual enterprise companies.

WHAT PROBLEM ARE WE SOLVING?

"It's not who has the best algorithms that wins. It's who has the most data."
[Banko and Brill, 2001]

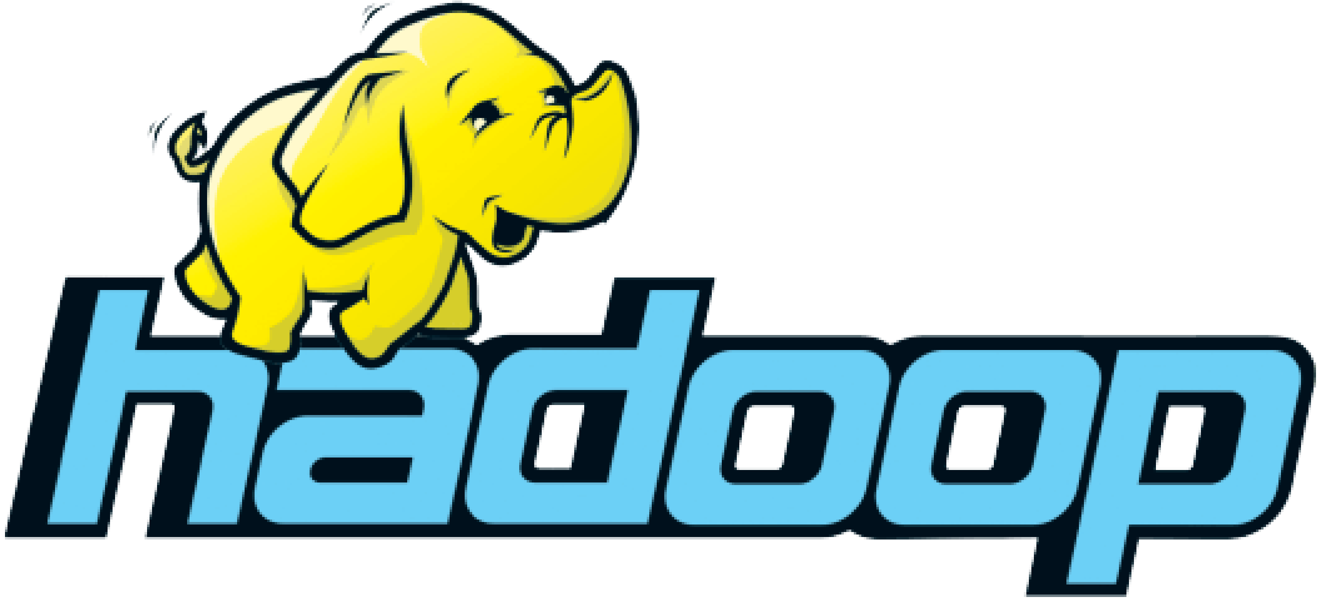


HOW IS THIS SOLVED OUTSIDE OF R?

HOW IS THIS SOLVED OUTSIDE OF R?

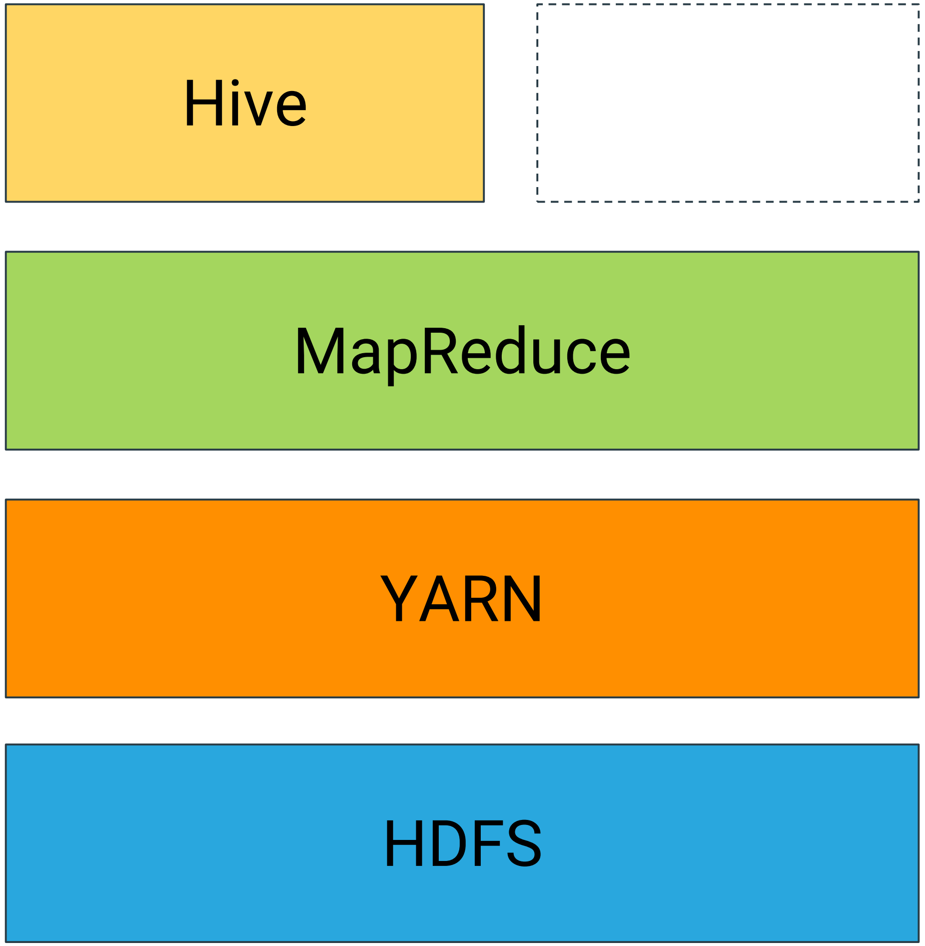
One of the many Jeff Dean facts: Jeff Dean's PIN is the last 4 digits of pi.

The rest are funny-ish, but a good way to kill 20 mins on a many person webex.



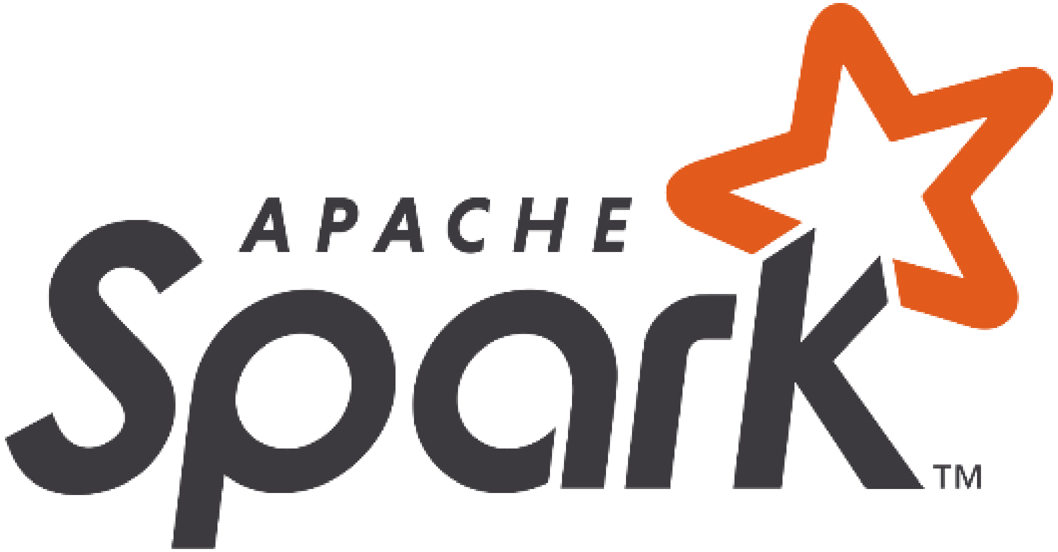
HADOOP STACK

Our template says I should put some related text here. How about - this layout was approved by Piet Mondrian.



HOW IS THIS SOLVED OUTSIDE OF R?

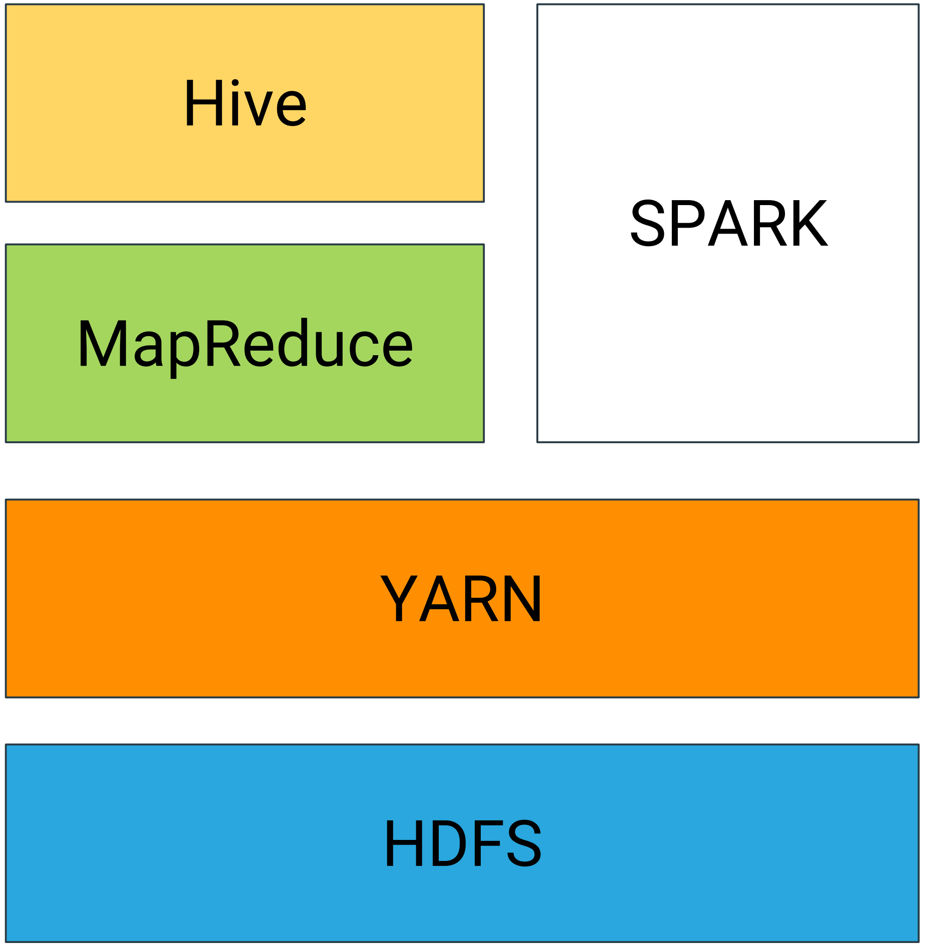
A Dutch painter and theoretician who is regarded as one of the greatest artists of the 20th century. Painty Piet I like to call him.



HADOOP STACK WITH SPARK

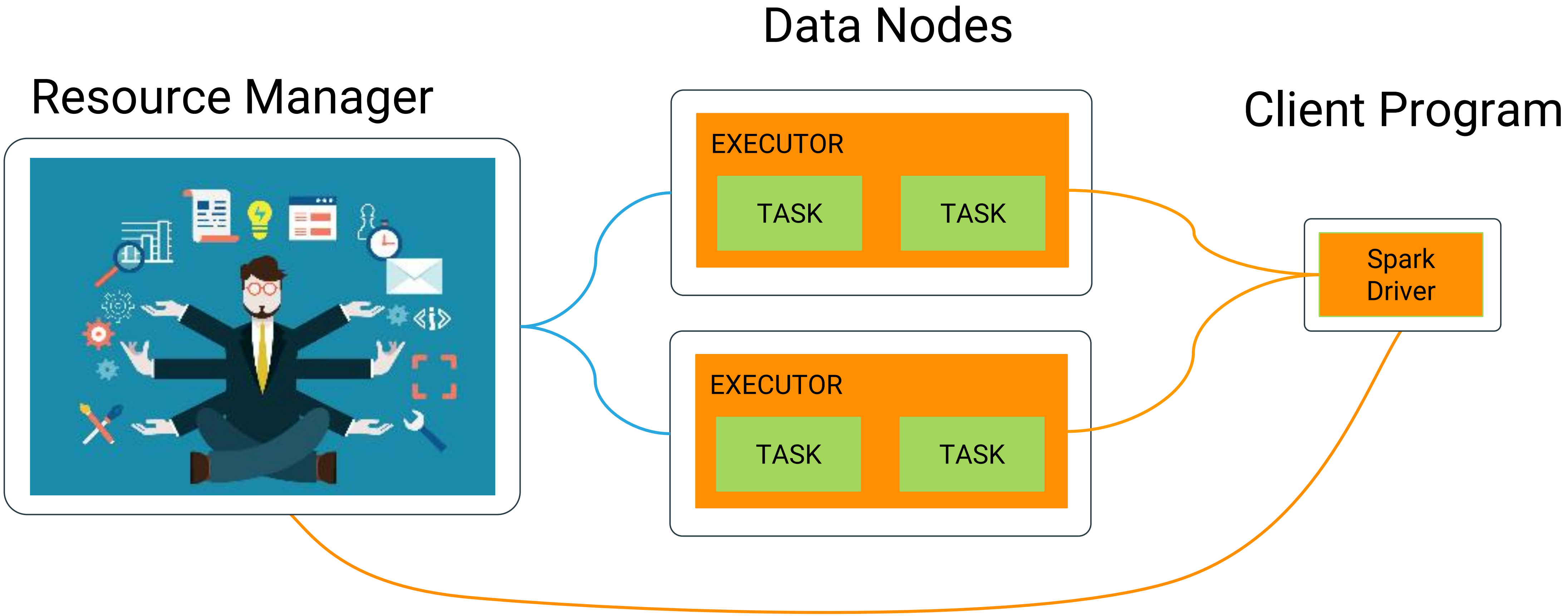
He proclaimed in 1914: Art is higher than reality and has no direct relation to reality…

According to Wikipedia anyway. I didn't hear him say it.



SPARK OVERVIEW

SPARK



Data Nodes

Resource Manager

Client Program

EXECUTOR

TASK

TASK

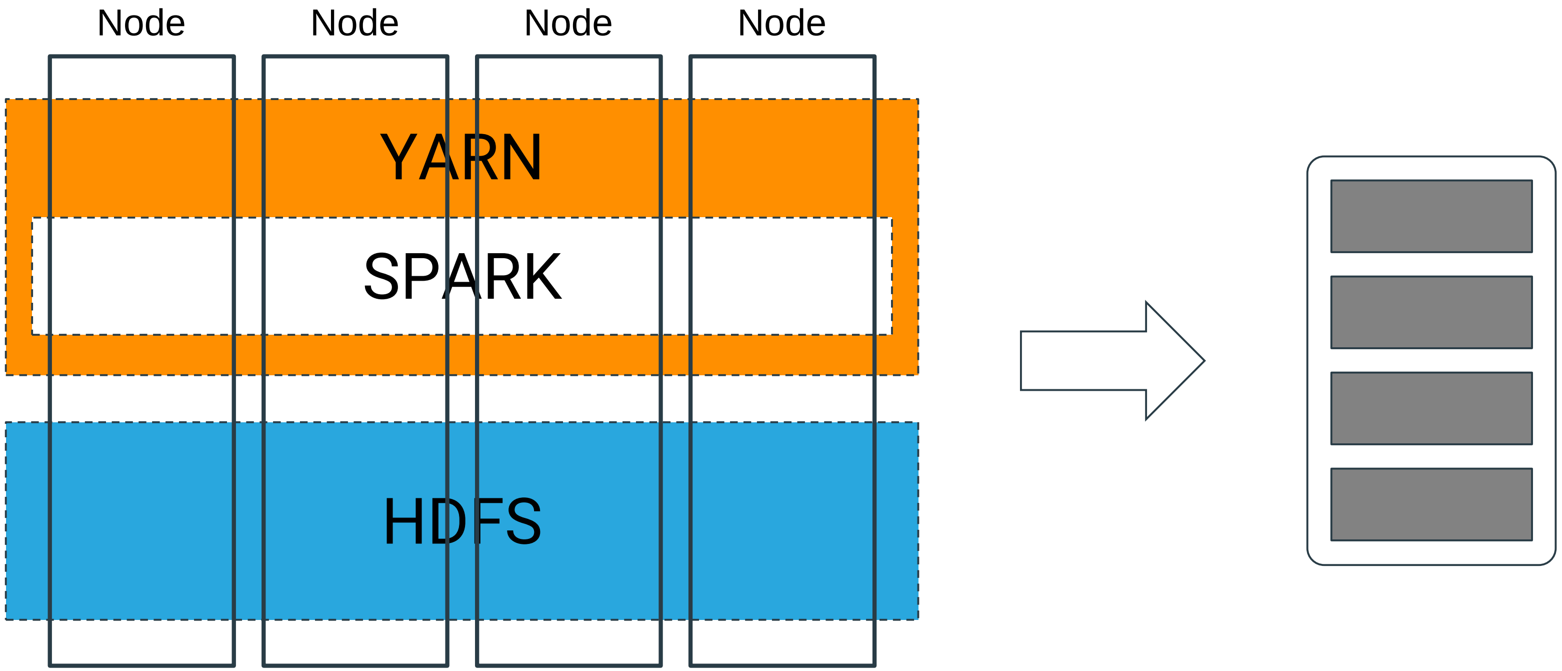
EXECUTOR

TASK

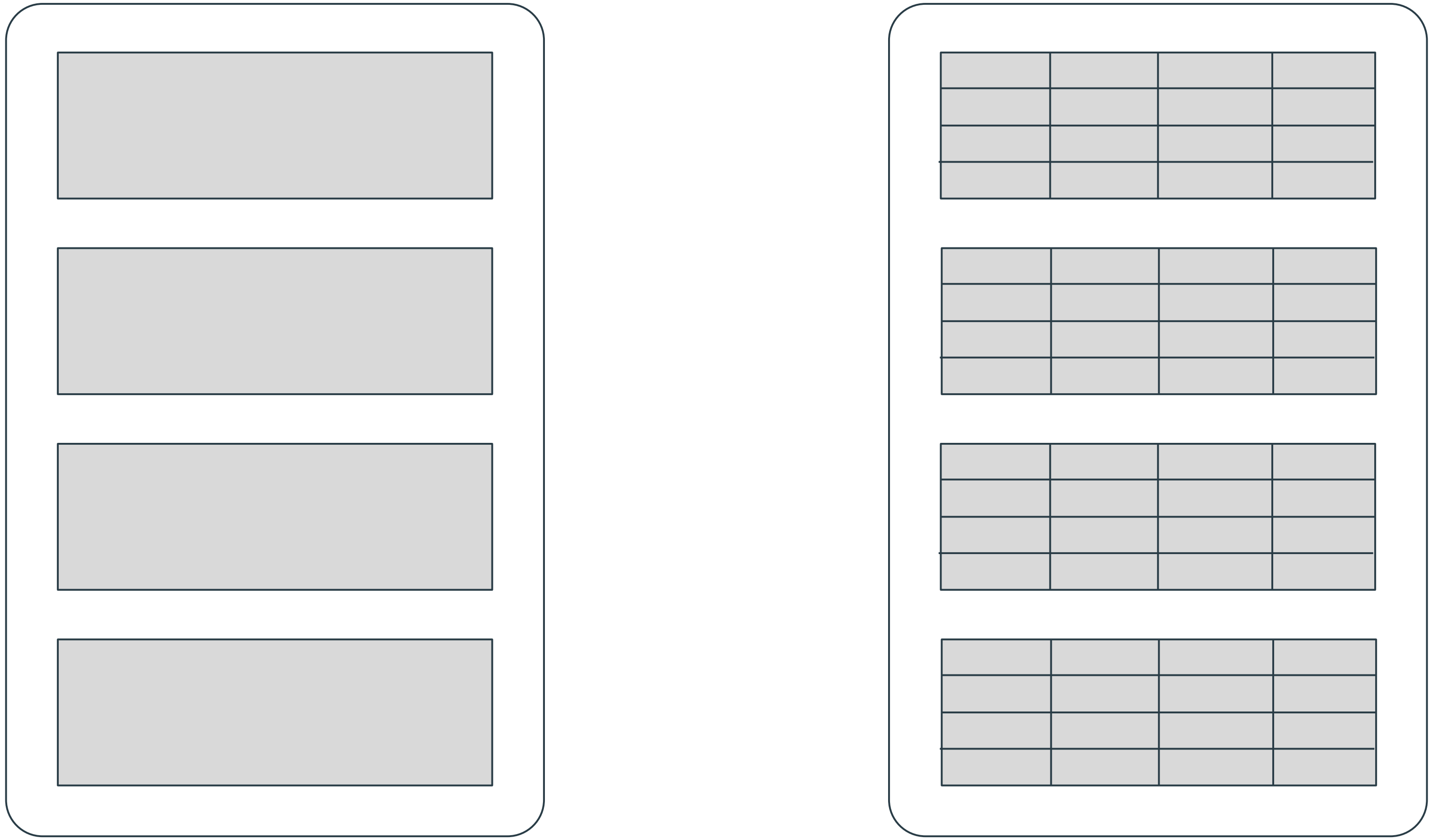
TASK

Spark Driver

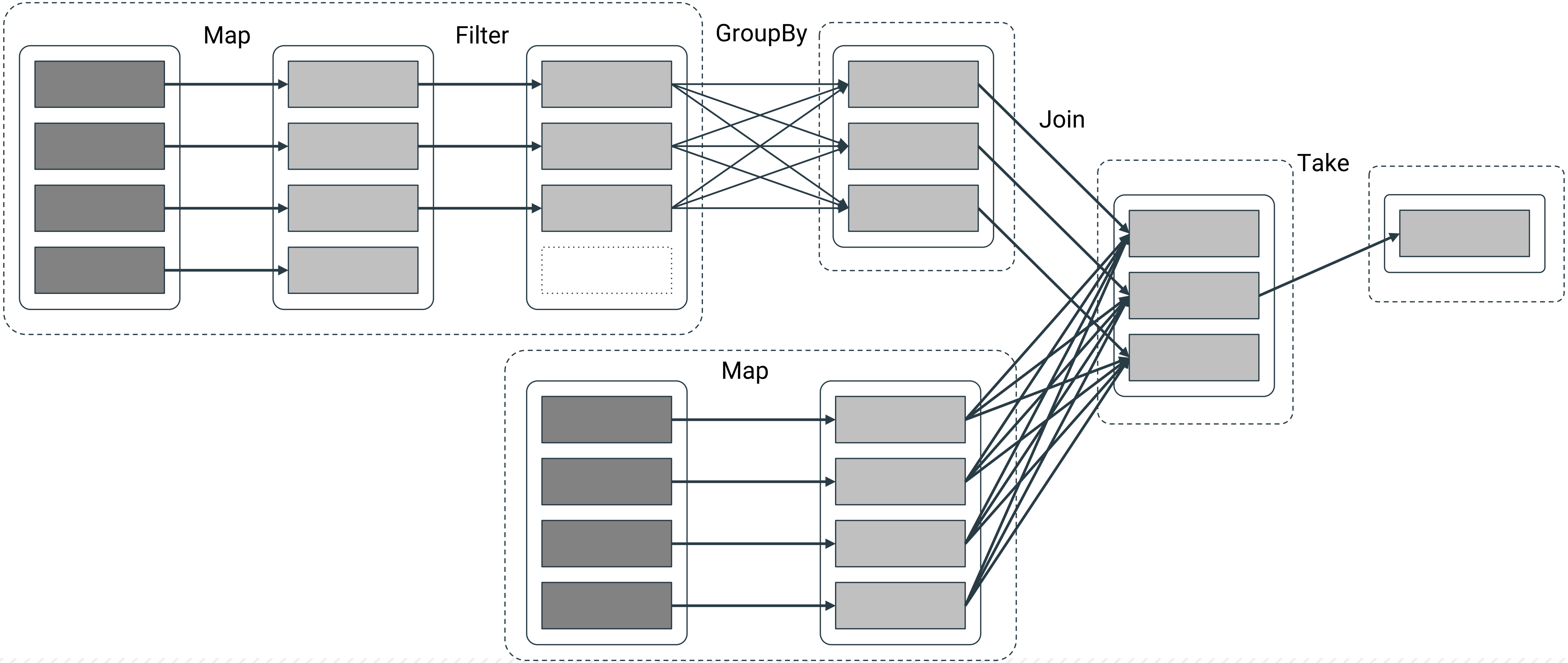
SPARK RDD



SPARK DATAFRAME

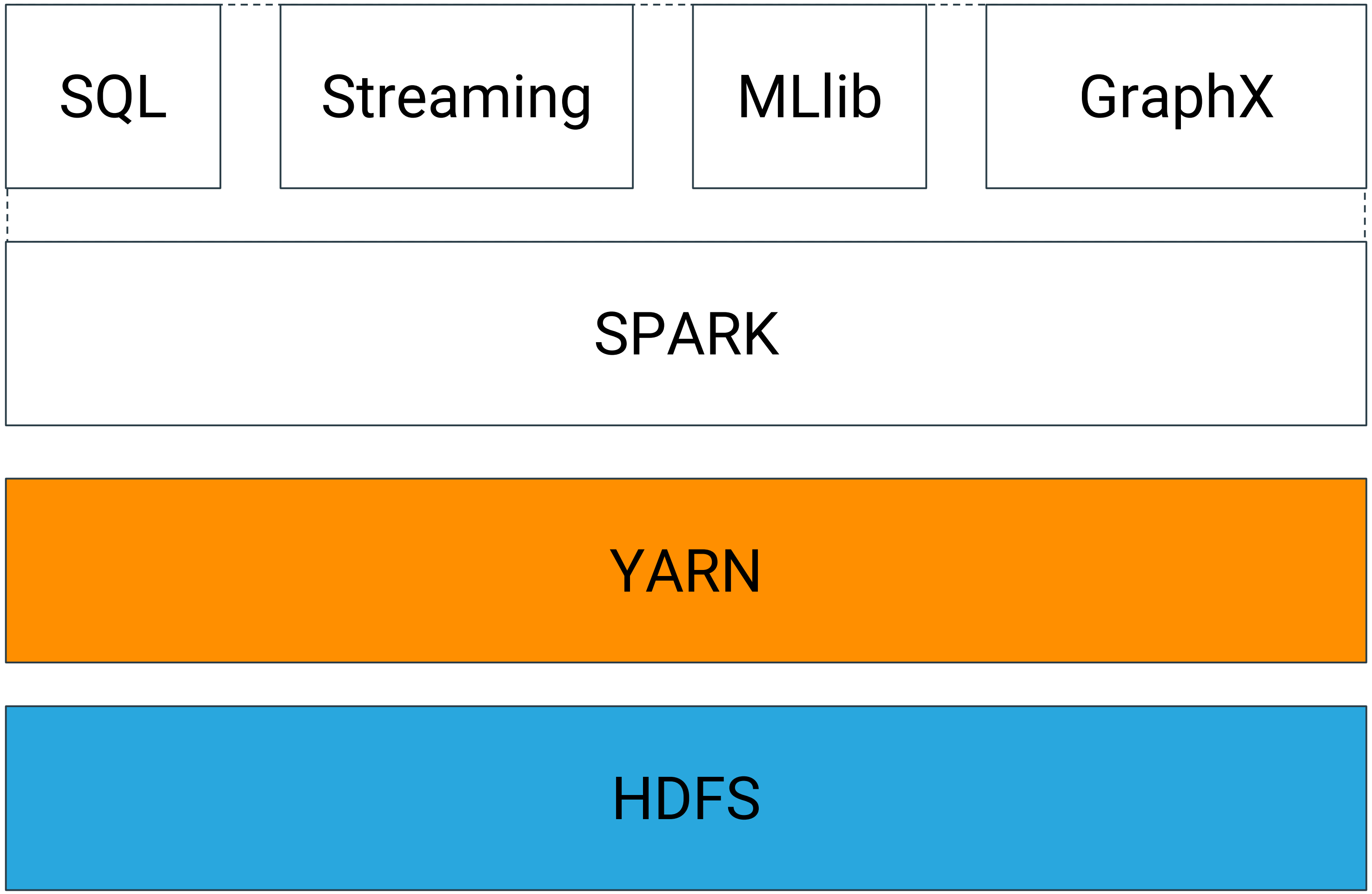


SPARK DATAFRAME OPERATIONS



SPARK LIBRARIES

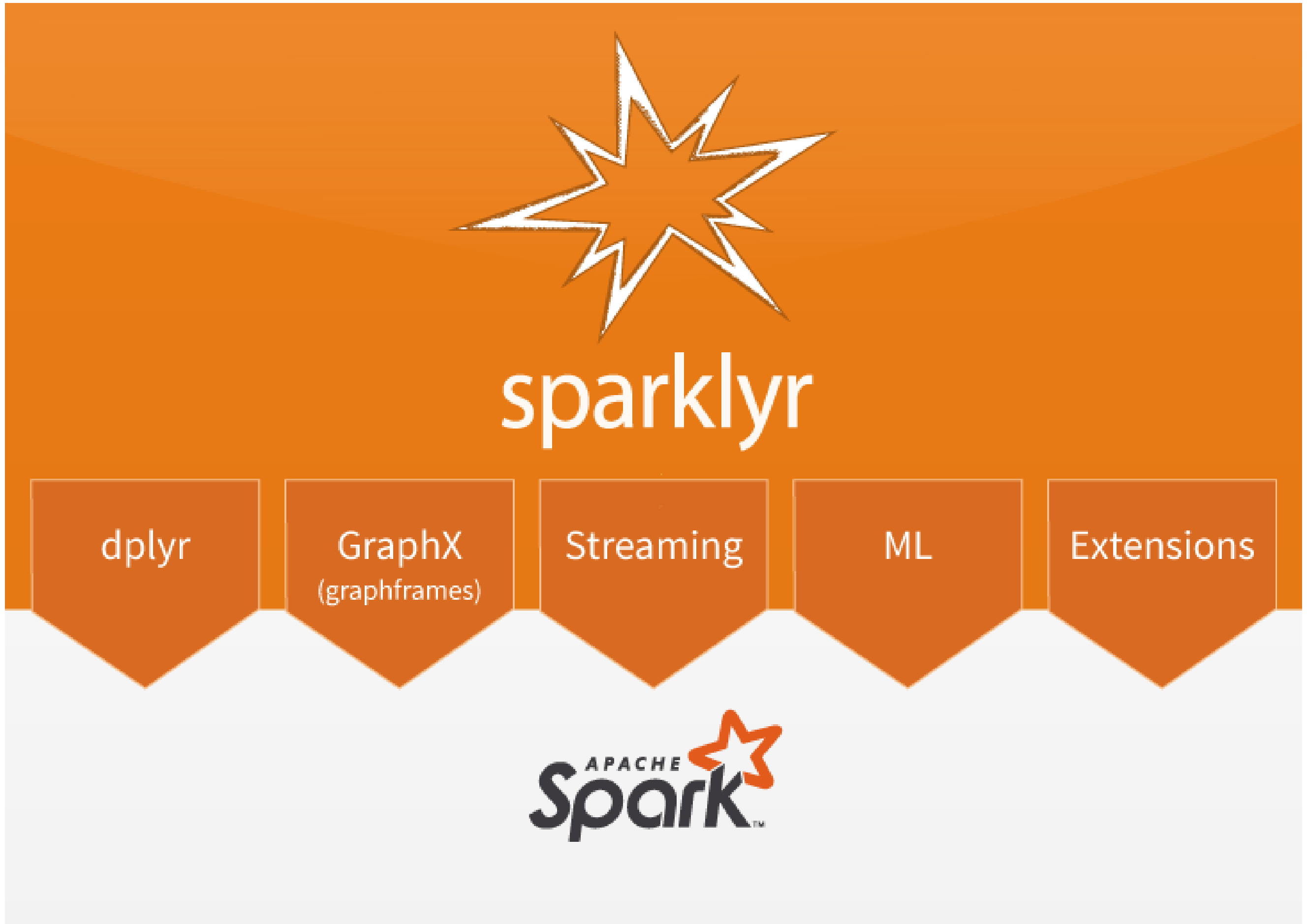
Apart from libraries, there are many books on Spark. Look for stuff by Holden Karau.



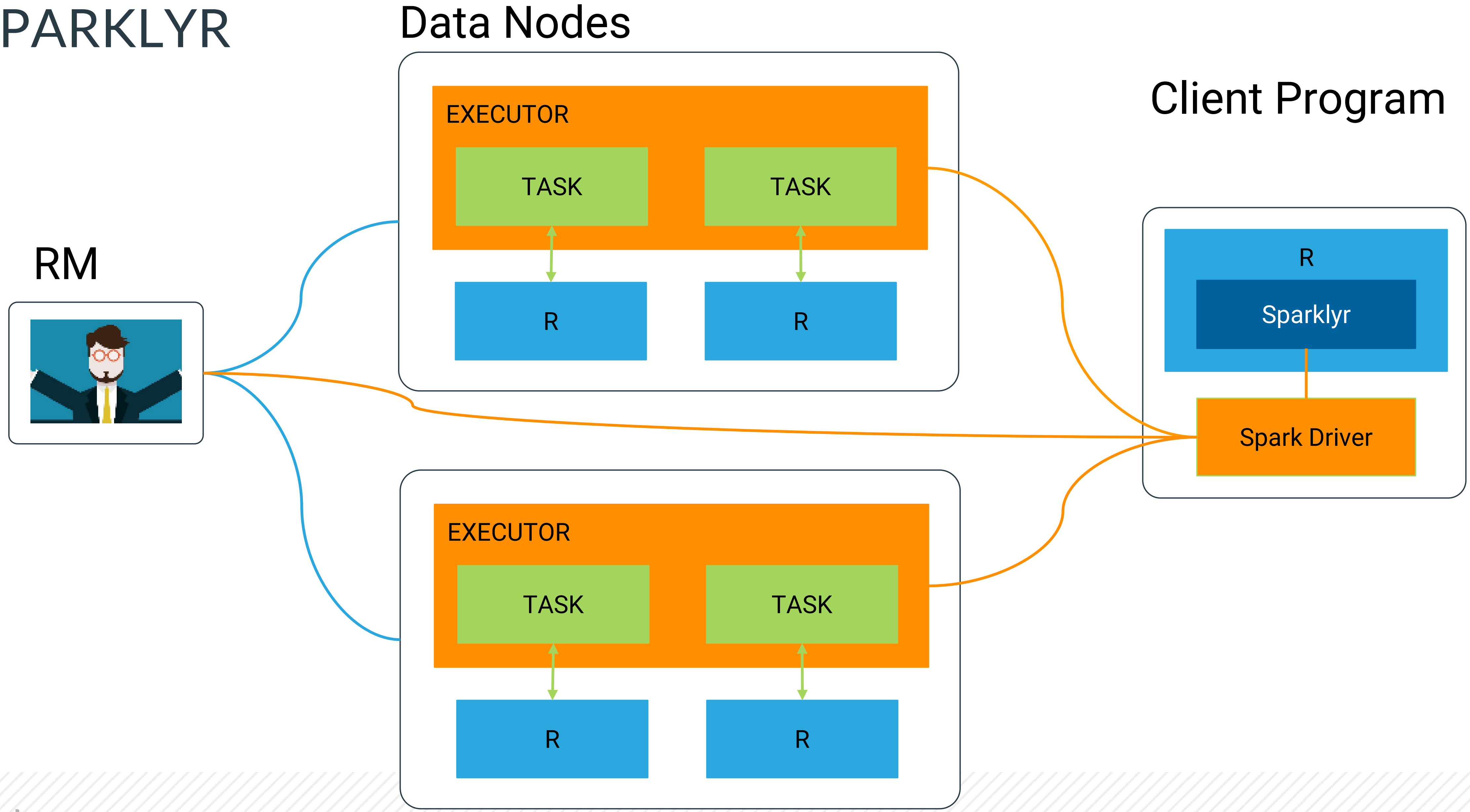
HOW DOES THIS WORK WITH R?

SPARKLYR

The silver medal goes to SparkR. Sparklyr 1.0.0 has most of the things you need and support for Arrow.



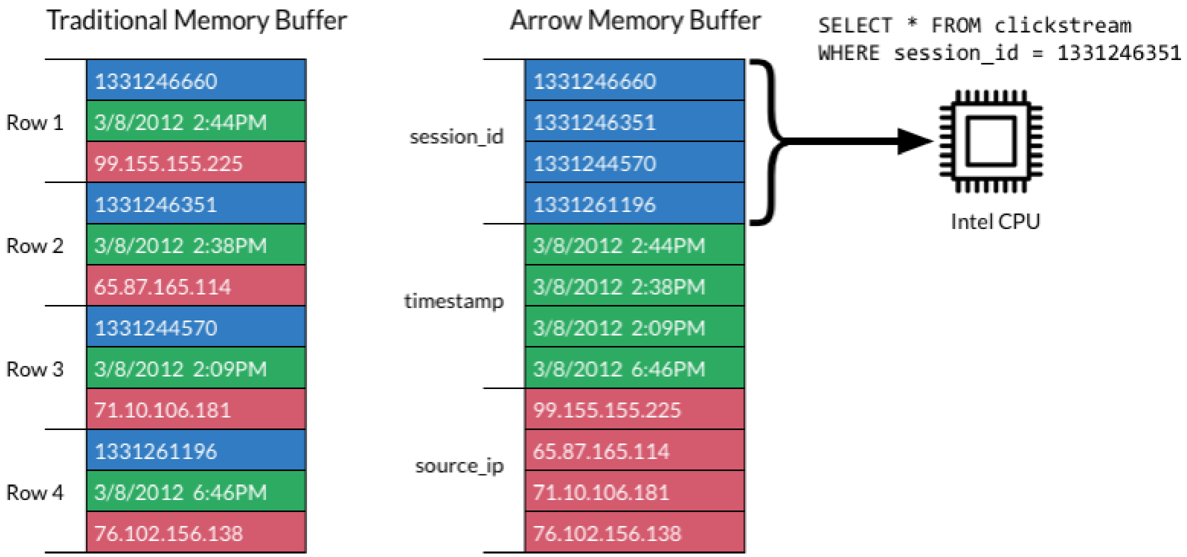
SPARKLYR


Data Nodes
EXECUTOR
TASK
TASK
R
R
Client Program
R
Sparklyr
Spark Driver
RM
EXECUTOR
TASK
TASK
R
R

APACHE ARROW

High speed in memory data processing for everybody!

	session_id	timestamp	source_ip
Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138



Traditional Memory Buffer

Row 1	1331246660
	3/8/2012 2:44PM
	99.155.155.225
Row 2	1331246351
	3/8/2012 2:38PM
	65.87.165.114
Row 3	1331244570
	3/8/2012 2:09PM
	71.10.106.181
Row 4	1331261196
	3/8/2012 6:46PM
	76.102.156.138

Arrow Memory Buffer

session_id	1331246660
	1331246351
	1331244570
	1331261196
timestamp	3/8/2012 2:44PM
	3/8/2012 2:38PM
	3/8/2012 2:09PM
	3/8/2012 6:46PM
source_ip	99.155.155.225
	65.87.165.114
	71.10.106.181
	76.102.156.138

```
SELECT * FROM clickstream
WHERE session_id = 1331246351
```

USEFUL THINGS TO KNOW

LAZY EVALUATION

This catches people out when working on the command line. Entered does not mean executed.

OVERLOADING THE DRIVER

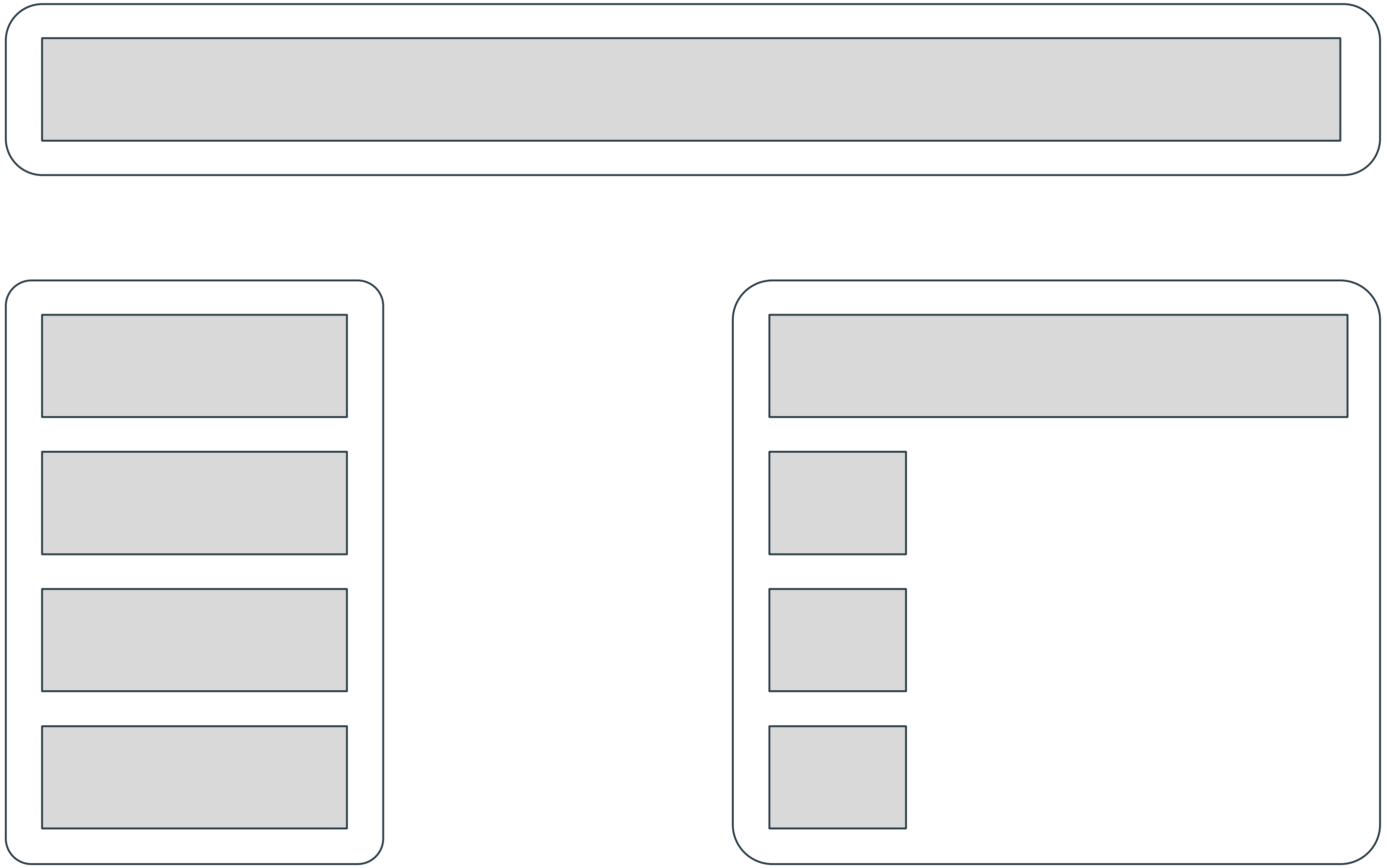
I see this often when talking with data science teams. It’s part of the standard operating procedure for “small data” data science work.

AVOID GROUPBYKEY

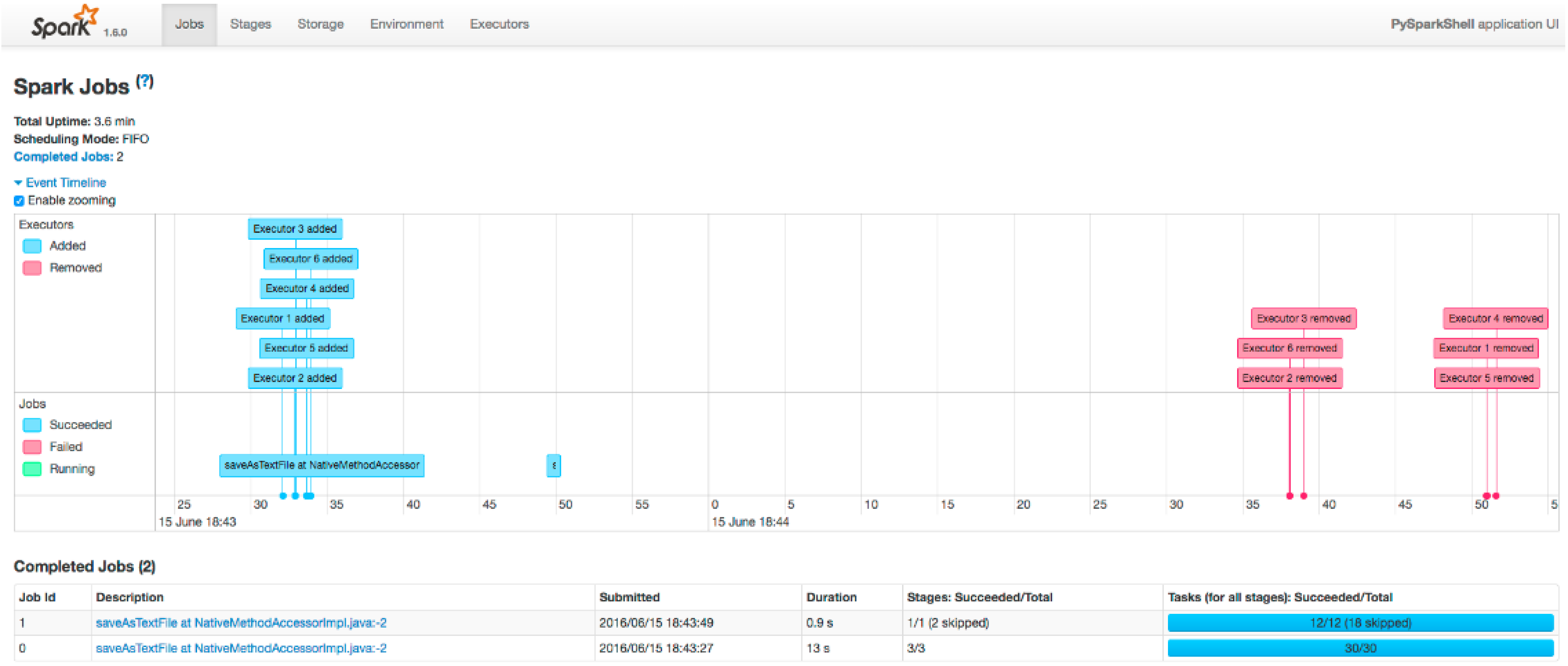
Use ReduceByKey. All the cool kids are doing it.

CHECK FOR DATA SKEW

I did not spend a lot of time drawing this. But Edward Tufte would be proud. Very favourable data to ink ratio.



USE THE SPARK UI



Spark Jobs (?)

Total Uptime: 3.6 min
Scheduling Mode: FIFO
Completed Jobs: 2

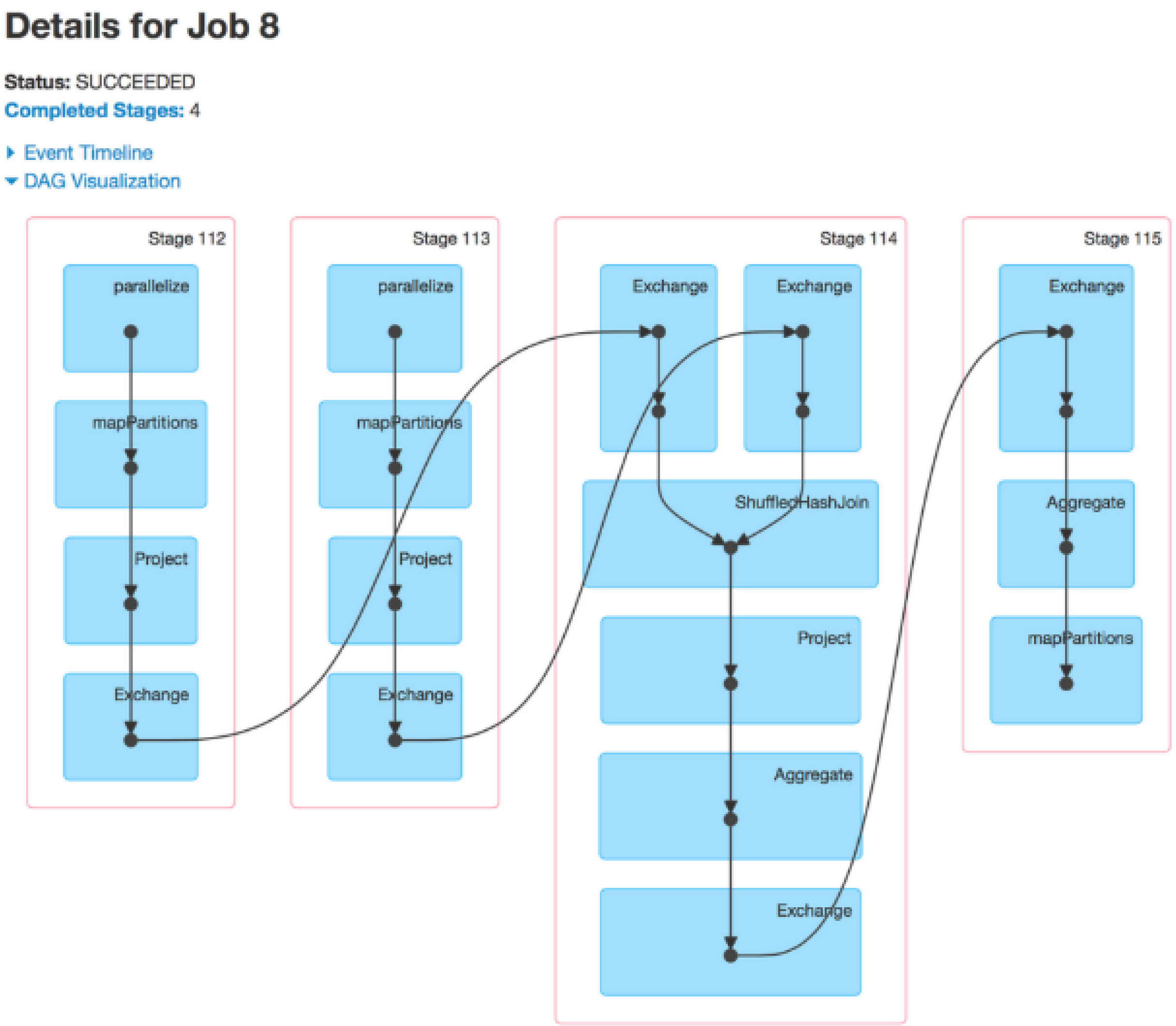
▾ Event Timeline
☑ Enable zooming

Completed Jobs (2)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	saveAsTextFile at NativeMethodAccessorImpl.java:-2	2016/06/15 18:43:49	0.9 s	1/1 (2 skipped)	12/12 (18 skipped)
0	saveAsTextFile at NativeMethodAccessorImpl.java:-2	2016/06/15 18:43:27	13 s	3/3	30/30

DAG

Directed Acyclic Graph - Use the full title if you want to sound smarter while talking data scientists.



Details for Job 8

Status: SUCCEEDED

Completed Stages: 4

▸ Event Timeline

▾ DAG Visualization

CHECK OUT IMPLYR

```
delay <- flights_tbl %>%
  select(tailnum, distance, arr_delay) %>%
  group_by(tailnum) %>%
  summarise(count = n(), dist = mean(distance), delay = mean(arr_delay)) %>%
  filter(count > 20L, dist < 2000L, !is.na(delay)) %>%
  arrange(delay, dist, count) %>%
  collect()
```

GETTING STARTED

GETTING STARTED

Implyr

https://blog.cloudera.com/blog/2017/07/implyr-r-interface-for-apache-impala/

https://github.com/ianmcook/implyr

Sparklyr

https://blog.cloudera.com/blog/2016/09/introducing-sparklyr-an-r-interface-for-apache-spark/
https://spark.rstudio.com/

Spark

https://www.cloudera.com/documentation/enterprise/5-16-x/PDF/cloudera-spark.pdf
https://www.cloudera.com/products/open-source/apache-hadoop/apache-spark.html

THANK YOU

IMAGES USED

Most diagrams I either drew or took from Cloudera's documentation directly. The rest come from here:

https://www.apple.com/shop/buy-mac/macbook-pro/15-inch-space-gray-2.6ghz-6-core-512gb

http://web.cs.ucla.edu/classes/winter13/cs111/scribe/10c/

https://commons.wikimedia.org/wiki/File:Special_Edition_NYC_2015_-_DC_vs_Street_Fighter_(18357923460).jpg

https://commons.wikimedia.org/wiki/File:Overloaded_truck.jpg

https://www.publicdomainpictures.net/en/view-image.php?image=220900&picture=two-lazy-bears

https://www.goodfreephotos.com/other-photos/lots-of-keys.jpg.php