A Science Benchmark

by

Michael Stonebraker
Sci-DB Benchmark

- Capture essence of science data processing
  - Raw data
  - Cooking
  - Querying
- Modelled on LSST data
- Make SciDB look good
  - And RDBMS look bad
- Can’t start with Skyserver queries
LSST Data
Raw Data

- Array of pixels
- Of a portion of the sky
- Time series of such arrays
Benchmark Model

- Big co-ordinate space
- Each image is positioned in the space
- Images are uniform in time (just to make life easier)
- Hot spots of interest
  - 80% of the images in a small area
Observations

- Images are “cooked” into observations
  - Basically feature extraction
  - Observations have a maximum size
  - Record a bunch of data (center, intensity, size, etc.)
- Effectively this is a spatial data base of polygons
Trajectories

- Same observation at different points in time
- Observations have a maximum velocity
- Basically a data base of polygons moving in time
Benchmark

◆ Group images into a collection of “units”
  ◆ “Batch” processing
◆ Cook each unit
  ◆ This is the load time
◆ At the completion of all of the units
  ◆ Run a collection of parameterized queries
  ◆ Fifteen times
  ◆ This is query time
Example Queries

- Recook portions of the imagery
  - With a different algorithm
- Regrid a portion of the raw data
- Find observations that intersect a “spatial slab”
- Find trajectories that intersect a “time cube”
- Total of 9 queries
Benchmark

- Small (100G), medium (1T) and big (10T) data
- Easy, medium and hard settings
  - How big the slabs are, maximum size of an observation, etc.
SciDB whomps “Sharded” MySQL

*Around a factor of 100*
Reasons

- Native array storage (not tiles simulated in blobs)
- Cell values stored “by column”
  - Read only those cells you need
  - Row store reads them all
Reasons

- Overlapping “chunk” storage with parallel execution
  - Too hard to code for shards
- Other parallel operations
  - Some things too hard to code on shards
- Compression built-in, not added on
  - MySQL is slower on the benchmark if compression turned on
Status

- Paper submitted to ICDE ’11
- Ask me for a copy
Next Steps

- How to get science buy-in for this (or another) benchmark?
- How to leverage it to get visibility?