S20 HSC PDR2 Reprocessing

The main objective here is to have the needed data for the Rubin Observatory Algorithms Workshop.

The catch-all ticket is

⚠ DM-23243 - Jira project doesn't exist or you don't have permission to view it.

. Output repos will be inside /datasets/hsc/repo/rerun/DM-23243/

> To access all DRP data products, use the following paths on lsst-* machines to instantiate your Gen2 Butler instance:
>
> > (DEEP+UDEEP) **/datasets/hsc/repo/rerun/DM-23243/OBJECT/DEEP/**
> >
> > (WIDE) **/datasets/hsc/repo/rerun/DM-23243/OBJECT/WIDE/**
>
> Exceptions for QA outputs: pipe_analysis outputs are in /datasets/hsc/repo/rerun/DM-23243/ANALYSIS/DEEP/ and /datasets/hsc/repo/rerun/DM-23243/ANALYSIS/WIDE/ validate_drp outputs are in /datasets/hsc/repo/rerun/DM-23243/validateDrp/
>
> Job logs are at /datasets/hsc/repo/rerun/DM-23243/logs/
>
> See individual tickets linked in the Job Summary Table for details running each pipeline

Input dataset: HSC PDR2

1. What data products do we need for the Algorithms Workshop?

- Tracts & fields from NAOJ https://hsc-release.mtk.nao.ac.jp/doc/index.php/database-2/
- Do we need all three layers of HSC-PDR2 WIDE/DEEP/UDEEP? All fields? Eventually yes we want all. Will start all with sfm and use priorities after sfm
- Feb 17 DRP team starts a DRP analysis sprint

Number of visits read from /datasets/hsc/repo/registry.sqlite3 (These are processed by singleFrameDriver)

field\filter	HSC-G	HSC-I	HSC-I2	HSC-R	HSC-R2	HSC-Y	HSC-Z	NB0387	NB0816	NB0921	Total
SSP_UDEEP_SXDS	18	31	5	18		46	53		30	32	**233**
SSP_UDEEP_COSMOS	56	36	104	25	43	212	226		25	50	**777**
SSP_DEEP_XMM_LSS/ SSP_DEEP_XMMS_LSS	35	18		27		30	52	20	22		**204**
SSP_DEEP_ELAIS_N1	76	28	44	43	24	99	142		37	38	**531**
SSP_DEEP_DEEP2_3	48	32	6	47		75	108	28	40	33	**417**
SSP_DEEP_COSMOS	103	40	75	32	74	111	168	26		51	**680**
SSP_WIDE	2519	916	1863	1363	1356	3207	3216				**14440**
SSP_AEGIS	8	7		5		7	7				**34**
SSP Total	**2863**	**1108**	**2097**	**1560**	**1497**	**3787**	**3972**	**74**	**154**	**204**	**17316**
(UH) COSMOS	21	90				67	21			~~7~~	~~206~~ 199 (Ignore 7)
Total											**17151**

Number of visits that are used in coaddition: 2792 visits for DEEP+UDEEP; 11821 visits for WIDE. (Only used those from NAOJ's tract-visits list)

Tract list copied from the HSC release page, the table of "database records":

UDEEP+DEEP	Filters	Tracts
COSMOS	g,r,i,z,y,NB0387,NB0816,NB0921	9569-9572, 9812-9814, 10054-10056
DEEP2-3	g,r,i,z,y,NB0387,NB0816,NB0921	9219-9221, 9462-9465, 9706-9708
ELAIS-N1	g,r,i,z,y,NB0816,NB0921	16984-16985, 17129-17131, 17270-17272, 17406-17407

SXDS+XMM-LSS	g,r,i,z,y,NB0387,NB0816,NB0921	8282-8284, 8523-8525, 8765-8767

In total 39 tracts UDEEP+DEEP.

WIDE	Filters	Tracts
W01 (WIDE01H)	g,r,i,z,y	8994-8999, 9236-9242, 9479-9485, 9722-9728, 9964-9969
W02 (XMM)	g,r,i,z,y	8278-8286, 8519-8527, 8761-8769, 9003-9011, 9245-9253, 9488-9496, 9731-9739, 9973-9981, 10215-10223
W03 (GAMA09H)	g,r,i,z,y	9069-9092, 9312-9335, 9555-9578, 9797-9820, 10039-10051, 10053-10057, 10282-10293, 10296-10298
W04 (WIDE12H+GAMA15H)	g,r,i,z,y	9096-9136, 9338-9379, 9581-9622, 9824-9864, 10079-10084, 10101-10106, 10321-10326, 10343-10348
W05 (VVDS)	g,r,i,z,y	8984-8986, 9206-9233, 9448-9476, 9691-9719, 9933-9960, 10175-10195, 10417-10436, 10659-10677, 10899-10904, 10912-10917
W06 (HECTOMAP)	g,r,i,z,y	15808-15834, 15987-16012, 16162-16186
W07 (AEGIS)	g,r,i,z,y	16821-16822, 16972-16973

The tract IDs for which we have data products in the WIDE layer: tract_id_wide.txt

2. Stack versions, pipeline steps and configs:

To get this running asap, we are comfortable to use different versions for different steps this time.

These use the /software/lsstsw/stack_20191101 shared stack.

- singleFrameDriver.py w_2020_05 default configs
- skymap w_2020_05 default configs
- jointcal w_2020_05 default configs
- fgcm w_2020_06 for buildStars, w_2020_06 + DM-23526 ticket branch for fit and outputProducts.
- skyCorrection w_2020_05 default configs
- coadd w_2020_07   Use FGCM photometry: config.makeCoaddTempExp.externalPhotoCalibName='fgcm' config.assembleCoadd.externalPhotoCalibName='fgcm' config.assembleCoadd.assembleStaticSkyModel.externalPhotoCalibName='fgcm'

The following use the new shared stack at /software/lsstsw/stack_20200220

- multiband w_2020_08
- validate_drp   matchedVisitMetrics.py w_2020_08
- validate_drp   validateDrp.py   TBD
- pipe_analysis   Started with w_2020_08 stack + qa_explorer at commit ab69304 + pipe_analysis at commit 09a7675.   Updated to w_2020_08 stack + qa_explorer at commit 4a24f54 + pipe_analysis at commit c11be5b + obs_base at commit 5b52ea6. See each ticket for details. Use fgcm PhotoCalib.
- pipe_analysis colorAnalysis: lsst_distrib w_2020_15 + lsst_sims sims_w_2020_15 + qa_explorer 4a24f54 + pipe_analysis 6269f0b . Ignored HSC-I2, HSC-R2, NB0387, NB0816
- post-processing w_2020_08
- forcedPhotCcd w_2020_08

Pipeline commands: https://github.com/lsst-dm/s20-hsc-pdr2-reprocessing

Discussions:

- can start with w_2020_03? sfm difference betw 03 and 05: defects map larger in 05
- w_2020_05 is not verified with RC2 yet. But is targeted for starting sfm.
- want jointcal for astrometry & fgcm for photometry.
    - jointcal udeep takes days. Each filter can be on separate nodes. ~3 nodes 5days for the deepest tract. Give it 14+ days of walltime for udeep.
    - To parallelize better, can run photometry and astrometry separately. e.g. run one with doPhotometry=False and the other with doAstrometry=False
- There are long (>60s) and short (30s) exposures. All were processed in sfm. Only long exposures should go to coadd. All will be used in FGCM. Debatable for astrometry. jointcal were already being run with only the long exposures when this was discussed (02/11/2020). The team decided not to re-do jointcal astrometry. Maybe in a new rerun we will include all exposures for jointcal and learn from that.
- Want to run validate_drp on all tracts?
    - validate_drp on master today does not need coadd & multiband. It only needs sfm & jointcal outputs.
    - Jeff's ticket branch adds 4 new metrics (DM-22310) Will use r-band as a reference. All filters depend on r-band data. No other new data dependencies. Maybe only want the new metrics in a few patches.
    - If using the new metrics ticket branch, need to understand the new data flow of validateDrp.py
    - Only validateDrp.py needs the DM-22310 ticket. matchedVisitMetrics.py can start with a weekly release.

    - matchedVisitMetrics.py is now finished for all tracts. On 4/21/2020 we decided not running the new validateDrp.py for this PDR2 reprocessing.
- Want pipe_analysis too. Though lower priority than coadd. Need DM-21052 merged. visitAnalysis and compareVisitAnalysis are the two lowest priorities.
- For the QA dashboard test, expedite the XMM-LSS field for visitAnalysis, coaddAnalysis, matchVisits.py, post-processing
- Question for next time: should we have not produced separate HSC-I & HSC-I2 (R & R2) coadds? (4/7/2020)

3. Infrastructure: compute & disk space – Michelle B is aware and has it under control.

- 2018 reprocessing HSC-PDR1 (DM-13666):   9227.15 node-hour ; output repo ~123 TB
- PDR2 is ~3 times bigger in raw inputs.
- Michelle can get 20 more nodes
- Hsin-Fang's idea is to have a reservation to create a new queue:

  ⚠ IHS-3422 - Jira project doesn't exist or you don't have permission to view it.

- A scheduled maintenance happened on Feb 27 and lsst-dev* were rebooted. Jobs on the worker nodes were not interrupted. Starting Feb 28 a rolling reboot is done on the worker nodes (DM-23690)

4. Human resources from NCSA?

- Michelle is very happy to have Hsin-Fang coordinate, check for errors and that everything is running correctly, but would like to keep Monika involved doing the running to continue building up experience.
- Michelle wants to try to include Felipe

5. Waiting for:

☑ Paul's new calibration set. Paul is copying into /scratch/pprice/CALIB-20200115 everything included. There may be missing data? The calib repo will be at /datasets/hsc/calib/20200115/

☑ sky correction is waiting for sky frame calibration

☑ ~~NAOJ's tract-visits mapping list: Yusra will follow up~~ **Tract-visit mapping:** https://www.dropbox.com/s/f1kv05k5vqv42pv/visitsFormatted_s19a_20200131.lis?dl=0

☑ About the above visit list: We don't have data with visit ID > 138618. Do we simply ignore those new visits? Yes

☑ The above visit list also includes some UH cosmos data (not SSP).   Want to include them too

☑ Need to replace transmission curves per RFC-656 before sfm.

☑ Want DM-23331 & RFC-668 & DM-23434 for fgcmcal

6. Job status and summary

	DEEP & UDEEP	WIDE	Total node-hours
singleFrameDriver	☑ DM-23303 154 jobs slurmIds_udu_all.txt	☑ DM-23301 724 jobs slurmIds_wide_all.txt	2758.08
skymap	☑ slurm job ID: 229995	☑ slurm job ID: 229996	0.02
jointcal	☑ DM-23323 756 jobs slurmIds_jointcal_deep.txt	☑ DM-23395 jointcal_WIDE_success_slurmIds.txt	3466.34
fgcmcal	☑ DM-23394 slurmIds_fgcm.txt		83.45
skyCorrection	☑ DM-23522 31 jobs slurmIds_skyCorr_deep.txt	☑ DM-23522 145 jobs slurmIds_skyCorr_wide.txt	369.50
coadd	☑ DM-23602 378 jobs slurmIds_coadd_deep.txt	☑ DM-23605 3411 jobs slurmIds_coadd_wide.txt	3735.56

multiband	☑ DM-23639 39 jobs slurmIds_multiband_deep.txt	☑ DM-23655 slurmids_multiband_wide.txt	20792.75
post-processing	☑ DM-23856 slurm job ID: 245355	☑ DM-23856 slurm job ID: 246730	152.68
forcedPhotCcd (low priority)	☑ DM-23867 slum job IDs: 246937,247134,247135,247264,247740,247762,247852		3050.69
matchedVisitMetrics (validate_drp)	☑ DM-23654 slurmIds_mvm_deep.txt	☑ DM-23654 slurmIds_mvm_wide.txt	1233.11
~~the new validateDrp.py?~~			
visitAnalysis	☑ DM-23579 slurmIds_visitAnalysis.txt		4298.16
CompareVisitAnalysis (low priority)	☑ DM-23580 slurm job IDs: 247627,247769		724.71
colorAnalysis	☑ DM-23866 slurm job IDs: 47430,247453,247456		26.06
coaddAnalysis	☑ DM-23807 slurmIds_coaddAnalysis.txt		2231.57
matchVisits (qa_explorer)	☑ DM-23831 slurm job IDs: 245159,246742,246820		15.42

7. Reproducible Pipelines Failures - **singleFrameDriver**

DEEP+UDEEP:

301 CCDs failed in UDEEP and their data IDs are in fatals_id_udeep.txt 1730 CCDs failed in DEEP and their data IDs are in fatals_id_deep.txt
Among these 2031 reproducible failures:

- 297 No matches to use for photocal
- 221 RuntimeError: Unable to measure aperture correction
- 28 RuntimeError: Unable to match sources
- 67 No objects passed our cuts for consideration as psf stars
- 1415 InvalidParameterError 'Only spatial variation (ndim == 2) is supported; saw 0'
- 2 TaskError: Fit failed: median scatter on sky = [] arcsec > 10.000 config.maxScatterArcsec
- 1 TypeError 'The metadata does not describe an AST object'

WIDE:

1390 CCDs failed in WIDE. Their Ids are in fatals_id_wide.txt

- 260 : InvalidParameterError: 'Only spatial variation (ndim == 2) is supported; saw 0'
- 1 : RuntimeError: No good PSF candidates to pass to PSFEx
- 839 : RuntimeError: No matches to use for photocal
- 16 : RuntimeError: No objects passed our cuts for consideration as psf stars
- 16 : RuntimeError: Unable to match sources
- 4 : RuntimeError: Unable to measure aperture correction for required algorithm 'base_GaussianFlux': only 0 sources, but require at least 2.
- 22 : RuntimeError: Unable to measure aperture correction for required algorithm 'base_GaussianFlux': only 1 sources, but require at least 2.
- 10 : RuntimeError: Unable to measure aperture correction for required algorithm 'base_PsfFlux': only 0 sources, but require at least 2.
- 35 : RuntimeError: Unable to measure aperture correction for required algorithm 'base_PsfFlux': only 1 sources, but require at least 2.
- 10 : RuntimeError: Unable to measure aperture correction for required algorithm 'ext_photometryKron_KronFlux': only 0 sources, but require at least 2.
- 22 : RuntimeError: Unable to measure aperture correction for required algorithm 'ext_photometryKron_KronFlux': only 1 sources, but require at least 2.
- 6 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_dev': only 0 sources, but require at least 2.
- 26 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_dev': only 1 sources, but require at least 2.
- 6 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_exp': only 0 sources, but require at least 2.
- 26 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_exp': only 1 sources, but require at least 2.
- 10 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_initial': only 0 sources, but require at least 2.
- 37 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel_initial': only 1 sources, but require at least 2.
- 7 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel': only 0 sources, but require at least 2.

- 35 : RuntimeError: Unable to measure aperture correction for required algorithm 'modelfit_CModel': only 1 sources, but require at least 2.
- 2 : ValueError: cannot convert float NaN to integer

8. Reproducible Pipelines Errors - **Jointcal**

Seeing some   ERROR: Potentially bad fit: High chi-squared/ndof. Data IDs are attached in DM-23323 and DM-23395.

(Maybe only in tract with few visits??)

9. Reproducible Pipelines Failures - **skyCorrection**

visit=137268 and 137288 failed with error "No good pixels in image array"; only 1 and 2 calexps exist for these visits; DM-23551 is filed;

Both visits are 30s exposures in NB0387 from 2018-01-14; for continuing the reprocessing campaign, they are not needed in the coadd.

10. FGCM

fgcm_photoCalib products were not written for some visits. See DM-23394 and DM-23698

In total 138 visits miss some fgcm_photoCalib products. Some visits miss fgcm_photoCalib for all CCDs and some for selected CCDs.

The data IDs missing fgcm_photoCalib are

(DEEP+UDEEP) https://jira.lsstcorp.org/secure/attachment/42853/42853_fgcmNoPhoto_deep.txt

(WIDE) https://jira.lsstcorp.org/secure/attachment/42854/42854_fgcmNoPhoto_wide.txt

The missing fgcm_photoCalib means no downstream data for those visits/ccds.

11. Reproducible Pipelines Errors - **coadd**

Among many warnings some also mentioned errors:

- "All pixels masked. Cannot estimate background"
- "No PsfMatched warps were found to build the template coadd ...."   This happens when warp is made but psfMatchedWarp isn't.

See DM-23602.

12. Reproducible Pipelines Failures - **matchedVisitMetrics** (validate_drp)

If a tract+filter only has one visit, the task can't work: DM-23581   So we don't run those cases.

For WIDE, 15 failed with "FATAL: Failed: `ydata` must not be empty".

For DEEP,

- 3 jobs failed with "cannot do a non-empty take from an empty axes" (DM-23981).
- 7 jobs failed with OOM on the cluster workers and seem to require >192G of memory. We decided to only include a subset of the visits for those.

See DM-23654.

Also note that the output are not proper Butler rerun repos; the task isn't writing outputs using Butler.

13. Reproducible Pipelines Errors - **coaddAnalysis** (pipe_analysis)

- "UnboundLocalError: local variable 'axes2' referenced before assignment" DM-23829
- "RuntimeError: No good data points to plot for sample labelled: star"   DM-23894

14. Reproducible Pipelines Failures - **colorAnalysis** (pipe_analysis)

- Okay to let tracts lacking HSC-I/HSC-G/HSC-R data fail; the primary metric doesn't make sense unless you have all 3 of those bands.
- Need lsst_sims

- DM-24328 : Ignore HSC-I2 and HSC-R2. Ignore tracts without HSC-I. This means a much smaller tract list for WIDE. Still some errors look like data too sparse.

15. Reproducible Pipelines Failures - **forcedPhotCcd**

- DM-10755: 354 failed with errors of with lsst.pipe.base.TaskError("Reference %s doesn't exist" % (dataId,))
- 85 failed with lsst.daf.persistence.butlerExceptions.NoResults: No locations for get: datasetType:fgcm_photoCalib
- 2 failed with lsst.daf.persistence.butlerExceptions.NoResults: No locations for get: datasetType:jointcal_wcs

  See DM-23867 for the data IDs.

Compute Time

See the Job Summary Table for the breakdowns.

The total compute for HSC-PDR2 is 31205.7 node-hours up to multiband processing (that is, no forcedPhotCcd, no post-processing, no QA pipelines of any kind, -- same as in the S18 PDR1 run for a fair comparison). Before the execution it was estimated by a simple scaling of multiplying PDR1 by 3 times = 9227.15*3 = 27681.45; that was only ~13% more.

All non-QA pipelines sum to 34409.07 node-hours.

All jobs sum to 42938.10 node-hours.