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?
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.
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

Resource Manager

Data Nodes

EXECUTOR

TASK

TASK

EXECUTOR

TASK

TASK

Client Program

Spark Driver
SPARK RDD
SPARK DATAFRAME
SPARK DATAFRAME OPERATIONS
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

Client Program

RM

EXECUTOR

TASK

TASK

EXECUTOR

TASK

TASK

R

R

R

R

Sparklyr

Spark Driver
APACHE ARROW

High speed in memory data processing for everybody!
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.
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

Executors
- Added
- Removed

Jobs
- Successes
- Failed
- Running

Completed Jobs (2)

<table>
<thead>
<tr>
<th>Job ID</th>
<th>Description</th>
<th>Submitted</th>
<th>Duration</th>
<th>Stages: Succeeded/Total</th>
<th>Tasks (for all stages): Succeeded/Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>saveAsTextFile at NativeMethodAccessorsImpl.java: 2</td>
<td>2016/06/15 18:43:49</td>
<td>0.5 s</td>
<td>1/1 (2 skipped)</td>
<td>12/14 (10 skipped)</td>
</tr>
<tr>
<td>0</td>
<td>saveAsTextFile at NativeMethodAccessorsImpl.java: 2</td>
<td>2016/06/15 18:43:27</td>
<td>13 s</td>
<td>3/3</td>
<td>30/30</td>
</tr>
</tbody>
</table>
DAG

Directed Acyclic Graph - Use the full title if you want to sound smarter while talking data scientists.
```
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
THANK YOU
IMAGES USED

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

http://web.cs.ucla.edu/classes/winter13/cs111/SCRIBE/10c/
https://commons.wikimedia.org/wiki/File:Overloaded_truck.jpg