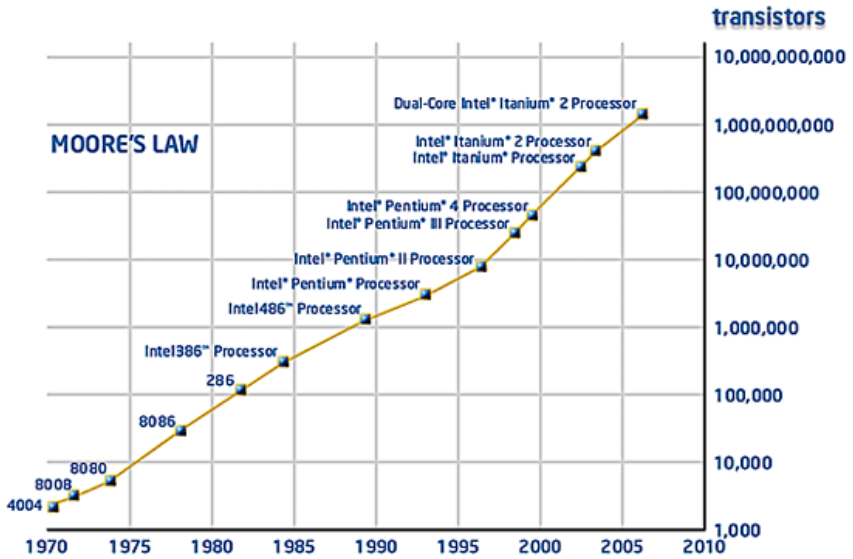
The State of the Apache Hadoop Ecosystem

Doug Cutting
Cloudera & Apache

Outline

- the ecosystem
  - why we need it
  - what it is
  - why its strong
  - how it can evolve
- highlights
  - current
  - next
- wrap up

Why are we here?



Hardware has improved

- exponentially for decades
- both storage and compute

We can now store and process **much** more!

- yet have been slow to leverage

Analyzing more data makes us smarter.

- Norvig's *Unreasonable Effectiveness of Data*

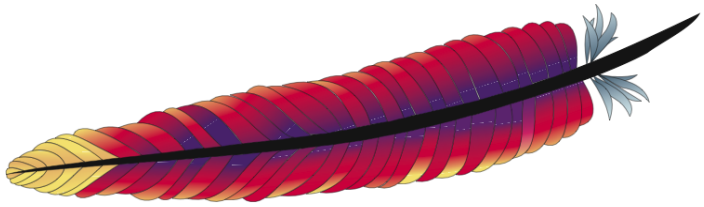
The Ecosystem is the System

- Hadoop has become the kernel
  - of the distributed operating system for Big Data
  - a de-facto industry standard

- No one uses the kernel alone

- A collection of projects at Apache

Strengths of Apache



Mandates diversity & transparency

- you control your fate

Insures against vendor lock-in

- can't buy the ASF

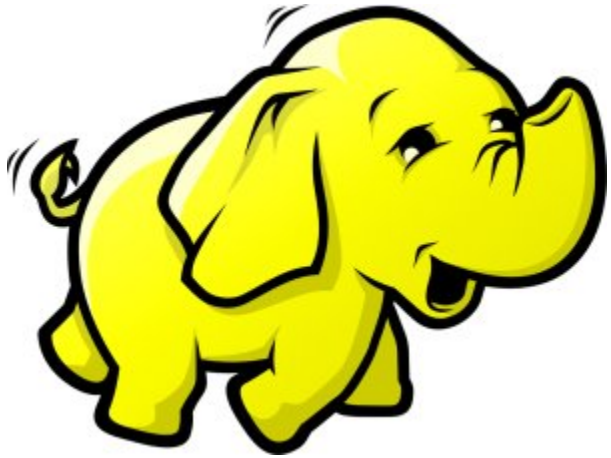
Allows competing projects

- survival of the fittest

Ecosystem as loose federation

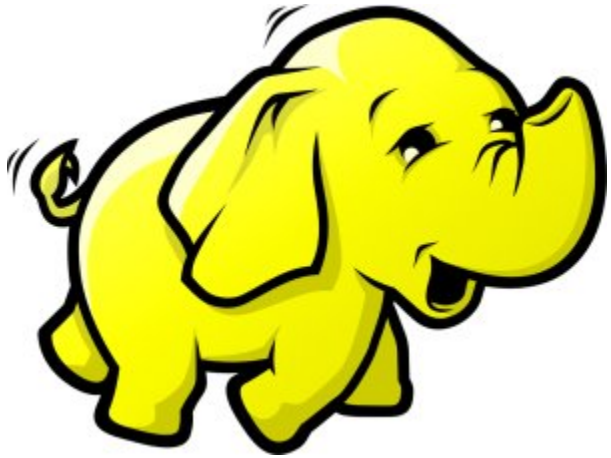
- lets platform evolve

What's new?



- Apache Hadoop 0.20.205
  - append
  - security
- CDH3
  - Mahout included
  - Avro support across components

What's next?



- Apache Hadoop 0.23
  - HDFS
    - performance
    - scalability (federation)
    - availability (HA)
  - MR2
- CDH4
  - includes Hadoop 0.23
  - BigTop-based
- S4, Giraph, Crunch, Blur, ...

Apache BigTop *(incubating)*

Ecosystem as a project

- integration tests
- compatible versions
- common packaging
- release is a set

Basis for CDH

- like Fedora is for RHEL

Community driven

Includes:

- Hadoop
- HBase
- Zookeeper
- Avro
- Hive
- Pig
- Oozie
- Flume
- Mahout
- ...

Join the community

Hadoop and Big Data are still young.

- Hardware trends will continue.

Hadoop started with just two developers.

- Now it has hundreds.
- You can be the next.
- What do you need?

Thanks!

Questions?