Hadoop:
Industrial Strength Open Source for Data Intensive Supercomputing

Doug Cutting,Yahoo!
CIKM '08, Napa, CA, USA

Hadoop: What?

- A software framework for distributed computing
- For brute-force, ad-hoc computations
- Use thousands of computers in parallel
- Write simple programs that can
  - Reliably (more nodes means more failures)
  - Process data at rate it can be read off all spindles
  - Compute global properties (sort, count, etc.)

Hadoop: How?

- Inspired by two Google publications
  - GFS in 2003, MapReduce in 2004
- HDFS
  - Hadoop Distributed File System
  - Spreads data over all drives in a cluster
  - Single namespace for files
- MapReduce
  - Simple metaphor for parallel data processing
  - Supports global analysis (sorting, counting, etc.)

Open Source: What?

- It's all Free...
  - As in beer.
  - Beyond that, different models, licenses, etc.
    - e.g., read-only code dumps
    - Or, restrictions on use
- Apache
  - Community based
  - Collaboration by diverse parties
  - Meritocracy
  - Transparent process

Open Source: Why?

- For me
  - I want to write code...
    - That is maximally used
    - Does not become estranged
- For users
  - Free, as in beer
  - Transparent
    - Code and process
  - Mutable
    - With or without giving back

Open Source: Why?

- For Y! *(for whom I do not speak)*
  - Better code
    - Publication encourages quality
    - Sharing encourages generality
  - Happier Employees
    - Respect from wider peer pool
  - Lower costs
    - Shared with collaborators
    - QA, documentation, features, support, training, etc.

Open Source: How?

- Build it and they will come
  - Create something that's generally useful
  - Give it away
  - Treat others as welcome peers
    - Include in the process
      - No backroom planning
      - But not served as a customer
- Incrementally improve
  - Long-term planning hard w/o central control
- Have faith, not fear

Hadoop: Status

- Project nearly 3 years old (1k days)
  - ~4100 issues opened, ~3000 resolved
  - ~100 contributors, ~60,000 email messages
- Runs well on up to 4k nodes
- Terabyte sort world record, 209 seconds
- Used by:
  - Y!, Facebook, Amazon, AOL, IBM, Google, MS, ...
- Basis for:
  - HBase, Pig, Hive, Jaql, Mahout

Hadoop: Future

- Still lots of hard problems!
  - Security (authentication, authorization, etc.)
  - HDFS
    - High-Availability
    - Distributed Namespace
  - MapReduce
    - Better scheduling

Thanks!

http://hadoop.apache.org/