Data Challenge 3b Performance Test 1.1

Background

DM is motivated to provide science-grade data of limited content and quality as soon as possible to the Collaborations. This will serve two purposes:

- Assist in debugging and measuring DM progress on algorithms
- Assist in science collaborations assessment of whether LSST data products will support their science cases.

Progress with middleware and infrastructure tasks and the application framework has been acceptable and a solid and flexible foundation has been created. Measurable, science-usable output data is just beginning. This means that improving data quality and data release processing of more frames (that can be stacked) is more important than adding more processing stages at the present time. This plan is based on this prioritization, and we do not see a contradiction in this with PDR priorities, in fact we will be much stronger at PDR if we can show this progress. Science Collaboration analysis of PT1.1 results would be highly desirable for PDR. It would also be beneficial to get rapid turn-around from Scientists shortly after processing begins as a sanity check of results.

This page covers the PT Improvement Plan prior to PDR (aka PT1.1 and 2 month plan). It is based on the results of the PT2 Scoping Breakout at the 2010 LSST All Hands Meeting, plus subsequent discussions within the DM team.

Goals

- Run full ImSim data through all current stages (i.e. including through single frame measurement and source association, even though operational system may do this differently, see PT1.1 Pipelines and Stages below for full list)
- Required capabilities/data quality metrics:
    - Science collaborations can create color-magnitude and color-color diagrams
    - ImSim object catalog to compare to output object catalog is required (ImSim instance catalogs are not useful for data analysis)
    - Isolated point source photometry: 0.05 mag
    - Galaxy photometry: 0.07 mag for color or repeatability, measured on ensemble
    - Revisit/improve flexible metadata captured for DC3b data quality requirements/metrics scorecard values (e.g. astrometry, ICP)
    - PSFs with 2 algorithms (Princeton and U Penn)
    - Improve the processed image headers (especially the comment field) to support scientist users (per existing? ticket)

(Note: star-galaxy separation a stretch goal for the 2 month plan; this is ready for integration testing but may not be fast or reliable)

Plans and Actions

Schedule Summary

- 2 months for code changes (15 Aug ? 15 Oct)
- 1 month for runs (15 Oct ? 15 Nov)
- 1 month for analysis and report with overlap (1 Nov ? 1 Dec)

For detailed project plans refer to: DC3 Schedule

For detailed action list refer to: DC3b PT1.1 Actions

Data requirements

ImSim data is deemed most useful by Collaborations; further CFHT-LS data processing is deferred to 6 month plan

- ~600 visits in 7 contiguous pointings (central field and 6 surrounding), all filters, current catalogs
- At 10 visits/day, ImSim team can produce 600 in 2 months wall clock, can do in 1 month wall clock with more disk
- Purdue/OSG and GPU clusters may offer some improvement in rate, and DM will position to be able to process more if it is available, but we will not require more at this time
- Include fixes to missing stars, changes in FITS file headers
- Not addressing Y-band response

(Note: ~400 visits are complete, but it is not clear how many are useful for 600 visits needed, as they must be contiguous, probably an upper bound of 25%. Current plan is to generate all fields, but that requires more work moving/renaming/directory structure/staging/etc.)

User Support

- Data Challenge Handbook needed within ~1 month, with at least basics
- Source (and configuration) documentation to be started. Various procedures (new users accounts, etc.) need to be finalized and documented
- Documentation improvement will be continuous and ongoing, currently being done by scientists, developers themselves. (Requested dedicated resources added in FY12 request).
- Science collaborators will interact on science wiki forum, PT1 team will periodically review forum, a digest will be sent from forum to those who subscribe to it. This is to facilitate both pushing information from DM to the Collaboration scientists, and getting the various collaborations to coordinate cross-collaboration efforts.
- There will be a ?help now? button on the forum for immediate forwarding request to PT1 team via email. This email will go to TBD User Support contact (Dick Shaw plus TBD others on list) for triage & forwarding to appropriate team member. This link will also exist on Gator page. A help account already exists on IPAC server.

(Note: the email address needs to reflect the LSST domain, not IRSA. Also, it was not clear whether the Help Desk software has been configured at IPAC, or indeed if this is still the plan.)

Performance & Reliability

- Reported performance issues with Middleware have improved significantly since last monthly report according to Ray/Robyn?
- Need review all existing usage of logging facility

- Need full abe runs
- Significant improvement expected by grouping 6 pipelines into 2
- Joboffice work is to be completed for these runs
- Run master role is to be performed by NCSA
- Need performance statistics from full abe runs for PDR, including ability to independently assess middleware vs apps, SDQA vs Data Release Production, I/O vs processing, etc.
- Need to achieve level of instrumentation we had in DC3a, i.e. instrumented the code and loaded stats into database
- Speed up convolution, warping

Environment and Tools

- At least weekly buildbot for performance testing / regression testing of performance
- Upgrade of stack to be approved by TCT asap: includes gcc 4.4 plus boost patch, eigen 2.0.15. We will still support gcc 4.3.3 (with work-around from Mike Jarvis, Robert Lupton).
- Continuous integration of the *trunk* pipelines Robyn (build off development modules, trunk against trunk)
- Guarantee consistent weekly builds/build manager, including automation with buildbot
- The ability to get how-to-repeat cases (tarballs?) that can repeat failures on at least the same cluster as original run, and much more preferred on a smaller/other machine.
- A way for developers to run the "latest" code on a specified volume of data, with outputs in a well-defined place, documentation/training on existing scripts that do this.

PT1.1 Pipelines and Stages (Data Release Production)

Pipeline: Instrument Signature Removal

- isr_initialize: Acquire a single channel's image data
- isr_saturation: Do saturation correction
- isr overscan: Do overscan subtraction
- isr_variance: Calculate variance from image counts
- isr_dark: Do dark subtraction
- isr_flat: Do flat subtraction
- isr_sdqa: Generate SDQA metrics
- isr_output: Output initial corrected channel image (post-ISR) and SDQA metrics

Pipeline: CCD Assembly

- ca_initialize: Acquire a CCD's worth of post-ISR channel image data
- ca_assembleCCD: Assemble appropriate channels into a CCD
- ca_isrCcdDefect: Mask out CCD defects
- ca_isrCcdSdqa: Calculate additional metrics characterizing assembled image
- ca_sdqa: Package metrics for output
- ca_output: Output assembled image and SDQA metrics

Pipeline: Cosmic Ray Split

- cs_initialize: Acquire a visit's worth of post-ISR CCD image data
- cs_backgroundEstimation: Do background estimation

- cs_reject: Estimate and possibly subtract cosmic rays from exposure
- cs_output: Output final modified image and SDQA metrics

Pipeline: Image Characterization

- ic_initialize: Acquire visit image data
- ic_sourceDetection: Detect 'best and brightest' sources on an exposure
- ic_sourceMeasurement: Measure 'best and brightest' sources on an exposure
- ic_psfDetermination: Given exposure and sources measured on that exposure, determine a PSF for that exposure
- ic_wcsDetermination: Validate Wcs for image using astrometry.net package and calculate distortion coefficients
- ic_wcsVerification: Compute statistics that indicate if calculated WCS is a good measure
- ic_photocal: Calculate magnitude zero point for a SourceSet for an image that has been matched to a corresponding SourceSet for a catalogue
- ic_output: Output measurements and SDQA metrics

Pipeline: Single Frame Measurement

- sfm_initialize: Acquire visit image data
- sfm_sourceDetection: Detect all sources on an exposure
- sfm_sourceMeasurement: Measure all sources on an exposure
- sfm_computeSourceSkyCoords: Compute the sky coordinates of sources
- sfm_output: Output source catalog and SDQA metrics

Pipeline: Source Association

- sa_initialize: Acquire source catalog
- sa_SourceClustering: Determine which sources belong to the same object
- sa_SourceClusterAttributes: Characterize the objects
- sa_output: Output object catalog and SDQA metrics