cloudera

First things first...

On behalf of the Hadoop community in general and of Cloudera in particular, I acknowledge that data parallelism is a technique with a long history in the operating systems and data management literature.

I am pretty sure that the guys from Google would admit this, too, but it is very hard to get them to say anything useful in public about the computer software that they write.

Now get over it.

Is Hadoop Relevant?

- Absolutely, among the big Web properties:
  - Facebook, Yahoo!, LinkedIn, all the rest
- Increasingly, among ordinary enterprises
  - Financial services, pharma and biotech, intelligence, energy, all the rest
- Terabyte scale storage and analytics using complex algorithms poorly suited to non-procedural languages and conventional query optimization and execution
- One size does not fit all!

The Character of Hadoopable Problems

- Large data volumes
  - Right now, two terabytes to start
  - No apparent upper bound
- Exhaustive analytics
  - Touch every single byte
- Often -- but not always! -- complex analytics
  - Don’t just touch it. Manhandle it.
- Often -- but not always! -- multi-stage analyses
  - Chain together a series of procedural transformations

- This is much more than “ETL on steroids”
  - Complements current columnar, relational, other tools

The Data Center of the Future

- A variety of storage and analysis tools:
  - The old-guard relational vendors will continue to make a lot of money solving critical business problems with their excellent products
  - A new generation of relational descendants -- columnar stores, for example -- will pick up a subset of the workloads and will deliver real value
  - Concepts that offend the elderly, like “no schema”, “no SQL” and “eventual consistency,” will gain a foothold for some workloads
    - Distributed hash tables. And give it up for Berkeley DB!
  - Hadoop is going to be huge.
- Did I mention that one size no longer fits all?

Thank you!