ALMA Pipeline

Crystal Brogan (for the PLWG)

ObsMode 7(+8) Meeting

Santiago, April 24-26, 2018

Pipeline Design Fundamentals - I

1. ALMA Pipeline (PL) operates deterministically based on **data, metadata, scan intents, and ancillary information stored in the SBSummary.xml table**

2. Many PI entries in the OT are not currently available – representative spw is a good example

3. The PL is *LINEAR* - it is a fundamental aspect of the current design that precludes going back and repeating steps. Implementation of loops would require major development

4. The pipeline produces calibrated data (in an image-ready state), and cleaned images of per-spw continuum, aggregate continuum, and spectral cubes (with continuum subtracted) for as many science target spws and fields as deemed currently feasible
    - Feasibility is determined by total product size (a rough proxy for imaging run time) and the maximum size of a cube that a PI is likely to be able to deal with
    - The pipeline/CASA is currently expending SIGNIFICANT effort to improve runtime
    - Making PI able to deal with 50 GB+ cubes is NOT on the near horizon

ALMA Pipeline Flow Diagram

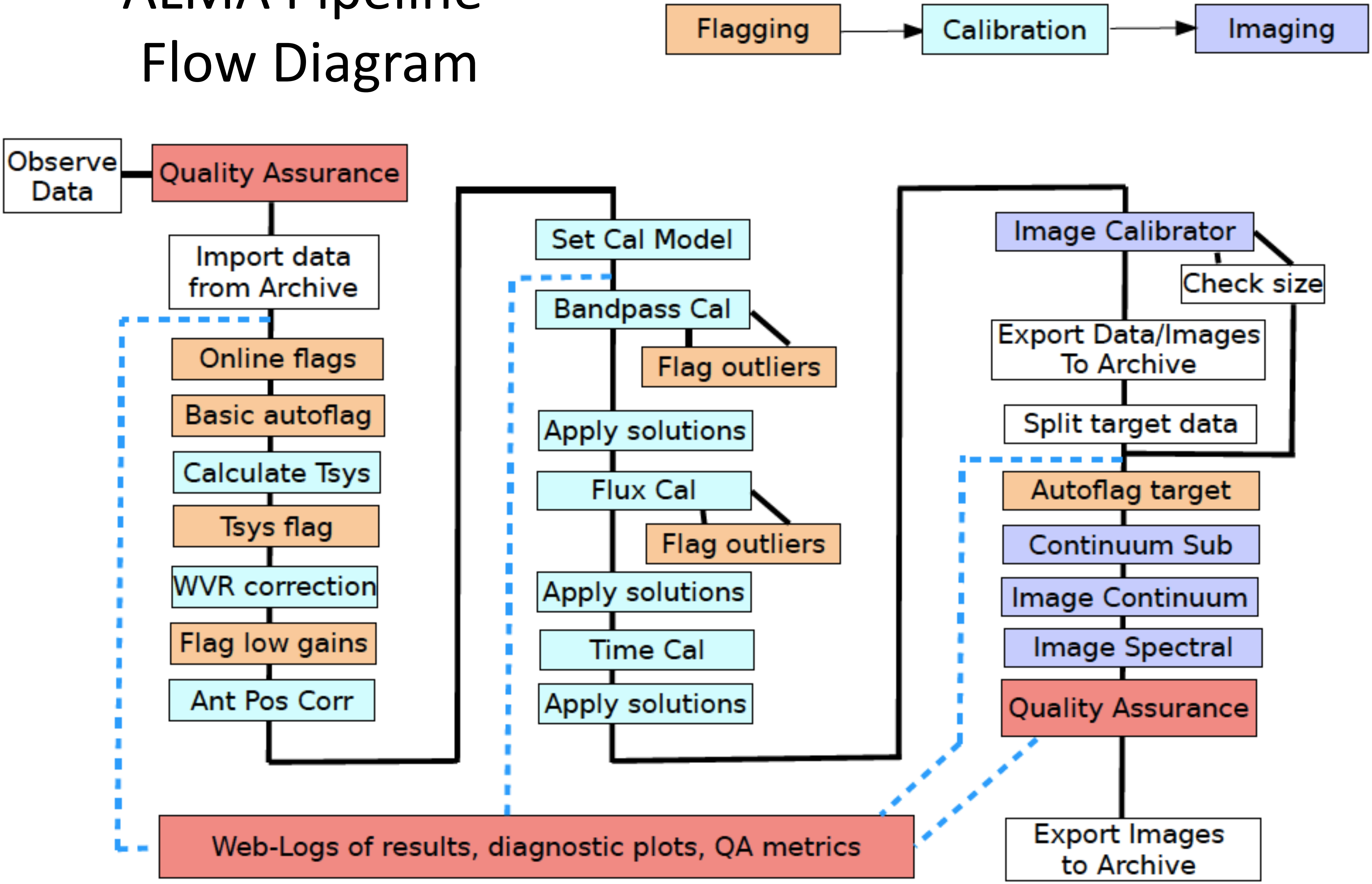

Figure courtesy of Urvashi Rao

Pipeline Design Fundamentals - II

As a high level requirement, PL aims to produce **calibrated data and images on the path to data products that are Science-Ready while being cognizant of data throughput issues:**

1. The imaging PL is designed to be consistent with the OT and hence what the PI believes they will get based on their Control & Performance entries

   Actions which break assumptions and degrade image fidelity:

   - Data collected without the required full range of spatial scales for each MOUS
   - Using significant departures from robust=0.5, or significant uv tapering

2. The imaging PL heuristics are optimized for a well-formed 12m-array array (near-Gaussian uv distribution) – and the more limited uv-coverage of the 7m-array.
   - Automatic clean masking and clean thresholds are fairly sensitive to the properties of the PSF at the MOUS level.
   - We do not have the resources to adapt these heuristics to handle a semi-random collection of antennas.

Major Pipeline development Principles

Pipeline development encompasses two needs:
Improvements and New Modes, each must proceed through similar stages:

1. Research (analysis of issue) and data collection
2. Development and definition of Solution
3. Verification and Quantitative Demonstration
4. Implementation
5. Validation

To date, much of 1, most of 2, and all of 3-5 have been done by PLWG

Scripts developed by EOC and at some level even DRM have little relationship to the pipeline procedure, provide a small fraction of the required information, and/or violate aspects of pipeline design such that we basically have to start from scratch

Ex: Bandwidth switching (BWSW).
This has been on the request list for 2 years, but to date, a significant fraction of BWSW data is not observed using the best practices observing sequence: we have had difficulty even finding a representative sample of properly observed use cases

Major Pipeline development: Improvements

1. Research and data collection phase

Result = clear description of the issue.
Ex: “the pipeline does not perform adequately for low-signal-to noise cases” is not sufficient

a. Scope of the issue: What **fraction** of PI data is affected, averaged over a cycle
b. Scope of the issue: What **type** of data are affected? Is it only HF data, or only low-signal-to-noise data? Multiple examples of both the suspected data types, and other data types spanning parameter space need to be analyzed.

- Calibrated amplitude flagging took >50 datasets to define the problem
- Optimizing the Cycle 6 findContinuum took > 600 cubes spanning 3 Cycles of data and spanning all configurations

If data have to be re-reduced because they’ve been deleted, and analyzed by PLWG, need another 2+ months of a >1/3 time expert (to date, has been PLWG)

If DRs would analyze data in while it’s on disk, and *synthesize* the information, that timeline could be reduced. Decreasing overall human effort to regular processing could allow this more of the time.

Major Pipeline development: New Modes

1. Research and data collection phase

Clear description of what EXACT observing strategy the new mode is meant to encompass, including sequence of calibrators, their intents

Datasets for new capabilities (or heuristics) must be collected, analyzed, and synthesized, typically over ~6 months of PI data reduction

Ex 1: “Make High Frequency run in the pipeline”
Reality: need more concrete examples of why and when this doesn’t work dealt with a systematic approach to evaluation

Ex 2: “Polarization calibration and imaging”
There is a well developed DRM script for polarization calibration. However:

- There is nothing for circular even for manual at this moment
- There is much discussion about whether current calibration produces optimal results and now there is a high level request by the project to carry out CASA research and development on a whole new procedure.
- **Any mode requiring CASA development: add +1 Cycle**

Major Pipeline development

2. Definition of solution (page 1 of 2)

– 4 months of research (has been PLWG to date)

2a. What are the precise conditions under which a new heuristic should be applied, using ONLY the data itself and metadata present in the ASDM. No external information can be used to make decisions.

2b. What is the state of the data just before the new heuristic is to be applied – what state of calibration, flagging, etc?

2c. What data and metadata parameters need to be present or calculated to decide whether to apply the new heuristic?

2d. What is the precise quantitative threshold for action?

2e. What is the basis for calculating a score (or scores) to assess success?
What are the quartile breakpoints? What should low score warning messages say?

Major Pipeline development

2. Definition of solution (page 2 of 2)

2f. A script which automatically (no human intervention) processes data using the new heuristic. Script should follow general pipeline processing path – no extra splitting of the data, or loops, or a detailed description of why this isn't possible

2g. The script should either run completely through science target imaging, or be demonstrated that the script can be inserted into a regular pipeline recipe, and that all required pipeline tasks which follow the new heuristic still run to completion.

> Both this and 2f above are more easily accomplished if the new script uses exclusively pipeline tasks before the new heuristic, and then follows with more pipeline tasks after the new heuristic.

2h. What parameters will be required to be exposed for testing/commissioning?

2i. What parameters will be required to be exposed to the user in the PPR (casa_pipescript.py) and pipeline task interfaces?

Major Pipeline development

3. Verification and Quantitative demonstration of effectiveness:

3a. The prototype script must be run on at least 10 datasets, and results analyzed. If this analysis results in ANY changes to the observing best practices, the testing must start again with data taken using the new observing sequence.

3b. If applicable, the results must be shown to be quantitatively better than the existing pipeline heuristics

3c. It must be shown that the new heuristics have no harmful effects

> Data spanning parameter space must be tested with the new heuristic, and the results analyzed – depending on the parameter space in question, this generally requires analysis of another >10 datasets.

Major Pipeline development

4. Implementation in the real pipeline by pipeline developers – could be 2-4 person-months, and all PL developers other than the lead are part-time, so scheduling can be challenging.

Major requirements are due by the end of November

If CASA development is required, then the requirements are due SOONER so that Development can be sequenced: CASA first, then pipeline

5. PLWG testing in the production code

repetition of both the effectiveness and “no-harm” validation, on 10s-100s of datasets – typically concentrated during the summer. It can take weeks of computer time to run the data, let alone the analysis time.

Major Pipeline development: FTE estimates

One major project takes ~ 1 person-year PLWG plus 3-6 person-month developer.

There's always 1-2 FTE of minor research, development, and testing in addition.

Augmented resources were ***temporarily*** applied in C4,5 in particular for imaging.

C4-C5 PLWG effort >4FTE:		C7 PLWG effort ~2FTE:
Hibbard 0.8		Hibbard 0.25 (QA)
Indebetouw 0.3		Indebetouw 0.5
Brogan 0.8		Brogan 0.3
Hunter 0.5		Hunter 0.2
Mason 0.2		Kepley 0.2
Kepley 0.4		
Videla 0.2		Videla 0.2
Villard 0.2		
Humphreys 0.2		Humphreys 0.05
Egusa 0.3	Single Dish	Egusa 0.3
Miura 0.2		Tafoya 0.2

C4-5 Dev. effort ~3FTE		C7 Dev. effort ~2.5FTE
Davis 1.0		Sugimoto 1.0
Muders 0.3		Muders 0.3
Williams 0.5		Williams 0.5
Geers 0.5		Geers 0.5
Sugimoto 0.5	Single Dish	
Nakazoto 0.3		Nakazoto 0.3
Kosugi 0.1		Kosugi 0.1
Yoshino 0.2		Yoshino 0.2