



20 Spatial Queries for an Astronomer's Bench (mark)

María Nieto-Santisteban¹

Tobias Scholl²

Alexander Szalay¹

Alfons Kemper²

1. *The Johns Hopkins University,*
2. *Technische Universität München*

Motivation

- Most astronomical data is or must be online
- Data & catalogs are growing
- No single best data management solution
 - Different projects approach the data access challenge in different ways
 - Different users and science problems have different needs
- **A well defined “test” suite to provide objective comparison would be useful**

Benchmarking

Key criteria for a domain-specific benchmark

- **Relevant:** It must measure the peak performance and price/performance of systems when performing typical operations within that problem domain
- **Portable:** It should be easy to implement the benchmark on many different systems and architectures
- **Scaleable:** The benchmark should apply to small and large computer systems. It should be possible to scale the benchmark up to larger systems, and to parallel computer systems as computer performance and architecture evolve
- **Simple:** The benchmark must be understandable, otherwise it will lack credibility

“The Benchmark Handbook: For Database and Transaction Processing Systems,” Jim Gray

- **Repeatable**

Data Corpus (+ outputs)

Table	nCol	rSize(B)	nRow	tSize(GB)	Notes
<i>ROSAT</i>					
DS1	30	193	125K	0.02	RASS
<i>FIRST</i>					
DS2	14	93	811K	0.07	
FP1	5	32	52	~ 0	Footprint
<i>SDSS DR6</i>					
DS3	63	252	377M	89	PhotoTag
DS4	63	252	116M	27	I < 20.0
FP2	5	32	43	~ 0	Footprint
<i>2MASS</i>					
DS5	63	482	471M	211	PSC
...					
* raw data size without indexes					

The “20” Spatial Queries

- Single/Multi Catalog & Regions
 - Cone search: Find objects within a circle
 - Find objects within a circle satisfying a high multi-dimensional constraint
 - Find the closest neighbor
 - Find objects within a region
 - Find objects in/outside masked regions
 - Find objects near the edges of a region
 - Compute the area of a region
 - Find surveys covering a given region
 - Find the intersection between several surveys
 - Count objects from a list of regions

The “20” Spatial Queries

- Find these 1k - 100k objects in these catalogs
- For all catalogs, extract a random sample of existing objects within a given region
- Cross-match 2 catalogs within a given region
- Cross-match n catalogs, $n > 2$, within a given region
- Find objects which are in A, B, and C but not in D
- Given a sparse grid, find the closest grid point for all objects in the catalog
- Find multiple detections of the same object with given magnitudes variations
- Find all quasars within a region and compute their distance to surroundings galaxies
- more . . . **open to discussion** ...

Query Specification

- Inspired by the TPC Benchmark tm H spec.

Queries are defined using:

- ✓ **The business question**, using natural language sets the context and functionality, and specifies the inputs and outputs parameters
- ✓ **Substitution parameters**
 - Random input parameters
 - Random output columns
- ✓ **Validation**
 - Provide a few queries with known results
- ✗ **Functional query**, defines the business question using SQL

An Example: Cone Search Query

■ Business Question:

- The query describes sky position and an angular distance (both in degrees), defining a cone on the sky. The response returns a list of astronomical sources from the SDSSDR6 catalog, ds4 in the corpus, whose positions lie within the cone. The response shall include: **objid, Ra, Dec, d[deg], modelMag_u|r|i|z|g**

■ Substitution Parameters:

- Runs for 1000 random points
- Runs for Radius = 4", 16", 30", 1', 4', 30', 1 deg, 4 deg
- Returns 10, 20, 30 additional randomly selected columns

An Example: Cone Search Query

- Validation:
 - RA = 195.0, DEC = 2.5, Radius = 1'

objid	ra	dec	d[deg]	mMag_i	mMag_g	mMag_r	mMag_u	mMag_z
587726032791994481	194.999	2.5131	0.0132	18.6113	19.9490	18.9565	22.3014	18.4255
587726032791994487	195.006	2.4900	0.0116	20.1843	20.2789	20.1739	20.9002	20.2222
... 24 rows ..								
587726032791995406	194.991	2.5086	0.0121	21.0228	22.5842	22.1353	22.4252	21.4586

The Metrics

- Quantitative (min, max, avg, geometric mean, stdev)
 - Elapsed time [s], CPU [s], I/O
 - Throughput: Data [Mbs], Work [Jobs/s]
 - Total space: Data + Index [GB]
 - Number of concurrent users
- Qualitative (capabilities)
 - Query flexibility
 - Single catalog
 - Multi-catalog
 - Inter-catalog
 - Data flow: upload and download, *MB, GB?*
 - Distributed access to data and services
 - Workflow definition
 - Data, results & code sharing
 - Job management

Final Remarks

- Benchmarking is difficult and controversial
- Different hardware, different software, changing environments, ...
- Scalability is not guaranteed



Copyright © 1997 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited