NOAO SDM Document DEC001

Integrating the DESDM-CP Into NOAO/E2E

F. Valdes[1]

National Optical Astronomy Observatories
Science Data Management

V1: May 7, 2012

[1]NOAO Science Data Management, P.O. Box 26732, Tucson, AZ 85732

Table of Contents

Abstract

This document describes how the Dark Energy Camera Community Pipeline (DECCP), provided by DESDM, is integrated into the NOAO/SDM end-to-end (E2E) system. This includes use of the NOAO Science Archive (NSA), the Pipeline Scheduling Agent (PSA) and Pipeline Scheduling Queue Database (PSQDB), and the ”Save-the-Bits” (iSTB) archive submission gateway. It also makes use of an NOAO High Performance Pipeline System (NHPPS) pipeline application to provide an additional, higher-level orchestration wrapper to the DECCP. The integration will then be operated very similarly to the other NOAO/SDM operated pipelines (MOSAIC, NEWFIRM).

Keywords: DECam, pipeline, E2E

Purpose of this Document

This is a design document laying out the strategies and justifications for how the DECCP is integrated into the NOAO E2E system.

1 Introduction

The Dark Energy Survey Data Management Community Pipeline (DESDM-CP) is an externally developed pipeline delivered to NOAO to calibrate Dark Energy Camera (DECam) data. At NOAO this pipeline is called the Dark Energy Camera Community Pipeline (DECCP). When operated by NOAO the DECCP must be integrated with other NOAO pipeline systems and components. A goal of this integration is for this pipeline to appear and be operable in the same way as other NOAO operated pipelines (e.g. MOSAIC and NEWFIRM). This allows operators to be more easily trained for multiple pipelines and to use common SDM operations tools.

The primary system for the DECCP must be integrated with is the NOAO Science Archive (NSA) which provides the raw data to be processed and consumes the pipeline data products. In outline, this involves the pipeline operations interface querying the archive for available data based on knowledge of the telescope schedule, accessing the data through an archive interface, receiving submissions of data products, and ingesting the data (placing in permanent storage and registering metadata about the data with the archive database).

At the next level down the components the pipeline integrates with are the NOAO pipeline scheduling agent (PSA) and queue database (PSQDB), the archive data service or mass-store system, and the *save-the-bits* (iSTB) data handling system with handles submission of files to the archive.

The integration described here makes use of another component, the NOAO High Performance Pipeline System (NHPPS) [2]. While this is an orchestration system and the DECCP also has its own orchestration system, NHPPS provides a integration layer between the DECCP and the other components. The ”pipeline application” which it defines provides pipeline activities which are outside of the DECCP and for which there already exists E2E integration methods. Since an NHPPS pipeline calls host commands, and the current DECCP native interface consists of host commands the transition between the two is trivial.

The NHPPS pipeline application performs the prequisites for running the DECCP – staging of data, entering this into the DECCP database (ingest), setting up a pipeline configuration file, and submitting the configuration file for execution with the `dessubmit` command. At the end of the DECCP processing a trigger event is generated for which the NHPPS pipeline is waiting. After receiving this trigger, a set of final steps are performed. This includes notifications to the PSA/PSQDB, setting up operator review information, and setting up the data products for submission. In the NOAO operations model [5] the actual submission of data to the archive is initiated by the operator, after a review, through an operator interface rather than automatically by the pipeline.

This integration using an NHPPS pipeline ”wrapper” application is another example of the concept of ”marrying” two pipeline systems to make use of what is best from both for a particular sit-

uation. The other case, [3] and [4], is using the OGCE workflow orchestrator as a wrapper around NHPPS pipelines for processing ODI data in the XSEDE environment. In a sense the DECCP integration is the converse of that, namely converting a pipeline system designed for XSEDE for use on a dedicated cluster.

2 PSA and PSQDB

The PSA/PSQDB system is the primary operator interface for defining, submitting, and tracking datasets for NOAO pipelines. The interface is currently through a browser which has buttons and links to perform various tasks. For more details see [1].

The database contains definitions of datasets (also known as campaigns in the DES lexicon) based on the telescope schedule. The definition is interpreted by the PSA (activated either manually by the operator or automatically based on the current time or completion of a previous dataset) to produce a query to the NSA which returns references to raw exposures available from the archive. The list of references is passed through a script that can eliminate exposures; usually because they have been successfully processed already. The remaining list of references is the trigger for an NHPPS pipeline application.

At the end of processing the NHPPS pipeline will ultimately notify the PSA of the completion of the processing so it can proceed to another dataset. During the processing and at the end, processing status information is also captured and used to update status information in the PSQDB. The operator can then monitor this status in the database through the operator interface. Note that this is high level status dealing with the record keeing of success or failure of the processing and not the lower level monitoring of the processing for which a different monitoring tool is appropriate.

3 The NHPPS DECCP Pipeline Application

As described in the introduction, the NHPPS DECCP Pipeline Application is a wrapper layer for the DECCP. While it is more than just a simple scripting wrapper, it logically has the same function as any wrapper paradigm in that it interfaces inputs and output to an eventual call to the component it wraps.

The DECCP has the concept of blocks which execute host commands to perform a particular operation that is part of a larger set of operations. One is free to put these blocks together as needed to form a single orchestrated pipeline application. So a large number of blocks can be combined to do everything in one submission, a smaller set of blocks for specific categories of processing, or even just one block at a time. Note that the DECCP orchestrator is based on using a job submission system, which at NOAO is *condor*.

NHPPS pipeline applications also have similar concepts of pipeline applications and smaller pieces called pipelines and modules. Pipelines perform some particular function and internally may consists of a small or large number of steps. Each of these pipelines can be distributed and/or run in parallel.

Without going into great detail, we have the flexibility to define how many DECCP blocks are peformed in a DECCP submission and combine these at a higher level as needed. The integration design is to use NHPPS to convert and orchestrate the processing of calibrations (biases and dome flats) and science exposures by filter. Note biases can be considered a case of a ”dark” filter. This structure is basically the same as all the NOAO pipelines including that for ODI (which is designed by NOAO but is not operated by NOAO). The principle behind these is the PSA/PSQDB tracking of dataset processing which extends down to tracking the processing of each of these groupings and therefore allows reprocessing of just those subsets requiring it. In other words, once a dataset has been processed the operator can decide to reprocess just one filter while submitting other filters to the archive.

3.1 Staging

The result of submitting a dataset or subset of a dataset by the PSA is a list of archive file references. The first thing the NHPPS wrapper application does is check if the data has already been staged and registered for the DECCP and, if not, take care of this.

The concept of ”staging” is a logical one and not necessarily physical. In particular, NOAO and DESDM-CP discussed the possibility of staging as consisting of mounting the required files as read-only directly from the archive mass storage system. In this case staging is just this mount but there would still need to be an interface step to make files appear to be staged, as currently required, as a subdirectory of a higher ”Archive” directory. For physical staging the NHPPS pipeline would do just what it does with MOSAIC and NEWFIRM data of using the archive data service to get the files and move then to a staging area (i.e. the ”Archive” structure of the DECCP processing model).

Another prepatory step is separating the data into the logical groupings to be processed. As discussed above, NOAO will do this by filter. So the data will staged to different archive directories (again either with links or physically). Note that in the DECCP lexicon these directories are called ”run” directories but a run can be defined as desired and not just by night or campaign of nights as done for the DES processing.

Once the files are ”staged” the DECCP provides a command to ”ingest” the files, which means registering the locations and some other metadata in a database. The DECCP system then makes use of this database for identifying files rather than explicit path arguments.

3.2 Orchestration

The NHPPS pipeline will orchestrate the groups as is done for other NHPPS pipeline applications. Specifically it will process biases, then dome flats, and then science exposures. In a purely NHPPS pipeline application one or more pipelines are ”called” in parallel. NHPPS has a useful feature of controlling how many may be in parallel and saving those not yet executed to be submitted when another group completes.

Whether we will use a layer of these ”calibration” and ”filter” pipelines or just directly execute the groups with the DECCP needs to be determined. Note that at whatever level the transition

between NHPPS and DECCP occurs it is just a matter of exchanging the NHPPS logical concept of *calling* another pipeline service by a trigger event with a ”dessubmit”. In other words a ”call” translated into a ”dessubmit”.

A ”return” is capturing of an event from the called pipeline, be it NHPPS or DECCP. The interface for a trigger from DECCP to NHPPS needs to be determined but NHPPS provides simple ways to generate the event along with status information so it is just a matter of deciding on an impplementation.

3.3 Post-DECCP Processing

Once the transition back to the NHPPS wrapper application is made steps are orchestrated for any data product handling, status updating, and possible archive submissions. This will be described in more detail in later versions of this document. For now we just say that essentially the same types of operations and modules as in other NOAO NHPPS pipeline applications will be performed.

References

[1] F. Valdes. Pipeline Scheduling Operator Interface. SDM Pipeline Document PL026, NOAO/SDM, Oct 2011. http://chive.tuc.noao.edu/noaodpp/Pipeline/PL026.pdf.

[2] F. Valdes, T. Cline, F. Pierfederici, B. Thomas, M. Miller, and R. Swaters. The NOAO High-Performance Pipeline System. SDM Pipeline Document PL001, NOAO/SDM, Oct 2006. http://chive.tuc.noao.edu/noaodpp/Pipeline/PL001.pdf.

[3] F. Valdes and S. Marru. The Marriage of Mario (NHPPS) and Luigi (OGCE). Draft SDM Pipeline Document PL023, NOAO/SDM, Dec 2010. http://chive.tuc.noao.edu/noaodpp/Pipeline/PL023.pdf.

[4] F. Valdes and S. Marru. The Marriage of Mario (NHPPS) and Luigi (OGCE). In I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots, editor, *Astronomical Data Analysis Software and Systems XX*, volume 442 of *Astronomical Society of the Pacific Conference Series*, pages 211–+, July 2011. Also http://chive.tuc.noao.edu/noaodpp/Pipeline/PL023.pdf.

[5] F. Valdes, D. Scott, N. Zarate, and R. Swaters. NOAO Science Pipelines Operations Model. SDM Pipeline Document PL012, NOAO/SDM, Jan 2008. http://chive.tuc.noao.edu/noaodpp/Pipeline/PL012.pdf.