Chris Olston

- Projects:
  - Pig: data flow scripting for Hadoop
  - workflow managers on top of Pig/Hadoop

- Research questions:
  - testing/debugging, scheduling, multi-query optimization
Old Techniques, New Bundles

• Most “cloud” data management techniques are not new
• The hadoop community is building a DBMS (in layers)
• The main contribution lies in unbundling the techniques, motivated by scenarios that only want a subset ...

• Scenario 1 (web data discovery workspace):
  – ad-hoc mining of transient data sets → not worth it to load/partition
  – users are experienced programmers → would rather hand-optimize

• Scenario 2 (web processing workflows):
  – strict consistency is obviously overkill, but do need some consistency semantics (e.g. join link scores with click scores)
  – variable latency requirements: per-pathway (e.g. crawler-to-index) and per-datum (e.g. news pages) SLAs
  → manage consistency and view update timing in the app logic layer