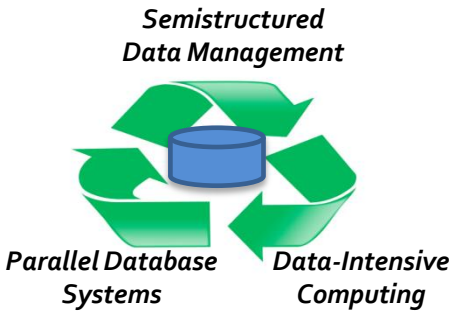
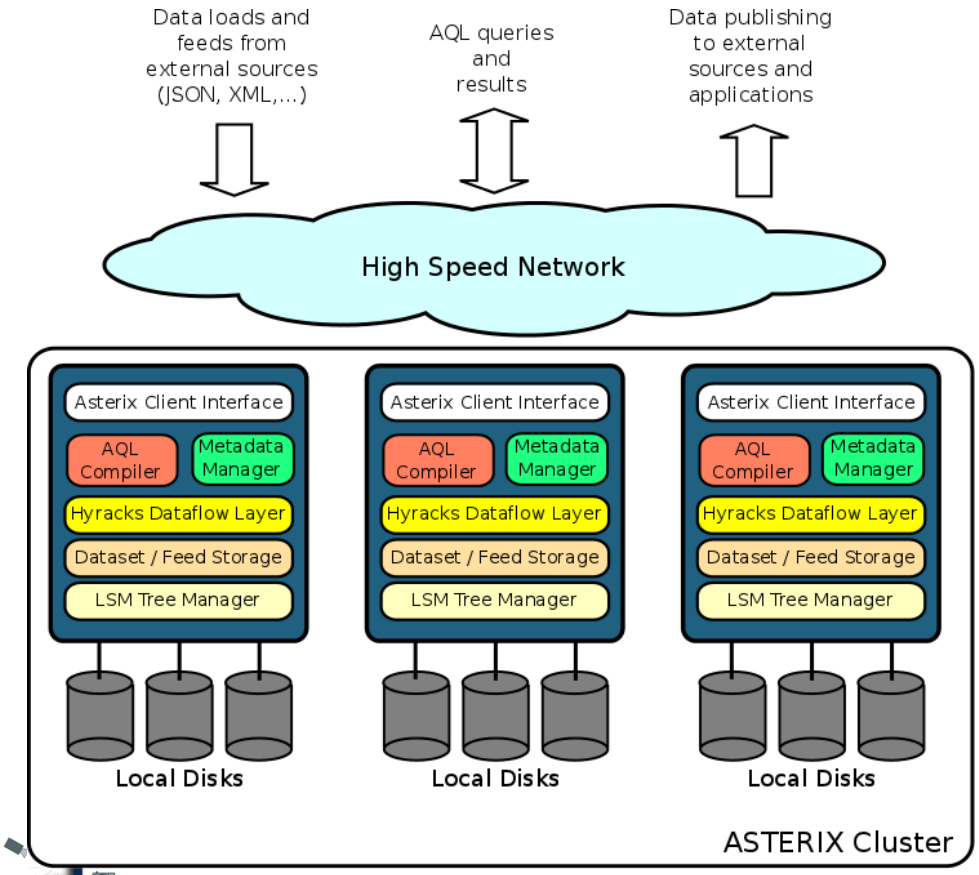
# Requirements for Meaningful Big Data Benchmarking

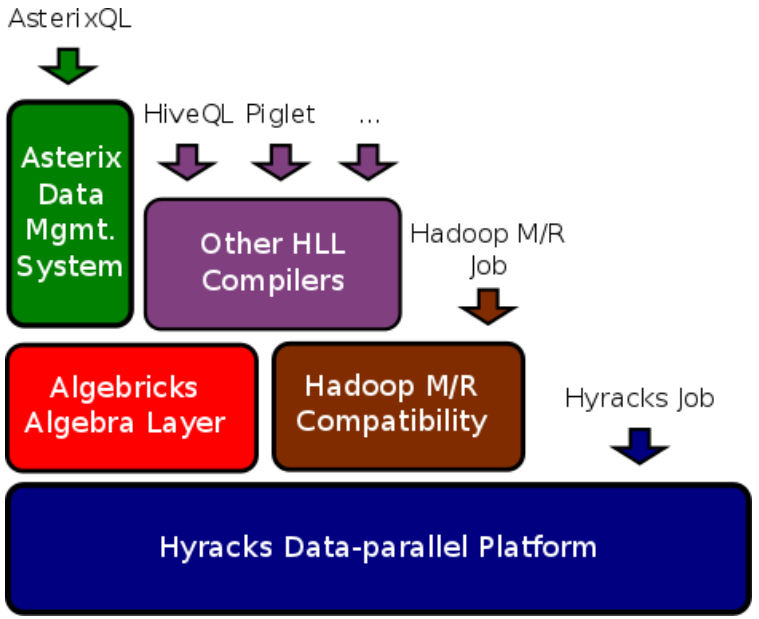
Michael J. Carey
*Information Systems Group*
*CS Department*
*UC Irvine*


http://isg.ics.uci.edu




# Context: ASTERIX @ UCI



# Pre-ASTERIX "Problem Survey"

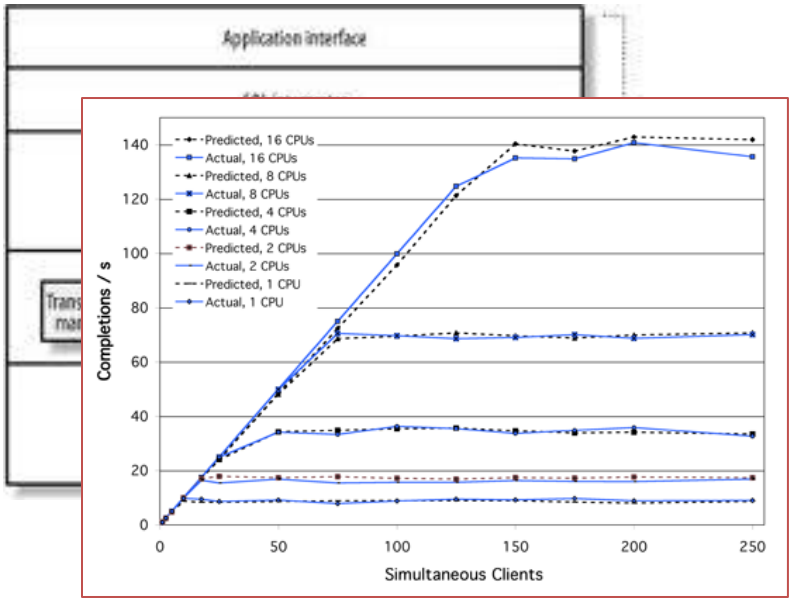
- We sought out Big Data problem input in 2009
  - Made pilgrimages to eBay, Facebook, Teradata, …
  - A common theme: Complex multiuser workloads!
  - (As opposed to single-problem/single-run studies)
- How do we benchmark meaningful scenarios?
  - Real clusters run a *mix* of jobs with different job sizes
  - Priority and responsiveness requirements vary as well
  - In 2009, eBay said nobody does this well (Teradata best, though)
  - From this we concluded:
    - We must have this problem on our research agenda!
    - We'll need to do benchmarks with such workloads…..

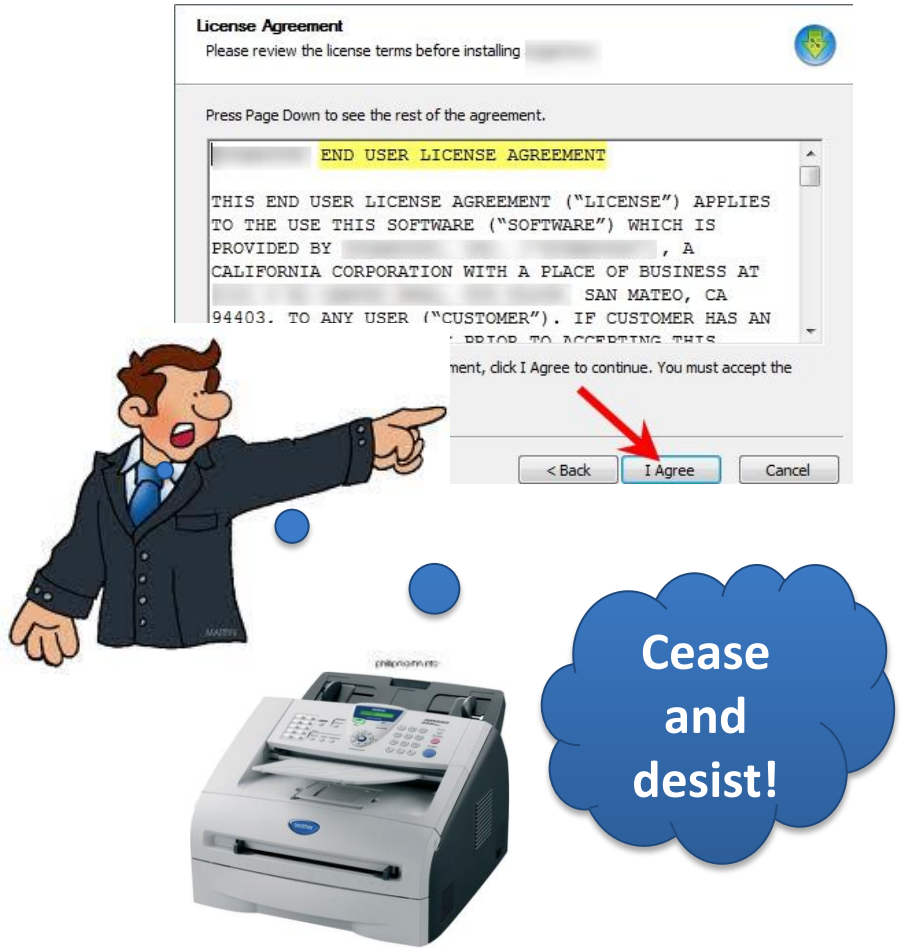
# Some Multiuser Considerations

- Workload generation
  - Multiple job classes (e.g., different job sizes)
  - Multiple job streams (bound statically to job classes)
  - *Ex:* Short update jobs mixed with long read jobs
  - Industry input on nature(s) of job mixes would be nice…!
- Performance metrics
  - Single user response time
  - Per-class multiuser response time
  - Per-class throughput (open vs. closed system)
  - Fairness (e.g., wait in proportion to request)
  - Careful consideration needed here!

# Will Open Source Be As Much Fun?

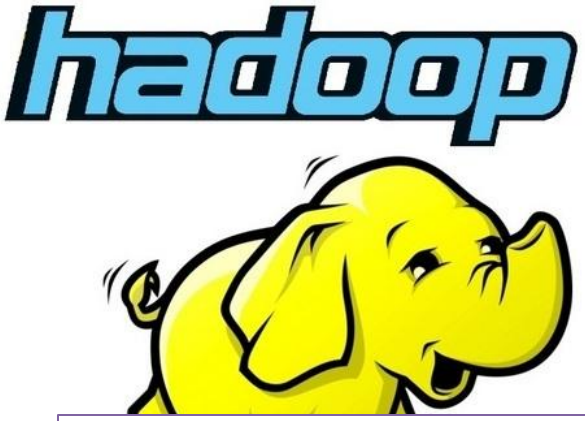
- Ah, the good old days:



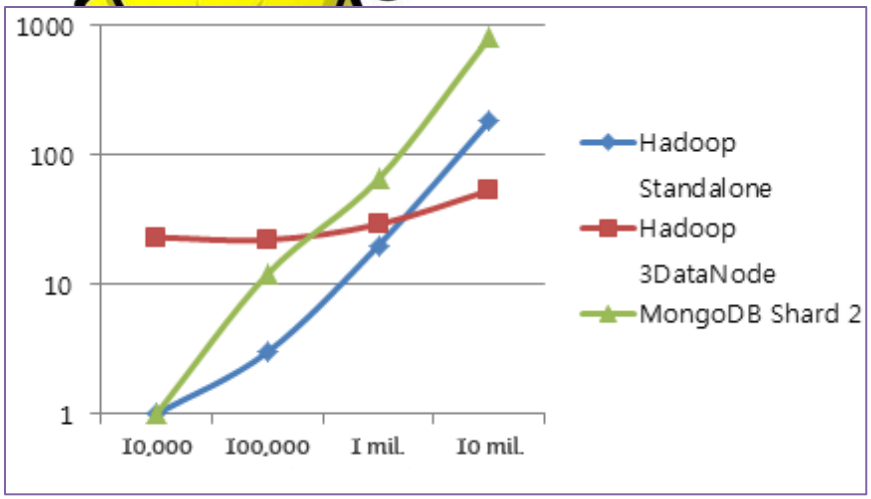



# Will Open Source Be As Much Fun?

- What now?








?