Yahoo! Experience with Hadoop

OSCON 2007
Eric Baldeschwieler
Why Yahoo! is investing in Hadoop

• We started with building better applications
  – Scale up web scale batch applications (search, ads, …)
  – Factor out common code from existing systems, so new applications will be easier to write
  – Manage the many clusters we have more easily

• The mission now includes research support
  – Build a huge data warehouse with many Yahoo! data sets
  – Couple it with a huge compute cluster and programming models to make using the data easy
  – Provide this as a service to our researchers
  – We are seeing great results!
    • Experiments can be run much more quickly in this environment
Yahoo! Inc.

Scaling Hadoop

Increasing size…

... while improving performance

• Hardware used in the benchmark
  – dual 2.8 GHz xeons with 4 SATA drives each
  – 10:1 network oversubscription (100mBit all to all)

Yahoo! Inc.
Yahoo! Inc.

Hadoop Clusters

- We have ~10,000 machines running Hadoop
- Our largest cluster is currently 1600 nodes
- Nearly 1 petabyte of user data (compressed, unreplicated)
- We run roughly 10,000 research jobs / week
Example: Web Crawl Problem Detection

• **The Problem**
  – Yahoo! crawls billions of pages per day, how do you detect when one site has a problem?

• **The Solution**
  – We load the crawl logs into Hadoop (via a map-reduce job)
  – We aggregate reports by site over time and flag sites where the crawl behavior has changed
  – This generates a report to customer service every day
  – They contact web masters and get sites fixed
Example: Web Survey

• The Problem
  – How do you know if new web technologies or products are gaining adoption on the web?
    • Is a micro-format being adopted by webmasters?
    • Which web2.0 site badges are being used?

• The Solution
  – We load our web crawl into Hadoop every month
  – We scan this for use of various technologies / products
  – Thus tracing the adoption of such technologies over time