Hadoop:
a brief history

Doug Cutting
Yahoo!

2002-2004: pre-history

- Nutch:
  - goal: web-scale, crawler-based search
  - open source, handful of part-time developers
  - distributed, by necessity
  - sort/merge based processing
  - demonstrated on 4 nodes
    - 100M web pages
    - operationally onerous
    - web scale still distant

2004-2006: gestation

- GFS & MapReduce papers published
  - directly address Nutch's scaling issues
- Added DFS & MapReduce impl to Nutch
  - two part-time developers, over two years
  - ported Nutch's crawler & indexer in 2 weeks
  - ran on 20 nodes at IA and UW
    - much easier to program & run
    - scaled to several 100M web pages
    - but still far from web-scale...

2006-2008: childhood

- Y! hired me & dedicated team, under e14
- Hadoop project split out of Nutch
- me: Apache/open-source liaison
- e14: engineers, clusters, users, etc.
- finally hit web-scale in early 2008!