# MapReduce, Pig, and RDBMS: Friends or Foes?

YongChul Kwon

CSE444 – Winter 2011

The title is adapted from the article
"MapReduce and Parallel DBMSs: Friends or Foes?" by Stonebraker e.a.

# MapReduce: A major step backwards

- Seminal debate in Jan 2008
  - http://databasecolumn.vertica.com/database-innovation/mapreduce-a-major-step-backwards/
- Five points
  - MapReduce is a step backwards in database access
  - MapReduce is a poor implementation
  - MapReduce is not novel
  - MapReduce is missing features
  - MapReduce is incompatible with the DBMS tools

# MapReduce is A step backwards in database access

- No schema or schema free
- Separation of the schema from the application is good
- High-level access languages are good

# MapReduce is A poor implementation

- No index. Only offers brute force access.
- Poor handling of skew
- Shuffle phase incurs a huge random access on disks

# MapReduce is Not novel

- User-defined functions have been around in database for decades
- Many of the parallel distributed processing techniques have been extensively researched in database literature

# MapReduce is Missing features

- Bulk loader
- Indexing
- Updates
- Transactions
- Integrity constraints
- Referential integrity
- Views

# MapReduce is Incompatible with the DBMS tools

- Report writers
- Business intelligence tools
- Data mining tools
- Replication tools
- Database design tools

# Questions

- Do you agree or disagree?
- How systems like Pig address the criticism?
- Can you find features and techniques from database in Pig? What are they?

# Follow-ups

- A Comparison of Approaches to Large-Scale Data Analysis
  - SIGMOD 2009
- MapReduce and parallel DBMSs: Friends or Foes?
  - Communications of the ACM, Jan 2010
- MapReduce: a flexible data processing tool
  - Communications of the ACM, Jan 2010