Is MR a DBMS?

Michael Stonebraker

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

Outline

- What is MR good at
- What are DBMSs good at
- Coupling the 2 together
- Yabuts

MR is a Parallel ETL tool

- Good at what ETL is good at
  - Transforming data
  - Data assembly
  - Low touch data

DBMSs are DBMSs

- Good at query, update
- High touch data

DBMSs do not try to do ETL

- No good at it
- They are downstream
- With good interfaces

MR should not try to do DBMS

- No good at it (X 50 slower)
- In spite of Google guys doing a DBMS benchmark in CACM Jan '08
  - Huge (not very productive) head fake
- Instead couple to a DBMS (downstream)

Real Answer

- Good interface between MR and DBMS
- E.g. Vertica. Asterdata, Greenplum, HadoopDB
  - Each system does what it is good at

Yabut

- MR has higher scalability
  - Nobody is currently asking for DBMS scalability about about 100
  - If they do, DBMSs will scale
  - If you are a factor of 50 slower, then you need 50X the nodes

Yabut

- MR provides intraquery recovery; DBMSs only do interquery recovery
- Nobody is asking DBMS for this feature; easy to provide if they do (make nodes in the query plan restartable)
- If you are a factor of 50 slower, you are 50X more likely to crash

Yabut

- Hadoop is open source
- So is Infobright, MySQL, Ingres, Postgres, SciDB, commercial H-store, ….

Yabut

- Hadoop is easier to use and set-up
- So are most other ETL tools
- It is (so far) difficult for DBMSs to work well in the high end corner cases without knobs. Challenge is to put them in only when needed

Yabut

- Hadoop allows semi-structured data
- So do most other ETL tools

Summary

- MR is an ETL tool
- Couple to a DBMS for DBMS stuff
- Lots of examples; more coming
- One system does not do everything