Pipeline Scheduling Operator Interface

F. Valdes[1]

National Optical Astronomy Observatories
Science Data Management

DRAFT: October 19, 2011

[1]NOAO Data Products Program, P.O. Box 26732, Tucson, AZ 85732

Table of Contents

List of Tables

List of Figures

Abstract

The Pipeline Scheduling Operator Interface is the primary operator tool for managing automatic pipeline triggering, tracking pipeline processing, reviewing results, and submitting pipeline data products to the archive. It has a number of other features, such as selecting subsets for reprocessing with parameter changes and notifying principle investigators of pipeline calibrated data availability. This document describes the look and options of the current NOAO PS interface using examples from the Mosaic Pipeline Application. Other pipeline developers may find this document useful as a set of features and functions found useful for pipeline scheduling and management.

Keywords: pipeline operations, pipeline triggering, pipeline scheduling

Purposes of this Document

This document describes the features of the NOAO pipeline scheduling operator interface for use by new NOAO operators and for collaborator interested in including similar functionality in their pipeline operations.

1 Introduction

In the NOAO pipeline operations model [1] observational data is processed in datasets. These datasets are a logical grouping of data, typically from multiple nights. Alternate names for this type of definition are blocks or campaigns. The definition of these datasets is provided by a pipeline scheduling queue database (PSQDB) [1]. In addition to defining the datasets the PSQDB has additional roles. These include triggering (aka launching) processing through a pipeline scheduling agent (PSA) [1] and keeping track of the processing status of various datasets.

Because the PSQDB is used as the main bookkeeping database for managing pipeline datasets it is natural for it to be the basis of a operator user interface [2]. This document describes and illustrates the user interface and the functions it provides. Because of its relatively simple design it is quite easy to add functionality. Basically, any operation that operates on a dataset can be linked to this interface.

The main functions of this pipeline scheduling operator interface are:

- displaying queue and dataset definitions
- managing the automatic pipeline triggering
- allow explicit triggering of a dataset
- tracking pipeline processing
- reviewing pipeline processing results
- submitting data products to the archive
- resubmitting data subsets for reprocessing with different parameters
- notifying PI's of completed processing if desired

These functions lead to the features of the interface. Table 1 provides a summary of the features described in more detail later in this document.

As a clarifying note, the structure of the information displayed in the operator pages is not precisely the structure of the tables in the PSQDB. They are related, of course, but the cgi script that reads the information from the database is free to display it as desired by the developer. We point this out if one is also studying the PSQDB documentation [1].

A convention is used where links to forms are shown as buttons while links to other pages are shown as normal browser links.

A useful capability of the interface implementation is that NOAO operators can access it remotely, using firewall control mechanisms such as VPN, at home and with mobile browsers such as an iPhone/iPad.

Table 1: Pipeline Scheduling UI Feature Summary

- display queues defined in the PSQDB
- enable/disable "queues" for automatic submission
- display defined datasets in the PSQDB by queue

  - show status, time submitted, completed, PI notification
  - show comments

- link to form to explicitly submit a dataset for processing
- link to form to change dataset status
- link to send notification to PI of dataset(s)
- display processing status and comments for filter subsets
- link to pipeline results for a filter subset
- link to telescope schedule and proposal
- link to

  - submit filter subset to archive
  - flag to resubmit for processing
  - flag to delete entry left over from failed processing

2 Queues Page

Datasets can be organized into queues. The grouping criteria for a queue is arbitrary. At NOAO the queues are defined by observing semester and telescope. Figure 1 shows a portion of the current NOAO Mosaic Camera Queue page. This page is ordered by state and then by queue name. For historical reasons the queue name and queue are identical and one of them should be eliminated.

There are two operator links on this page; one for focusing on a particular queue and one for modifying the state of the queue.

Selection of a queue link shows a page as described in §3.

Figure 2 shows the result of clicking on the state link. It shows an editable summary of the queue. The editable part is a menu for the state field. It currently has two values "enabled" and "disabled". The role of this field is for automatic triggering of datasets. It is ignored when the operator manually triggers specific datasets.

Pipeline Scheduling Queues

psqname	queue	application	pipeline	state
C4M04B	C4M04B	MOSAIC	dir	enabled
C4M11B	C4M11B	MOSAIC	dir	enabled
K4M04B	K4M04B	MOSAIC	dir	enabled
K4M11B	K4M11B	MOSAIC	dir	enabled
C4M05A	C4M05A	MOSAIC	dir	disabled
C4M05B	C4M05B	MOSAIC	dir	disabled
C4M06A	C4M06A	MOSAIC	dir	disabled

Figure 1: Queues Page

Updating Queue K4M11B

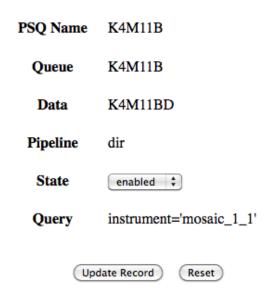|	
PSQ Name	K4M11B
Queue	K4M11B
Data	K4M11BD
Pipeline	dir
State	enabled
Query	instrument='mosaic_1_1'

Update Record   Reset

Figure 2: Modifying the Queue State

3 Queue Page

A queue page (see figure 3) shows the definition and status of datasets in a particular queue. The fields are the dataset name, the processing status, the processing submission and completion times, notification link or timestamp, the query defining the dataset, and operator comments.

The dataset name is a link. If the status is "pending" then the link is to the telescope schedule otherwise it is to the dataset processing page. The telescope schedule and the processing page are

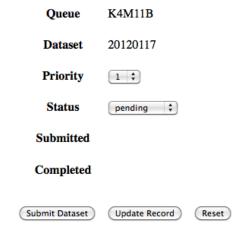| Dataset	Status	Submitted	Completed	Notified	Query	Comments
20111007	completed	2011-10-18T19:05	2011-10-18T20:11	2011-10-13T10:21:20	instrument='mosaic_1_1' and start_date between timestamp '2011-10-07' and timestamp '2011-10-09'	The gain changed on the last night requiring division into separate nights and sky flats to take the place of missing dome flats.
20111024	error	2011-10-18T19:05	2011-10-18T20:11		instrument='mosaic_1_1' and start_date between timestamp '2011-10-24' and timestamp '2011-10-25'	missing dome flats
20111026	hold				instrument='mosaic_1_1' and start_date between timestamp '2011-10-26' and timestamp '2011-10-27'	enter
20111028	resubmit	2011-10-18T19:05	2011-10-18T20:11		instrument='mosaic_1_1' and start_date between timestamp '2011-10-28' and timestamp '2011-10-30'	No sky flat
20111031	nodata				instrument='mosaic_1_1' and start_date between timestamp '2011-10-31' and timestamp '2011-11-02'	enter
20111220	submitted	2011-10-20T12:34			instrument='mosaic_1_1' and start_date between timestamp '2011-12-20' and timestamp '2011-12-23'	enter
20120117	pending				instrument='mosaic_1_1' and start_date between timestamp '2012-01-17' and timestamp '2012-01-18'	enter

Figure 3: Datasets in a Queue

Queue K4M11B

Dataset 20120117

Priority [1 ▼]

Status [pending ▼]

Submitted

Completed

[Submit Dataset] [Update Record] [Reset]

Figure 4: Status Update Page

Add comments for dataset: K4M11B_20110915

[Add] [Reset] [Cancel]

Figure 5: Comments Page

discussed in the next section.

The status field is important to indicate the processing status and to control the automatic triggering by the PSA. Note that manual operator triggering is not affect by this status. The possible status values are given in table 2. The status value is always a link to a form allowing the status value to be changed or to submit the dataset explicitly (see figure 4).

Table 2: Queue Page Dataset Status Values

pending: available for processing

submitted: currently processing

completed: processing completed

resubmit: all or some of the dataset needs reprocessing

hold: not available for processing

lock: not yet available for processing (being staged, etc.)

nodata: processing was attempted but no data was found

error: processing error occurred

The notified field allows the operator to notify the principle investigator when all the processing of a dataset has been completed, the results deemed acceptable, and the data products submitted to the archive. The link is to a cgi script that sends a standard automatic email to the PI. Figure 9 shows an example email. When the email has been sent the field is set to the time.

The "query" field is a fragment of SQL used to query the NOAO Science Archive (NSA) for data to process. The component that uses this query is the PSA which defines the rest of the SQL. As noted earlier, the NOAO convention for datasets is primarily a block of nights at a particular telescope (which is taken to be equivalent to the instrument). Therefore, the typical query, as seen in the example, selects by instrument and the calendar date starting at noon for the night. Note it might be tempting to have the UI allow the operator to modify the query but this has not been found to be needed and the operator may modify the PSQDB directly with the MySql client.

The comment field links to a page to enter comments as shown in figure 5. If there is no comment then a button is shown. If there is a comment then clicking on it allows the comment to be editted or expanded.

4   Dataset Processing Page

The dataset processing page is accessed through the dataset link in the queue page. This page is created, extended, and modified based on dataset processing by the pipeline. In other words, as a dataset starts processing a subset it creates a record in the PSQDB. When it reaches certain points

or finishes processing it updates fields in the record. Because new subsets only start when previous subsets complete the number of records grows during the processing of a dataset. Figure 6 shows an example of a dataset processing page.

plqueue	pldataset	plsubset	plprocid	plsubmitted	plcompleted	plarchived	plcomments
K4M11B	20111003	r1F	7db1e30	2011-10-07T12:51:38		delete	enter
K4M11B	20111003	Zero	7db1d30	2011-10-07T11:36:49	2011-10-07T12:02:38	2011-10-07T12:07:53	enter
K4M11B	20111003	IF	7db1d30	2011-10-07T12:02:46	2011-10-07T12:18:00	2011-10-07T12:24:45	enter
K4M11B	20111003	BF	7db1d30	2011-10-07T12:02:44	2011-10-07T12:18:12	2011-10-07T12:25:05	enter
K4M11B	20111003	RF	7db1d30	2011-10-07T12:18:04	2011-10-07T12:35:12	2011-10-07T12:42:17	enter
K4M11B	20111003	UnF	7db1d30	2011-10-07T12:18:17	2011-10-07T12:35:53	2011-10-07T12:42:45	enter
K4M11B	20111003	VF	7db1d30	2011-10-07T12:35:20	2011-10-07T12:51:06	2011-10-07T13:30:05	enter
K4M11B	20111003	gF	7db1d30	2011-10-07T12:36:00	2011-10-07T12:51:31	2011-10-07T13:30:26	enter
K4M11B	20111003	iF	7db1d30	2011-10-07T12:51:11	2011-10-07T13:07:04	2011-10-07T13:30:45	enter
K4M11B	20111003	zF	7db1d30	2011-10-07T13:07:08	2011-10-07T13:21:26	submit redo	enter
K4M11B	20111003	B	7db1d30	2011-10-07T13:21:32	2011-10-07T13:36:20	active	enter
K4M11B	20111003	I	7db1d30	2011-10-07T13:21:33	2011-10-07T13:38:45	redo	Poor dithering
K4M11B	20111003	R	7db1e30	2011-10-07T13:36:24	2011-10-07T14:07:38	alt version	enter
K4M11B	20111003	R	7db1d30	2011-10-07T13:36:24	2011-10-07T14:07:38	submit redo	enter
K4M11B	20111003	V	7db1e30	2011-10-07T13:38:55	dpsreview	delete	enter

Figure 6: Dataset Processing Page

Because entries are added to the PSQDB by the pipeline during and after completion of processing this UI page makes use of the simple html refresh function to keep the page current without requiring the operator to regularly reload the page. So in typical operation, the operator will have his/her browser showing the dataset processing page for the dataset being processed.

As datasets complete and the operator links described in this section appear, the operator can review the processing and either submit the subsets to the archive or mark them for reprocessing with any appropriate pipeline configuration parameters modified. This makes use of a feature in the NOAO pipeline applications that default parameters can be overridden by creating a file with a pattern that matches one or more datasets. This allows setting pipeline parameters specific to a single subset or all subsets in a dataset or all datasets from a particular queue.

This queue link in the first column provides a simple way to return to the queue page the the dataset belongs to. There is a link at the top of the page to go directly to the top level list of queues.

The dataset link opens a page to the telescope scheduling page for the observatory and date of the dataset. This is specifically tied to being able to parse the queue and dataset names to identify the observatory and data. The NOAO observatories, KPNO and CTIO, provide public telescope pages which include not only the current semester by back many years to the start of this capability:

```
http://www.noao.edu/kpno/forms/tel_sched/
http://www.noao.edu/ctio/forms/tel_sched/
```

The operator can find this function useful to check who the PI was and to understand the program or mix of programs and to understand the program goals through the public proposal abstracts

linked from there. Knowing something about the program can be interesting and provide clues about what might be the best way to process the data.

While the pipeline review page for a subset is not generated by the PSQ UI, the subset link provides access to these pages. This is one of the main purposes of the pipeline scheduling interface. Figure 8 shows an example of a review page from the Mosaic pipeline to illustrate why this is such an important part of the operator interface.

A subset record is created with the pipeline first starts to process it. The link associated with the value in the subset column only appears when the plcompleted column has a timestamp or a value of "dpsreview", indicating the pipeline review page has been created. If the value is "dpsreview" the review page may still be in the process of being populated or it may have been completed but downstream steps, at this point it is mostly the final retrieval of data products to the staging area, have not been completed.

The values in the plarchived column can take a variety of values, some of which are links, as shown in table 3. The operator may set other values or reset a value with other interface tools or direct sql updates to the PSQDB.

The pipeline does not automatically submit pipeline results to the archive. This is to allow the pipeline operator to review the results and decide if they are as a good as the pipeline can produce. So the PSQ UI provides a link in the plarchived column. This is done by subset so that only smaller pieces of the main dataset block can be archived or reprocessed.

If the submit link is selected the plarchived value becomes either "active" or "pending" and after completion it is the time the submission was completed or "error". Note that the time of completion for the submission is is from the pipeline's perspective. The data submitted through the pipeline-archive interface may take some significant amount of time to actually be completely ingested into the archive for users to retrieve.

The submission process can also report an error for the operator to handle. The main sources of an error are the existence of a earlier version of the data product or the data product not being found in the staging area. These conditions occur for various good operational reasons that the operator expects. The operator then uses the command line tool to reset the error to a timestamp.

Note that one point of the UI is that the submission link can only appear when it makes sense for the data to be submitted. If the dataset is marked for reprocessing or as an alternate version the operator cannot accidentally submit the dataset. The command line tool can be used to override the logic to force the submit link to appear.

Instead of the submit link the operator may select the redo button. When a subset completes processing a record of the raw exposures used is kept. When a dataset is (re)submitted the pipeline scheduling agent makes a list of all the exposures in the archive which satisfy the definition in the PSQDB. It then eliminates all those recorded as processed. What this means is that any resubmissions will only process those not already processed. Therefore, one of the things, the main thing, the redo cgi command must do is remove the record of those exposures having been processed. This is done in such a way that this can be undone by the operator if needed.

The redo button goes to an page to edit parameters which will be specific to that particular dataset. This page is currently in an early stage of development. An example is shown in figure 7. Using the Submit button activates a cgi-script which sets the parameters, sets the value in the

database for the subset to "redo", and set the status of the full dataset to "resubmit". This is then reflected in the UI for the operator to keep track of this.

The comment field behaves in the same way as described in the previous section and shown in figure **??**.
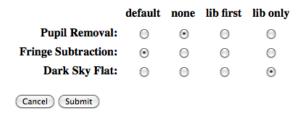

Figure 7: Example Redo Form (still under development)

5 Implementation

The pipeline scheduling database engine is MySql. The cgi scripts that render the http pages is in Python using the available MySql interface library. The action links, e.g. submitting to the archive, are cgi calls to C-shell scripts.

The html is fairly basic. The main setup requirements are to allow the web users (noboby) to access the database and pipeline review pages. The only security is that the interface is only available within the NOAO firewall.

6 Related Interface Tools

The tool that is used within the cgi scripts and the pipelines to add the time stamps or change the status are:

```
psqpltable
psqreset
```

These are C-shell scripts calling the MySql client program.

References

[1] 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.

[2] F. Valdes and N. Zarate. Mosaic Pipeline Operator's Manual. SDM Pipeline Document PL004, NOAO/SDM, Nov 2006. http://chive.tuc.noao.edu/noaodpp/Pipeline/PL004.pdf.

Table 3: Values and Links in the Archiving Column

delete: This link deletes the processing entry from the database. This is only used when the pipeline is stopped during processing and the operator needs to clean up history of this processing.

submit: This link triggers submission of the data products to the archive. This is one of the main operator functions. The cgi script triggers an NHPPS pipeline to allow sequencing of multiple submissions rather than possibly flooding the archive ingest interface. Because of this there are two possible values during submission as noted below.

> **active:** When the submit link is selected the field value is changed to active if data products for this subset are actively being submitted.

> **pending:** When the submit link is selected this value is set when another subset is currently active. When the dataset starts to be actively submitted the value will change to "active".

> **error:** This value is set if the archive submission script detects an error. The most common cases are when another version of the data product is found to be in the archive or when an data product listed in the pipeline manifest is missing from the staging area.

> *date*: A date is entered here when archive submission completes successfully.

redo: A form is given to edit the processing parameters. Upon submission a cgi script is called to have the subset reprocessed. The pipeline submission state in the queue page is also modified to "resubmit".

> **redo:** set when the redo submission is made.

alt version: Normally reprocessed datasets will appear after an earlier version has been marked as "redo". However, if there are multiple processing versions which have not been submitted (do not have any of the other values), the result of updating the plarchived value for one, for example by submitting to the archive, is to set the remaining versions to this value.
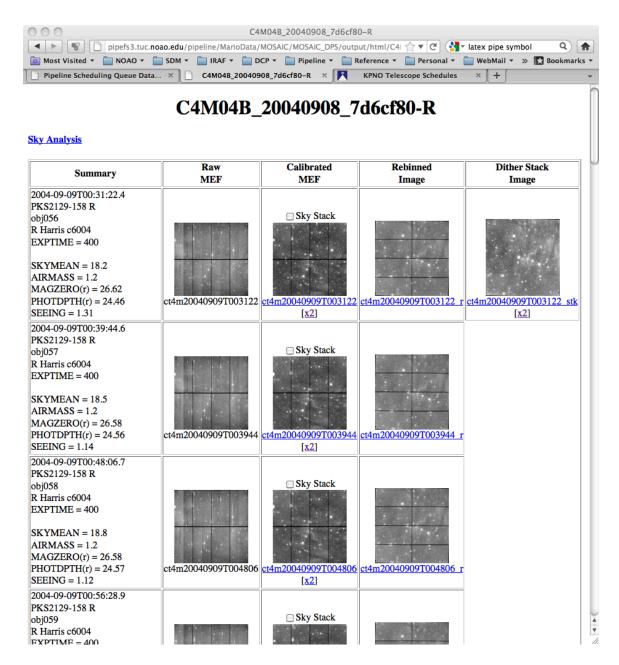
Figure 8: Example Pipeline Generated Operator Review Page

Figure 9: Sample E-mail Sent with the Notification Link.

```
Dear Dr. Walker,

This note is to let you know that pipeline calibrated data from your
proposal 2010B-0230 are available from the NOAO Science Archive.  You
may select and download these data from the NOAO Portal at

http://portal-nvo.noao.edu

There you will also find links for help with selection, staging, and
download options.  Note that you must be registered with the Archive
to retrieve your proprietary data and that for NOAO granted proposals
there is a proprietary period after which the data become public.  If
you have questions, forgotten your login information, or did not
receive an invitation to register please contact vohelp@noao.edu.
Once the proprietary period expires these data will be available to
anyone.

If you have questions about the pipeline data products or how these
data were processed, please contact vohelp@noao.edu.

A document describing the NEWFIRM science pipeline in detail is being
prepared and will be available in the coming weeks.

This message has been automatically generated based on processed data
found in the archive under your proposal number. This message covers
data from this and preceding semesters.  You may receive multiple
messages for data associated with different proposal numbers.  In the
future a message will also be sent as soon as new data becomes
available.

This message is automatically generated.
```