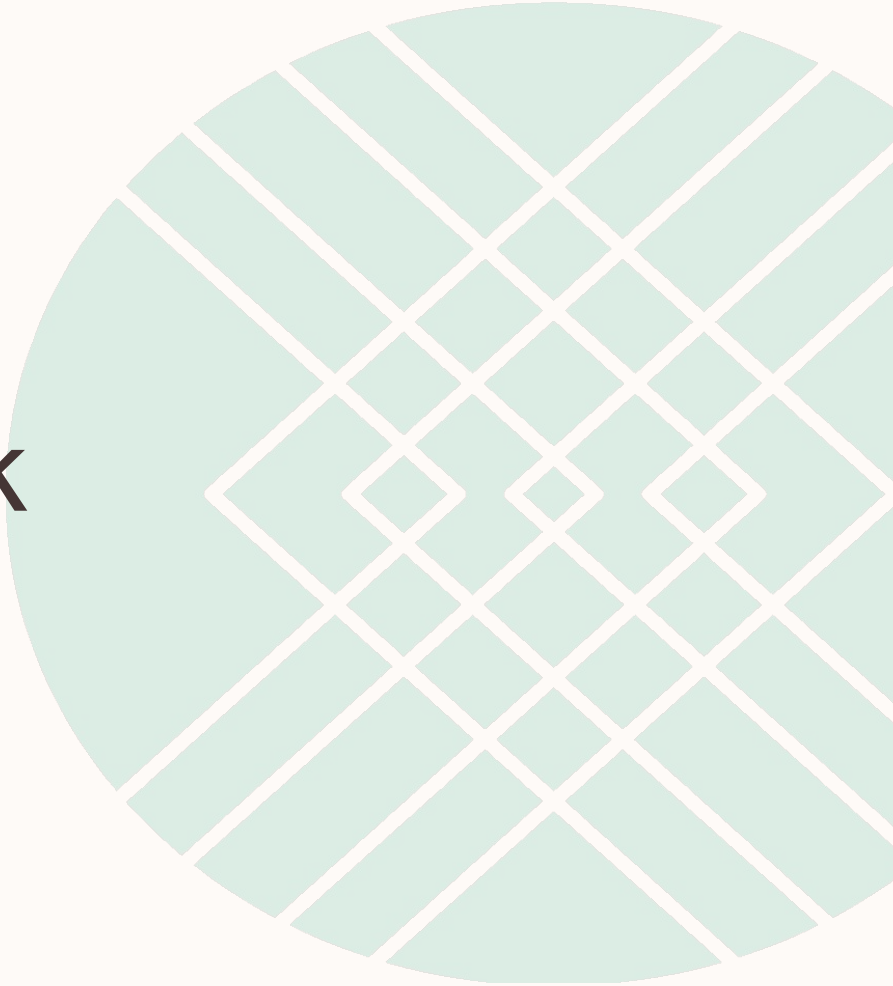
STITCH FIX

Optimizing Spark

Greg Novak



If you’ve thought about this at all, you won’t learn anything from me today

If you haven’t thought about this, you’ll learn a few principles to organize your thinking

Know what you want to measure

You don’t want to measure run times

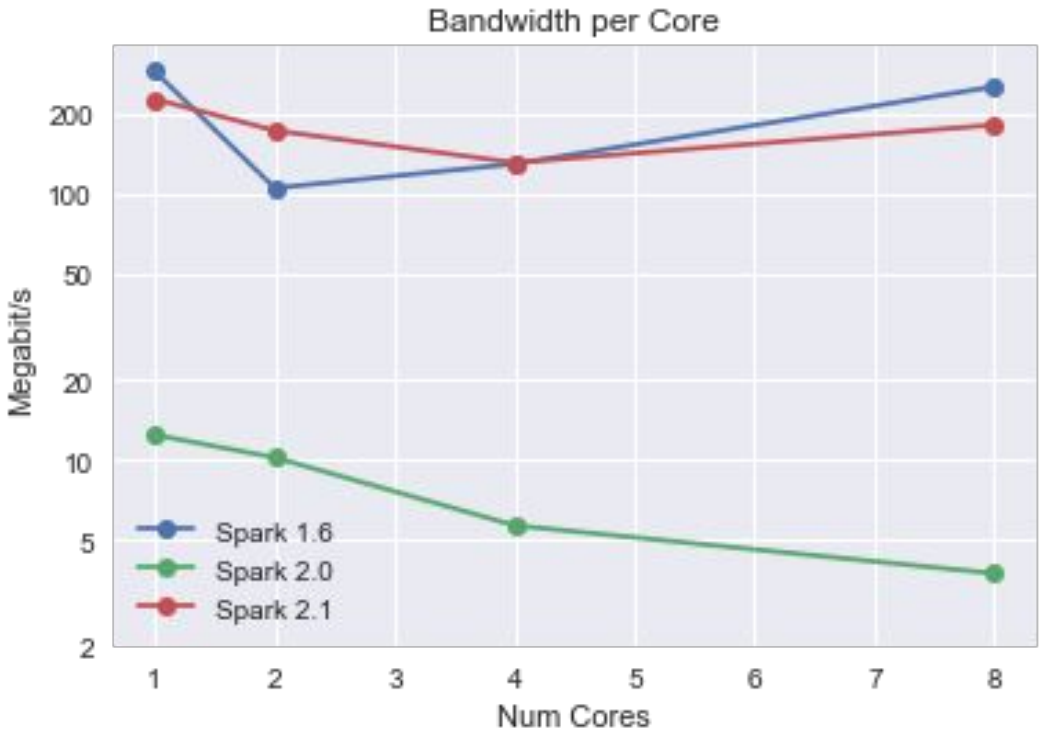
You want to measure effective performance of some machine characteristic: network bandwidth, file access latency, or CPU operations per second

You do this with carefully constructed data sets

To measure network bandwidth, construct a data set with the same number of files (so file access latency is constant) and do the same operation on it (so that cpu operations are constant) but force some extra data with variable size (e.g. random 1 byte ints vs. random 8 byte ints) to come along for the ride.

Then take difference of run times.

Case Study: Effective Network Bandwidth



Everything seemed to run slowly under Spark 2.0...

Latency and CPU performance looked fine

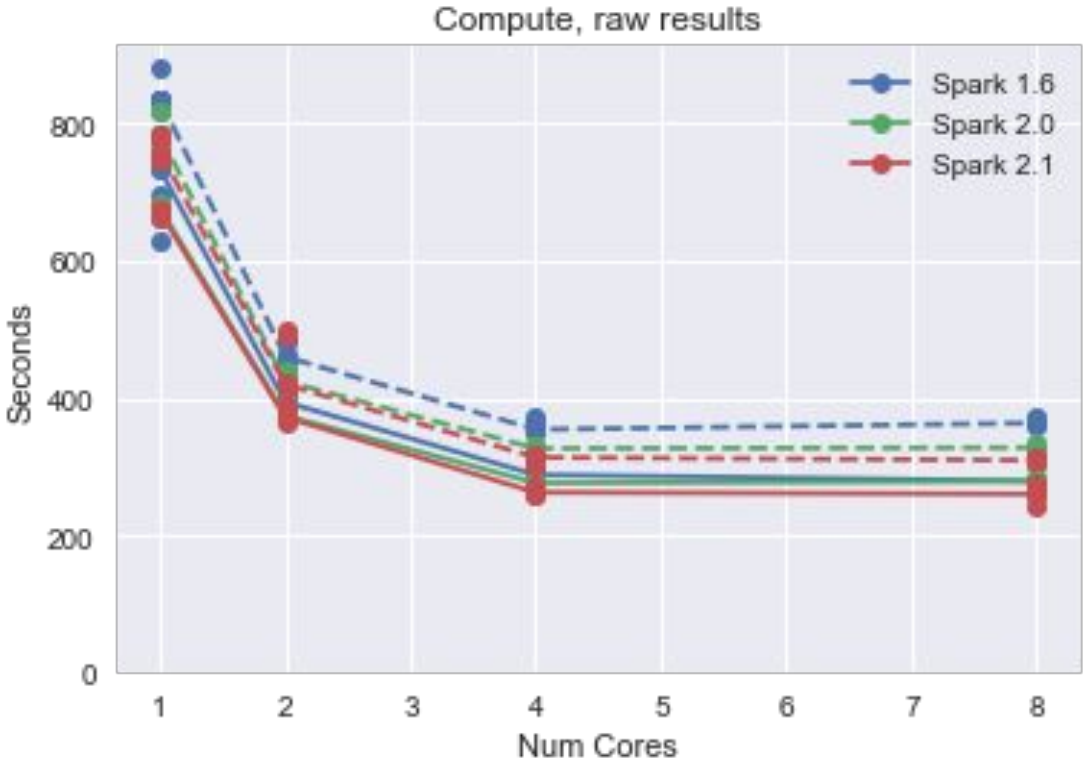
But we got terrible network bandwidth from Spark 2.0

Not necessarily intrinsic to Spark 2.0… could have been some detail of our setup

However Spark 2.1 worked fine, so we just decommissioned our Spark 2.0 setup

How do you know if you're getting your money's worth out of parallelization?

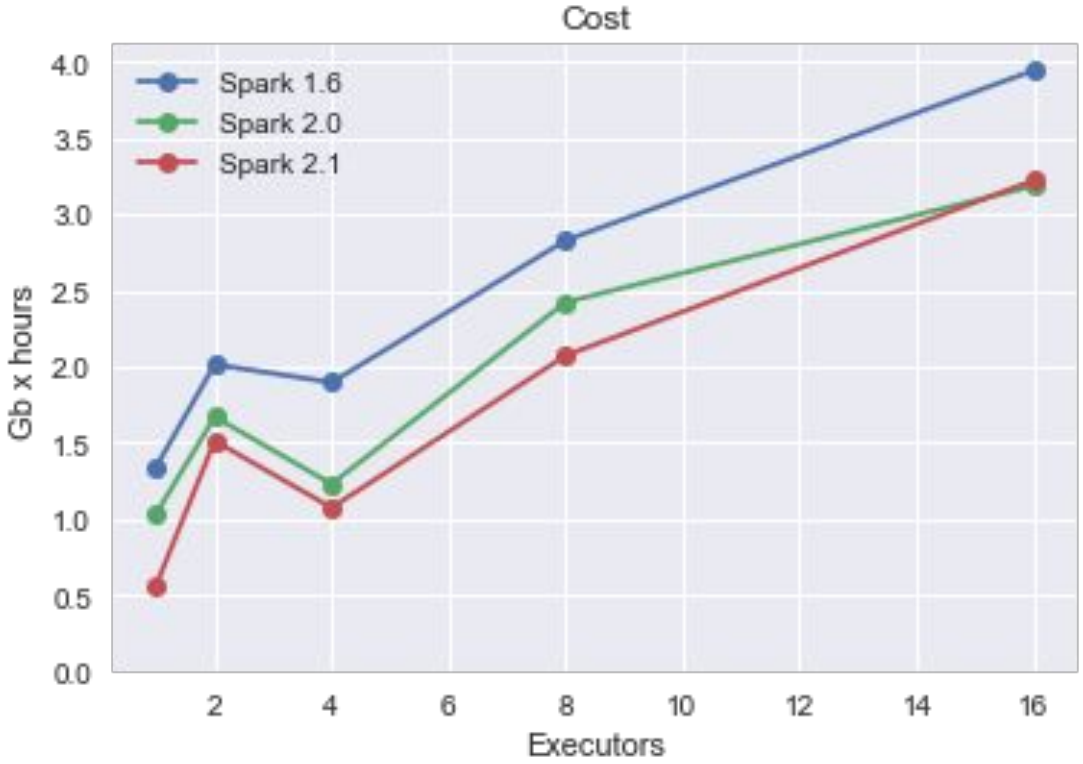
Run time vs. Number of Executors



Probably the first plot you draw…

but doesn't really tell you what you want to know

Overall Cost (in dollars if possible) vs. executors



In a perfect world (linear speed-ups) cost is independent of parallelism

In the real world costs generally rise with parallelism

Benefit: 1/walltime = answers per hour

1 hour vs. 2 hours:
Probably not a big deal

1 week vs. 2 weeks:
Probably is a big deal

1 minute vs 10 minutes is a huge deal:
Too easy to get distracted if your debug cycle is 10 minutes.

Once you are crisp on the costs and benefits, you will be in a position to say things like:

“If I double the amount of parallelism for this job, my AWS bill will rise by 30 pct and the job will run in 45 minutes instead of 60 minutes. Does that seem worth it to me?”

Recap

Focus on measuring performance of intrinsic machine characteristics like network bandwidth to characterize performance

Use carefully constructed data sets that change one and only one thing to do it

Be crisp on costs (dollars) and benefits (essentially debug cycles per hour) of parallelism to make informed choices about whether you want more or less of it.