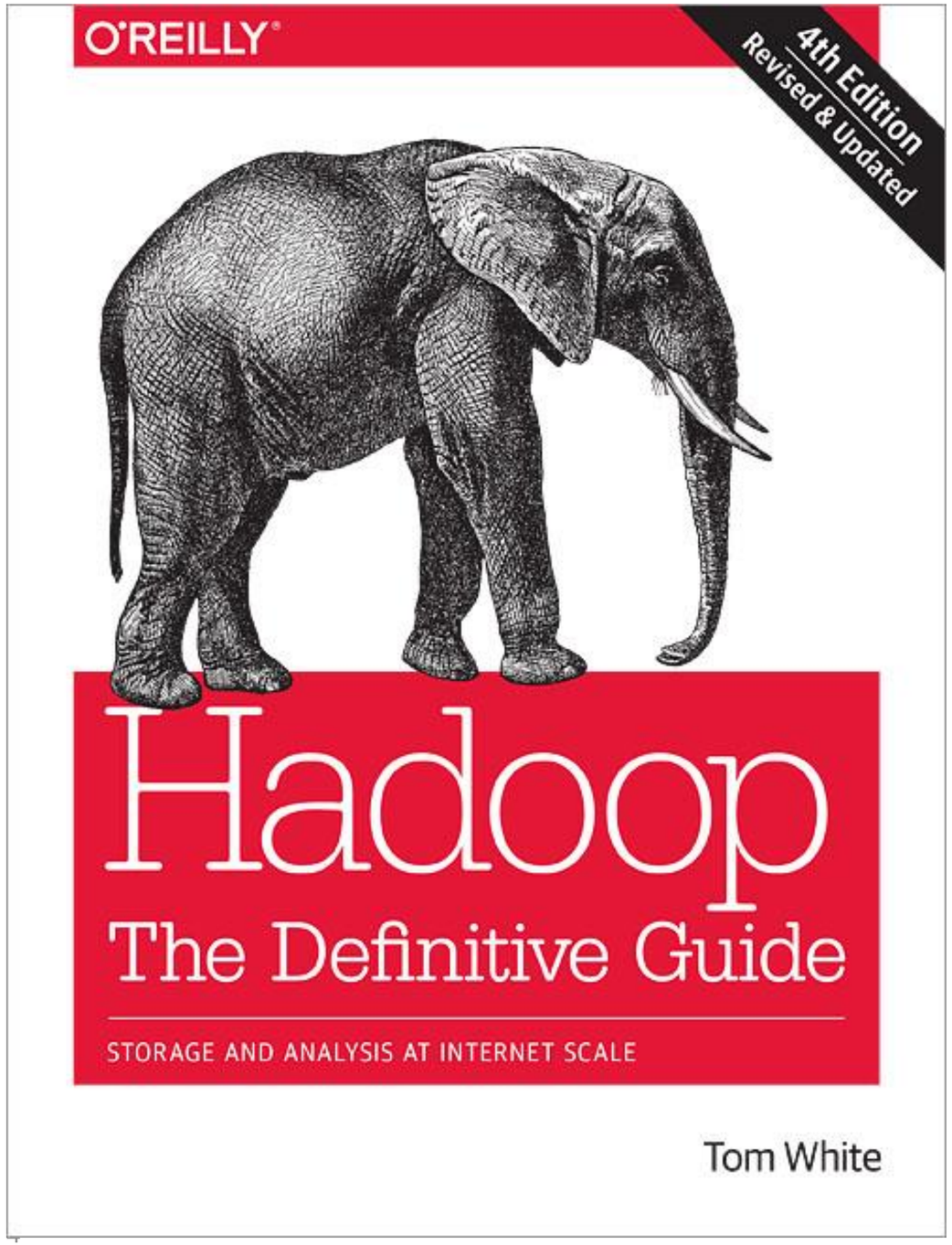
DLI Accelerated Data Science Teaching Kit

Lecture 10.2 - Why Hadoop?

How to Analyze Large Datasets?

- What software libraries to use?
- What programming languages to learn?
- Or more generally, what framework to use?



Lecture based on
Hadoop: The Definitive Guide

Book covers Hadoop, some Pig, some HBase, and other things.

http://shop.oreilly.com/product/0636920033448.do

Open-source software for reliable, scalable, distributed computing

Written in Java

Scale to thousands of machines

- Linear scalability (with good algorithm design):
  if you have 2 machines, your job runs twice as fast (ideally)

Uses simple programming model (MapReduce)

Fault tolerant (HDFS)

- Can recover from machine/disk failure
  (no need to restart computation)

http://hadoop.apache.org

Why Learn Hadoop?

- Fortune 500 companies use it
- Many research groups/projects use it
- Strong community support, and favored/backed my major companies, e.g., IBM, Google, Yahoo, eBay, Microsoft, etc.
- It's free, open-source
- Low cost to set up (works on commodity machines)
- An "essential skill", like SQL

http://strataconf.com/strata2012/public/schedule/detail/22497

Elephant in the Room

Hadoop created by Doug Cutting and Michael Cafarella while at Yahoo

Hadoop named after Doug's son's toy elephant

DLI Accelerated Data Science Teaching Kit

Thank You