STARS 2.6 in STN6:
Database Efficiencies for Hyper Suprime-Cam Multiple-Frame Exposures

Thomas Winegar
Subaru Telescope
STARS Administrator
winegar@naoj.org

Abstract:

In 2023, Subaru Telescope STARS 2.6 installed a modified STARS database adjusting Hyper Suprime-Cam (HSC) multiple-frame exposures, to reduce database size and for more efficient queries.

The modification involved deletion of 89% of HSC keyword=values, which were de-duplicated within each HSC exposure of 112 frames. In addition, an HSC Exposures table compiled details for each exposure, including starting & ending frames and total frame counts.

During registration of the first frame in each exposure, all keyword=values are stored. For subsequent frames in each exposure, only unique keyword=values are stored.

Now, STARS database collects two sources of keyword=values for each frame: all keyword=values from the 1st registered frame in each exposure and only unique keyword=values in each subsequent frame. This de-duplication resulted in a 40% reduction in database size.

Report:

Beginning in 2014, Hyper-Suprime Cam tested the limits of FITS registration into the STARS database and archive. The raw size and number of HSC FITS files collected in a single night, 15,000 frames of 18 MB each, comprised a 100x increase in a typical nights data acquisition (3 GB/nights to 300 GB/nights). However, focusing on the large data size was deceptive, as we successfully estimated and adjusted the archive expansion rate (~40 TB/year for HSC) without planning for increased database size.

By 2020, it was apparent that database was also increasing significantly, averaging more than ~100 GB/year. The 2013 database was 128 GB, the 2017 database was 370 GB. The 2020 database was 800 GB and on track for 1 TB within two years.

Investigation of the FITS keyword=values stored in the database revealed a highly duplicated set of records, with 89% of the HSC keyword=values duplicated within a single 112 frame exposure. The database was growing with duplicated values. HSC had copied a similar FITS Header into each of the 112 FITS Headers. For example, the same DATE-OBS value was stored for each of 112 frames. Of 195 total keywords in each HSC FITS Header, 174 keyword=values were identical for all frames in the same exposure. There was only 21 unique keyword=values in each of the 112 frames. For HSC, 89% of the database records were copied values.

The huge number of duplicated records in the database had the dual effect of slowing down display results by returning too many matches and drowning the user with unchanging results. The database was slow and the user could not distinguish variations in the data without excessive paging. Often the user did not attempt to page the results because the variations were not noticeable.

Beginning in 2020, STARS staff started planning for de-duplication of the HSC records to regain efficiency. By 2021, three changes were defined.

1. Only unique keyword=values were retained inside of one HSC exposure.

2. An exposures table was compiled to manage a single HSC exposure composed of 112 HSC frames.

3. An exposure relation was added to allow the option for the user to switch between grouping by frame or exposure.

Beginning in 2022, a de-duplication effort began organizing 2 sets of records for each frame. For the first frame in one exposure, all 195 keyword/values were stored. For subsequent frames #2 - #112, only 21 unique keyword=values were stored.

The de-duplication process resulted in a decrease from 21,840 keyword=values per exposure to 2,526 keyword=values per exposure. The Exposures table was used to confirm each keyword=value was unique and facilitated the deletion of duplicated values without any changes to keyword=values. The database records were organized and many were deleted without any loss of scientific information.

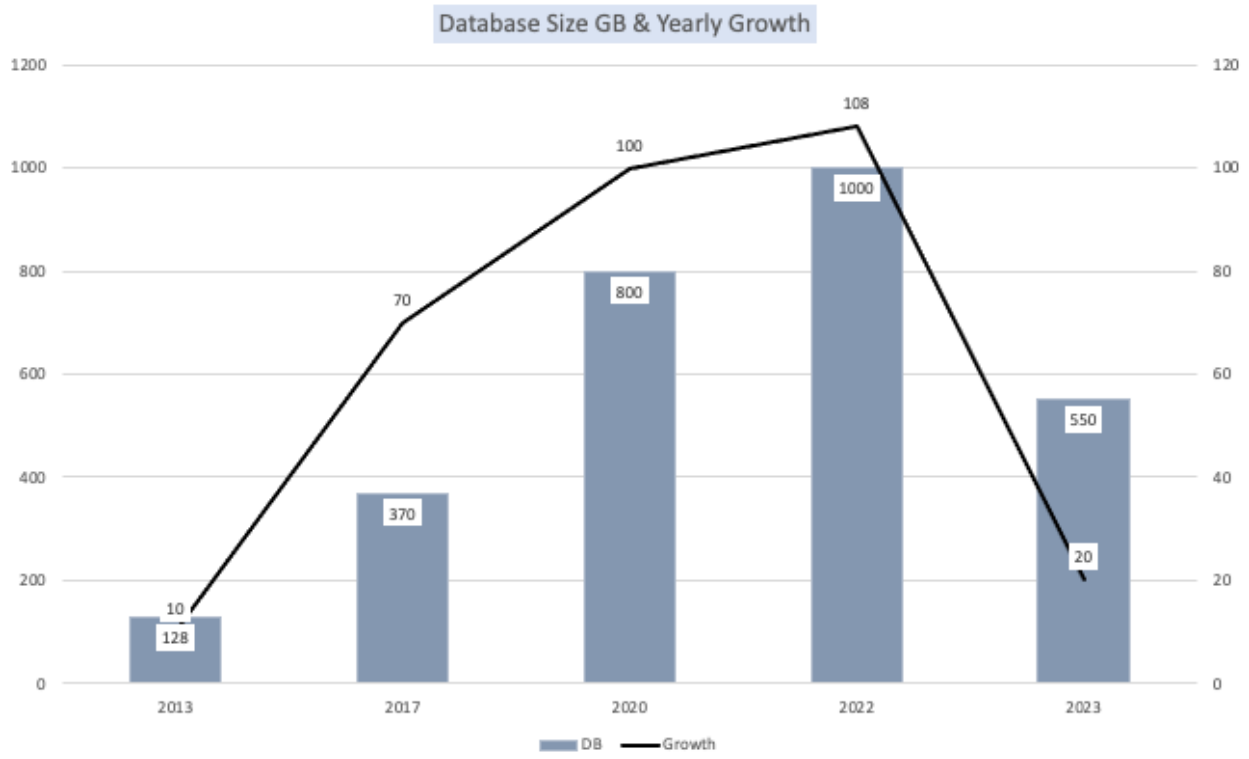
The final efficiency added to the database was an additional relation to allow organizing results by frame or exposure. Typically, the user maintains grouping by exposure as default, as this reduces a single night's frame count from ~14,000 to ~124 (for example, 2023-09-07) and achieves a more human-scaled result. The frame vs. exposure selection offers compiled or detailed results.

Conclusion:

By 2023, the 3 database efficiencies were completed and applied to primary and replica database servers. The de-duplication process deleted 140 million keyword=values and kept 110 million unique keyword=values. This deletion resulted in a 40% reduction in database size, from 900 GB to 500 GB. The disproportionately smaller reduction is because we did not delete any keyword=values for other instruments.

By modifying registration for HSC, the yearly growth rate of the database was reduced 80%, from 100 GB/year to 20 GB/year. For display of query-results, users experienced a ~3x reduction in elapsed time, from 20-60 seconds to 6-20 seconds. Finally, elapsed time for database backup administrative functions is cut in half.

At the time of this report in 2024, we have a smaller and faster database focussed on unique keyword=values and customized query-results. The database is growing at a slower rate and returning results quicker. The database efficiencies have more than repaid our efforts to our users and on disk.


Database Size GB & Yearly Growth
1200
1000
800
600
400
200
0
120
100
80
60
40
20
0
10
70
100
108
20
128
370
800
1000
550
2013
2017
2020
2022
2023
DB
Growth