HSC PDR1 Reprocessing

Hsin-Fang Chiang & John Swinbank, 2017-04-10 (299babb)

Summary and Timescale

This note summarizes the "HSC PDR1 reprocessing", which the DRP and NCSA groups aim to carry out on the LSST Verification Cluster in cycle S17B (ie, March, April, May 2017). It touches on how that work will transition into F17, and on its potential impact on other groups.

Goals

S17B

- Instantiate and run the Batch Processing Service as a façade at NCSA.
  - The Batch Processing Service is a production service to be provided by the LSST Data Facility at NCSA. The batch processing service is used for executing offline campaign processing, including large-scale pre-DRP tests, DRP, after-burners, etc. The service receives pipelines as payloads and runs them in the LSST production environment.
  - This specific activity deploys the first phase of the DPP Batch Processing Service. The goal is to begin prototyping production processes, including general operations of the batch processing service, incident response, problem management of both services and of data products, possibly insertion of changes, and interactions with science operations. In this phase, the service is presented as a "façade", as processes are maturing and the reliant services are still under development. The service provides feedback on operational feasibility to development
- Demonstrate large scale data release production processing and provide a large processed dataset for use by DRP and other developers in assessing and debugging large-scale performance and fidelity of the data release process.

Future Cycles

- Provide Science Pipelines developers with a convenient resource for testing and debugging.

  - To meet this goal, we should provide developers with convenient access to data processed through effectively arbitrary pipeline versions and configurations (e.g. latest release, latest weekly, using meas_mosaic, using Jointcal, ...).
  - The aim is to make access to data as low overhead and latency as possible, so that developers can make use of this as part of their regular workflow.
- Satisfy wider goals for test dataset availability which may have ramifications across the project. An obvious example of this is RFC-243, but we should expect that further use cases emerge in connection with the science validation effort, the PDAC, etc.

Effort available in S17B

NCSA

- Hsin-Fang Chiang has limited availability over the rest of the S17 cycle (to end May 2017). This is likely adequate to shepherd one data release through on the Verification Cluster, including checking that all processes have executed successfully and that data has been stored properly in the GPFS `/datasets` filesystem.
- Ongoing effort to maintain and improve the Verification Cluster.
- Ongoing effort to develop the workload management system.

Princeton

- Lauren MacArthur will devote 100% of her on-project time through S17 to the "QA" effort (loosely defined as finding and resolving issues apparent from large scale data processing). (It is likely that this will continue through F17.)
- Tim Morton will also contribute to this effort. At time of writing he is still onboarding, so his effort will ramp up given time.

Supporting Technical Infrastructure

S17B

- Work will be carried out using the ctrl_pool distribution middleware and pipe_drivers top-level scripts.
  - Using this system, it is possible to dispatch tasks to process the entire data release with a small number of commands.
  - However, it is relatively labour-intensive to check for successful pipeline execution. This is likely the major timesink when processing.
  - No effort is scheduled to improve the ctrl_pool package.

- Access to the Verification Cluster is available to all members of DM, and is documented at https://developer.lsst.io/services/verification.html. It already provides a SLURM environment which is compatible with the ctrl_pool middleware. All users may submit jobs without special permission.
- We do not foresee significant issues of contention or resource usage on the Verification Cluster during S17B.
- No effort is scheduled to ingest data into Qserv or any other database system during S17B.

Future Cycles

- Longer term, we will switch to a system based (probably) on SuperTask and Pegasus. That will resolve the problems discussed above, but the timescale is likely many months; we do not consider it further here.
- In order to provide adequate capacity for future processing, a more detailed understanding of resources needs and requirements must be developed.

Programme of Work

S17B

- Hsin-Fang will perform one complete reduction of HSC PDR1 using a TBD version of the stack.
  - The stack version will be a recent weekly at the time the work starts, and will likely include the full set of fixes included in the hscPipe RC. Hsin-Fang and the DRP team will directly agree on this in person.
  - A fixed set of configuration parameters will be provided to Hsin-Fang by the DRP team before work begins.
- Hsin-Fang will first process a small set of data, namely the HSC RC dataset. Lauren will perform basic QA with the RC results, and confirm the software version, setups, and configs, before Hsin-Fang starts processing the full set of data.
  - Hsin-Fang will monitor execution progress & resolve issues (calling on DRP / Lauren for help where necessary) to ensure that all data is successfully generated.
  - No science validation will be performed on the generated data products by the NCSA group.
- If the chosen version of the stack is unable to process some specific data, tickets will be filed with information to reproduce the problems. Those inputs may be removed from the S17B reprocessing.

- The procedure followed will be documented by Hsin-Fang (as a delta on existing pipelines documentation.
- No further support for large-scale processing is expected from NCSA during S17B.

F17

- The various stakeholders in the larger "QA effort" are asked to flesh out their requirements for large-scale processing through F17. These stakeholders are presumed to include:
  - The Data Release Production development team (action item on Lupton, Bosch, MacArthur Swinbank);
  - The DM Project Scientist and DM-SST (Juric, Lupton, etc, especially to include the Science Validation Scientist if available);
  - Other potentially interested parties, most notably those with an interest in RFC-243 (Dubois-Felsman, Wu, etc).

  The DMLT are asked to confirm and/or expand this list.
- These stakeholders are all requested to provide a proposed set of requirements and resource estimates for large-scale processing during F17 no later than end of April (so that it can be included in the NCSA cycle plan)
- We note that some of this may be dependent on the forthcoming V&V plan, but caution that the mandatory cycle planning cadence means that some requirements may need to be expressed before that is fully agreed.
- The plan for F17 will be driven by the requirements provided in response to the above.