SS-DB: A Standard Science DBMS Benchmark

Michael Stonebraker (et. al.)
What is SS-DB?

- Abstraction of LSST (astronomy) workload
  - Captures essence of several other fields as well (earth science, oceanography)....
What is SS-DB?

- End-to-end problem (as opposed to Sloan Sky survey and Jim Gray 20 questions)
  - Ingest of raw data
  - Cooking
  - Queries on cooked and raw data
SS-DB Data

• Raw imagery
  - 2D arrays of telescope readings
  - Considerable skew

• “Cooked” into observations
  - Image intensity algorithm
  - Spatial data

• Further cooked into “trajectories”
  - Similarity query
  - Constrained by maximum distance
SS-DB Benchmark

- 1 Tbyte of raw data (regular)
- Scales up and down (small, large)
- Can be made easier or harder by parameter settings (sizes of geometry)
Benchmark

• Load some data

• Cook what you loaded

• Run a collection of queries
  - To the raw data (regrid, recook, aggregate)
  - To the cooked data (mostly spatial queries)
Implemented on SciDB and MySQL

- 1 Tbyte – 10 nodes
- SciDB much faster (≈X 15)
  - Better parallelism (overlapping chunks)
  - Column storage of array values
  - Better compression
  - Better spatial data support
  - Better load speed
Details

• Data generator
• Cooking algorithm
• Queries
• Helper functions
• Detailed MySQL and SciDB numbers
• Will be available on SciDB website by November 1st
Next Steps

• Try it on HPC!
• Try it on your favorite DBMS
• Goal: Improve DBMS usability and performance on this style of application