SciDB A DBMS for Analytic Applications

by

Michael Stonebraker



Outline

- Context
- Application areas
- Why RDBMS Doesn't Work
- Our partnership
- **◆Status and future**



Context – One Size Does Not Fit All

- Vertica column store for warehouses
 - ♦50X the elephants
- VoltDB main memory, single threaded for (nearly partitionable) OLTP
 - ♦30-40X the elephants
- ◆At least one more vertical market template



Serious Analysis Applications

- ◆Science users (Astronomy, Earth Science,...)
- Web log analysis
- Medical imaging
- Drug Discovery
- Spooks



Three Lighthouse Customers

- ◆e-Bay
- **◆LSST**
- Russian astronomy project

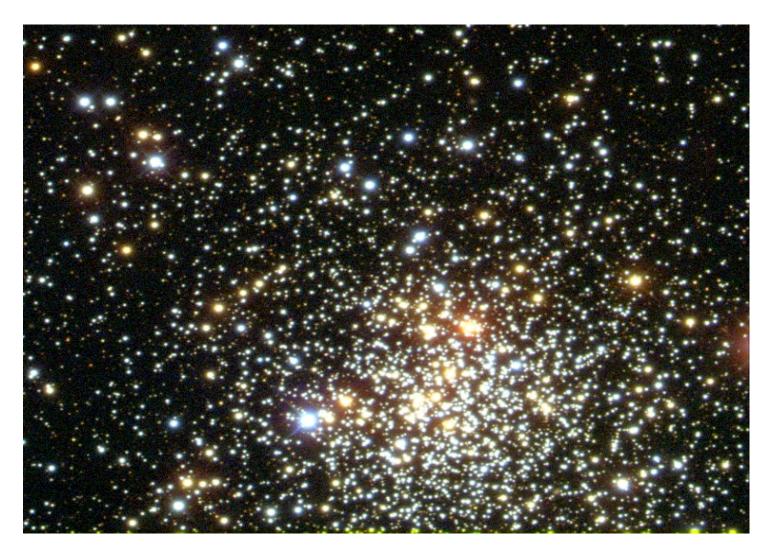


e-Bay Application

- Web log data (petabytes)
 - Sessionization
 - Clustering (in N-space)
 - Mining
 - Predictive analysis



LSST Data





LSST Application

- Raw telescope imagery (big arrays)
- "Cooked" into features (geographic data)
 - Data clustering algorithm
- Grouped together into observations of the same feature
 - Similarity metrics



LSST Queries

- Recook portions of the imagery
 - With different algorithm
- Trajectory queries
- Nearest neighbor queries



Why SciDB?

- RDBMS has the wrong data model
 - Arrays not tables
 - ◆Data clustering is natural in N dimensional space!!
 - ◆Tables impossibly slow at simulating arrays (x 100)



Why SciDB?

- RDBMS has wrong operations
 - ◆Regrid, cluster, not join (can't even wrap your mind around data clustering)
 - Parallel, user-defined functions a requirement
 - My-new-clustering-technique



Why SciDB?

- RDBMS has features missing
 - Named versions
 - Recluster just the tech stocks using my fancy algorithm
 - Provenance
 - What clustering technique was used?
 - Uncertainty
 - What is the error in my clustering?



Net Result

- ◆Roll-your-own on the bare metal
- Or put up with a horrible kludge on RDBMS
 - With mountains of app logic
 - And copying the world to app space



Design Team

- ◆Mike Stonebraker
- ◆Stan Zdonik
- Dave Maier
- ◆Sam Madden
- Magda Balazinska
- Dave DeWitt



Building SciDB

- e-Bay (partial FTE)
- LSST (2 FTE working on project)
- Persistence Software (committed 3 FTE)
- M.I.T (1 postdoc)
- University of Moscow (3 FTE)
- Paul Brown (world's best programmer)
- Washington (1 grad student)
- Brown (1 grad student)



What is SciDB?

- Nested array data model
- With bells and whistles
 - Non-uniform dimensions
 - Boundaries and holes
- Science specific operations (e.g. regrid)
- Array UDFs



What is SciDB?

- No overwrite
 - ◆Time is another array dimension.
 - New values written here
- Partitioning across nodes
 - Sharding in multiple (one or more dimensions)
 - ♦With overlap!!!!



Current Status

- "Sharded" multisite array system
- Which does a collection of interesting LSST queries
- Missing bunches of stuff (optimizer, parser, bulk loader, ...)

I.e. PoC demoware



- Storage manager
 - Big blocks (chunks)
 - Refining the node partitioning
 - Vertica-style compression
 - Replication by multiple copies with different partitioning
 - Each attribute stored in a separate physical array



- Fixed stride
 - Easy to index
 - But blocks may be highly variable in size
- Or variable stride
 - ◆Need an R-tree
 - But packing can be more uniform



- Optimizer design
 - Many "blocking" nodes in the plan (e.g. regrid)
 - Opportunity to repartition arrays
 - What cost function?
 - Deal with replication
 - ◆Not 2 phase



- Array UDFs
 - ◆No cell level UDFs
 - Experience of Postgres



- Uncertainty
 - Additional cell attribute
 - Uniform distribution
 - ◆Fast
 - Or something more complex
 - Can be arbitrarily slow



- Provenance
 - Want to trace backward from "bad" values to identify "patient zero"
 - Want to trace forward from patient zero
 - To fix all propagated errors
 - ◆Implementation?
 - ◆Trio a non-starter



Development Plan

- Complete system at end of Q1/10
 - But no transactions or recovery
 - Kludgy messaging system
- Currently being horribly under-managed
 - ◆No formal QA
 - ◆No formal doc



SciDB will be 100X RDBMS

- Optimized parallel UDFs
- Redistribution with overlap
- Multi-dimensional storage (not a column store; not a row store)
- Correct operations



What Else Have We Done?

- Science benchmark
 - Nearly done
- A bunch of use cases
 - See our web site (scidb.org)



What We Need

- Money
 - **◆NSF** does not like the word "infrastructure"
 - **♦VCs** worry about the size of the market

