

JDEM DB Development Status

Igor Mandrichenko, Vladimir

Podstavkov

6/30/2010

Where we are

- Work with Steve and Eric to implement the DB for slitless spectroscopy and other programs
- Trying JDEM use cases
 - Galaxy position reconstruction
 - Catalog cross-matching
- Comparing GreenPlum to plain Postgres

2 database configurations

- Jdemdb1: Postgres running on RAID-0 array of 8 disks, 7.2 TB in total
 - Was 8-segment GreenPlum, 2 segments per RAID-0 2TB logical disk
- Jdemdb2: 8-segment GreenPlum instance, each runs on its own physical disk of ~1TB

Plan: generation

- Generate 500 million galaxies as a “base catalog”. 500 million number comes from disk space limitation
 - Compare data ingestion performance of GreenPlum and Postgres (we pretty much know the result, GreenPlum is *much* faster due to the ability to ingest in parallel)

Use case 1: reconstruction

- Generate 1 day worth of “detections”
 - 20million (5 million galaxies by 4 roll angles) from 1 x 31 degree area of the base catalog and reconstruct galaxies coordinates
 - Goal is to reconstruct all in less than 24 hours
- Compare GreenPlum and Postgres performance and scaling
- Compare performance with and without blobs (move detection data into separate table with 1-to-1 relationship with the coordinate table)
- Compare 1 by 1 reconstruction algorithm with bulk reconstruction (say 500 detections at once)

Use Case 2: Catalog Crossmatching

- Take 1.0 by 0.5 degree area (1 exposure) of the base catalog and match all galaxies there back to the catalog, measure time.
 - Goal is < 350 seconds – to sustain 246 exposures/day rate
 - Compare Postgres and GreenPlum performance, scaling

FITS-Database tools

- sql2fits – dump data from arbitrary query into a fits file – done
- fits2sql – store data from a fits file into one of more database tables. More difficult, we are working on it
- Both tools will recognize and handle pgSphere data types
 - spoint – array[2]
 - scircle – array[3]
 - sbox – array[4]

Further Plans, Ideas

- Explore GreenPlum secondary segments
 - Internal data replication, redundancy
- Work with Steve and Eric to implement the DB for slitless spectroscopy and other programs
- Load SDSS data set
 - And then.. ?
- Load DES data archive
 - And then.. ?