abyte-Scale Processing: Big Ideas (courtesy of academia, Google, Yahoo, Facebook, etc.)

Commodity hardware

Share-nothing architecture

Move data to processors
Hadoop: HDFS and MapReduce

In source embodiment of those ideas)

**HDFS**

Distribute data blocks among servers

Application for reliability

**MapReduce**

- High-level programming abstraction
- Execution framework handles “details”
B rant Open Source Ecosystem
(right tool for the job)

ve

g

Base

keekeeper
Case Study 1: Yahoo

Hadoop runs production “Webmap”

Data Points

- Over 25,000 nodes running Hadoop
- Hundreds of thousands of jobs per day
- Typical HDFS cluster: 1,400 nodes, 2 PB capacity

Benchmarks (May, 2009)

- B sort: 1,460 nodes in 62 seconds
- B sort: 3,658 nodes in 16.25 hours
Use Study 2: Facebook

Data Points

100+ servers running Hadoop
- 1 PB under Hadoop/Hive management, +15 TB new data per day

What for?

- Data collection: server logs, web crawls, etc.
- Processing pipeline: ad optimization, summaries, etc.
- Ad hoc analyses
How much does one petabyte cost?

Considering hardware only…

Raw disks: ~$100k

2 TB SATA drive, ~$100

Lenovo X4540: ~$1m

2x6-cores, 32 GB RAM, 48 TB disk, ~$48k (21 units)
2 cores, 672 GB RAM

Hadoop cluster: ~$336k

Commodity server: 2x4-cores, 16 GB RAM, 12 TB disk, ~$4k (84 units)
2 cores, 1.3 TB RAM
How much does one petabyte cost?

Consider software now...

Parallel database vendor: ~$50m

Typical pricing: $50k per TB

Hadoop: $0

However, Cloudera would love to support you...