ThoughtWorks®

ARTIP

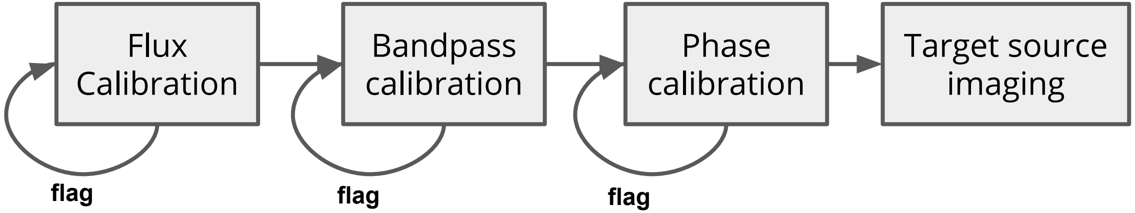
Automated Radio Telescope Imaging Pipeline

Presented by-
Dolly, Ravi
ThoughtWorks, Pune

AGENDA

1. Problem or Current Workflow of data reduction
2. What is ARTIP?
3. Key Features of ARTIP
4. Pipeline Architecture and Design
5. Performance
6. Hands-on

CURRENT DATA REDUCTION WORKFLOW



For an expert scientist, to manually reduce a dataset of 10 GB,

time taken is around 3-4 hours

CHALLENGES WITH NEW TELESCOPES

MEERKAT ABSORPTION LINE SURVEY

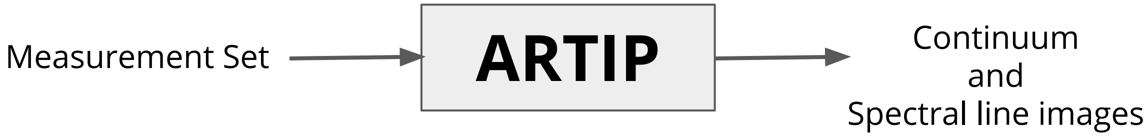
2 = **4TB** x **RAW DATA IMAGES**

Hours of Observation

ARTIP: PIPELINE OVERVIEW

A pipeline is a series of stages ran in a sequence, where each stage produces some artifacts which are then consumed by the downstream stages.

- ARTIP stands for **Automated Radio Telescope Imaging Pipeline**
- ARTIP is a fully automated end to end pipeline



Speed

Objectivity

Repeatability

PIPELINE OVERVIEW: DATASETS

Pipeline works on datasets containing :

- Flux calibrator(s)
- Phase calibrator(s)
- Target source(s)
- Spectral windows(s)

Time taken: ~20mins for 10 GB

Server Specs:

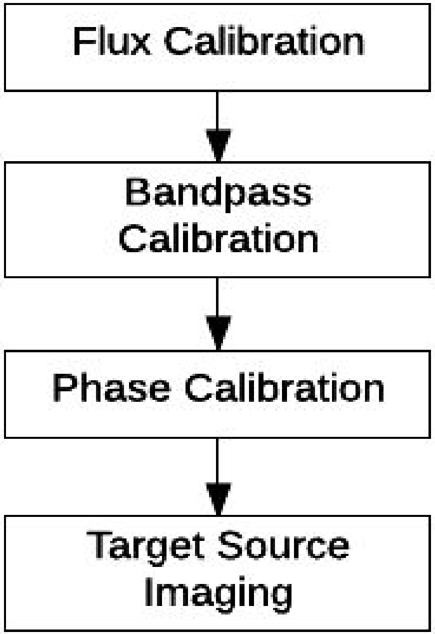
RAM - 256 GB

Cores - 40

Tested against GMRT (12) and VLA (13) datasets

KEY FEATURES OF ARTIP

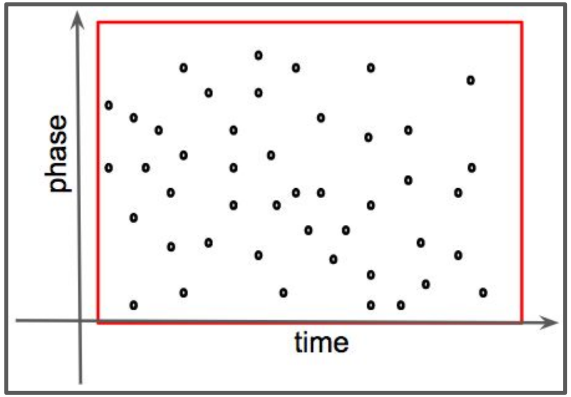
STAGE DRIVEN ARCHITECTURE



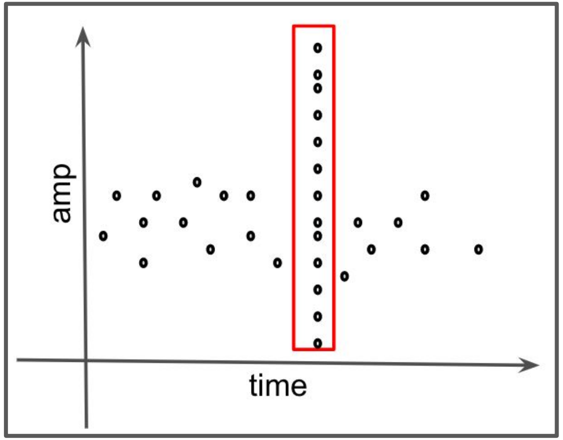
1. Outputs of each stage are persisted and used by downstream stages.
2. Quick feedback for the user (verification of output and quality check on each stage)
3. If the last stage fails, one doesn't need to run the entire pipeline all over again.
4. Stages further have substages which can also be toggled on or off.
5. Modularization and extensibility of code

PATTERNS OF BAD DATA (IN TIME) CAUGHT BY THE PIPELINE

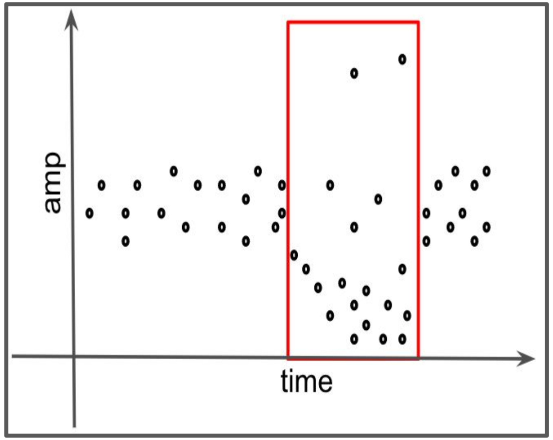
Bad Antenna



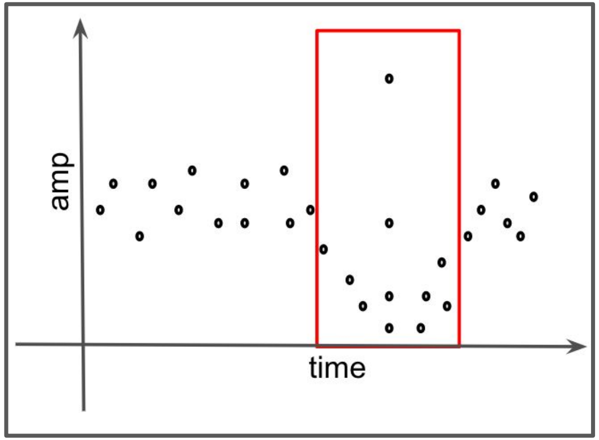
Bad Time



Bad Antenna Time



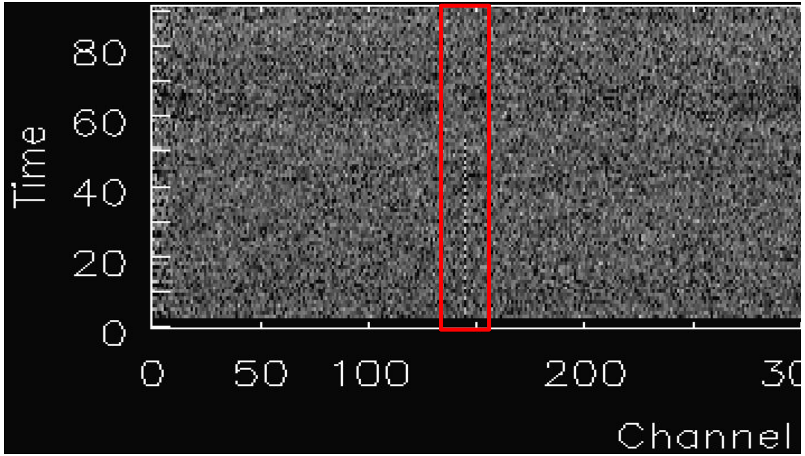
Bad Baseline Time



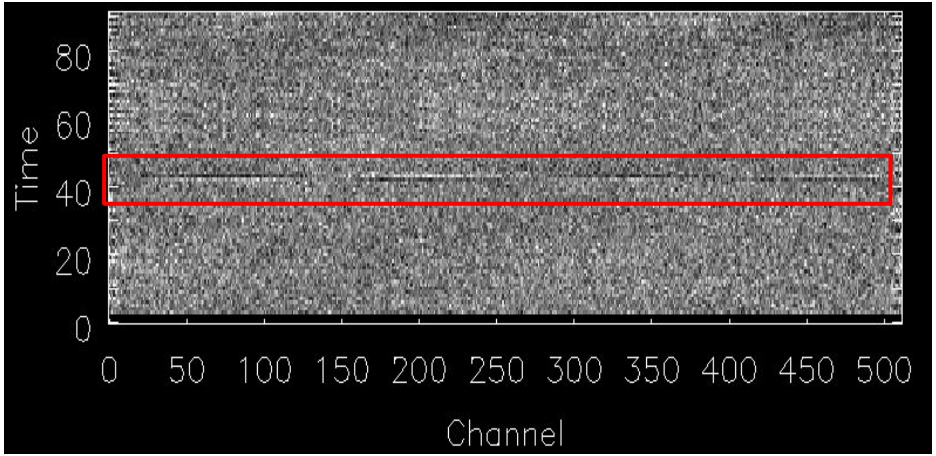
PATTERNS OF BAD DATA (IN FREQUENCY) CAUGHT BY THE PIPELINE

Tfcrop
Rflag

Bad Channel

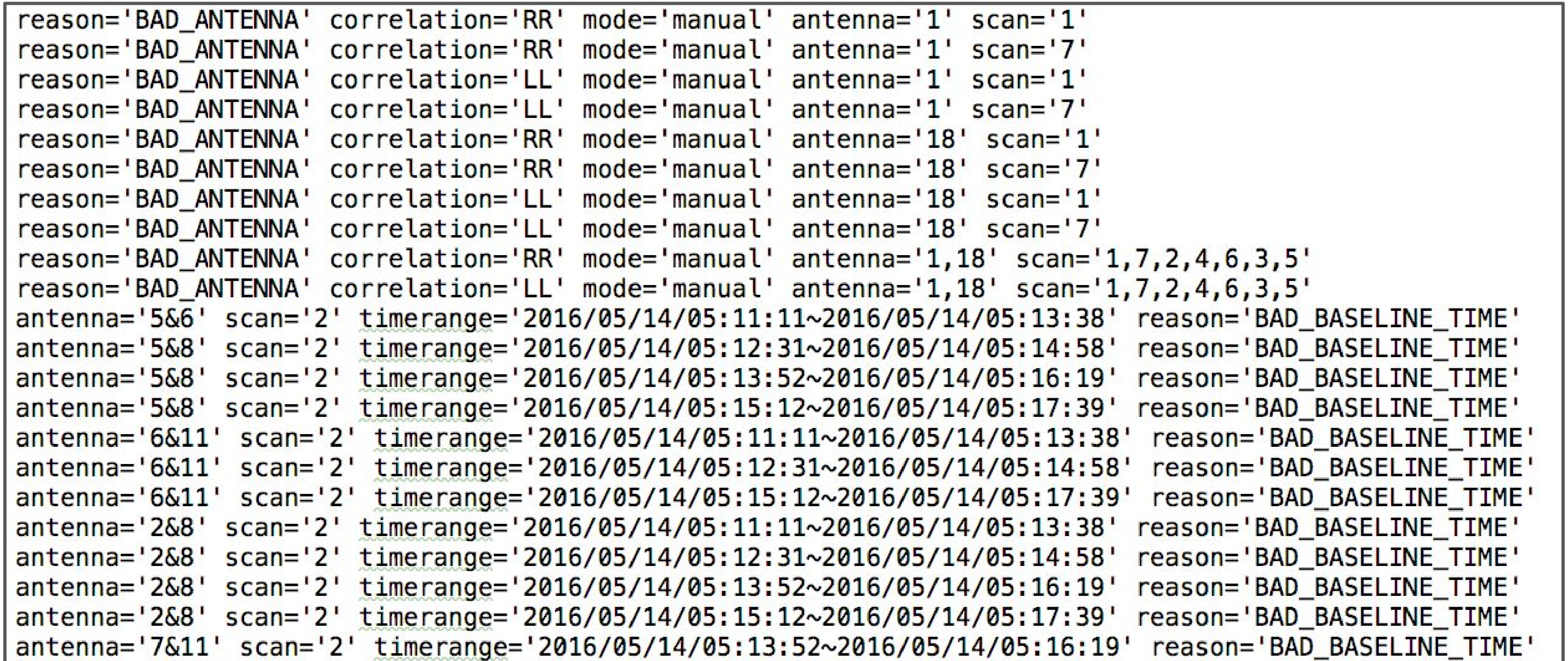


Bad Time Over Channels



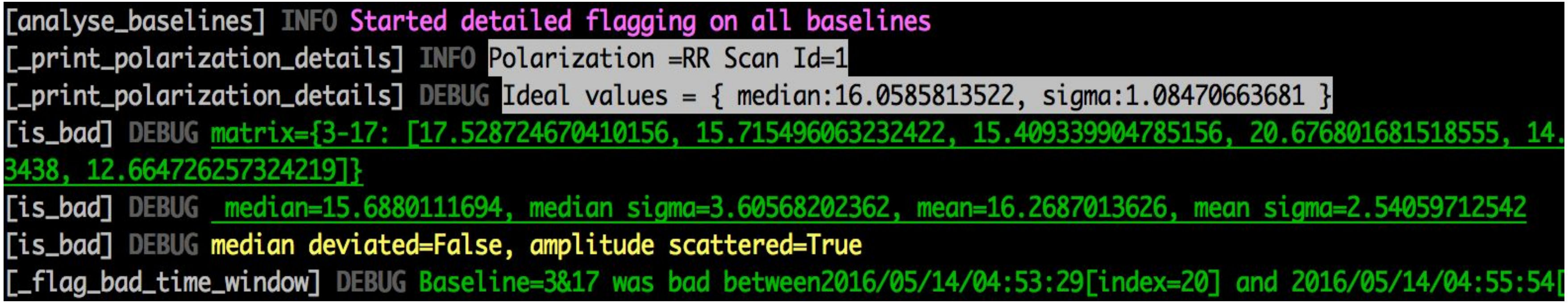
MAINTAINING FLAG REASONS

Flags.txt

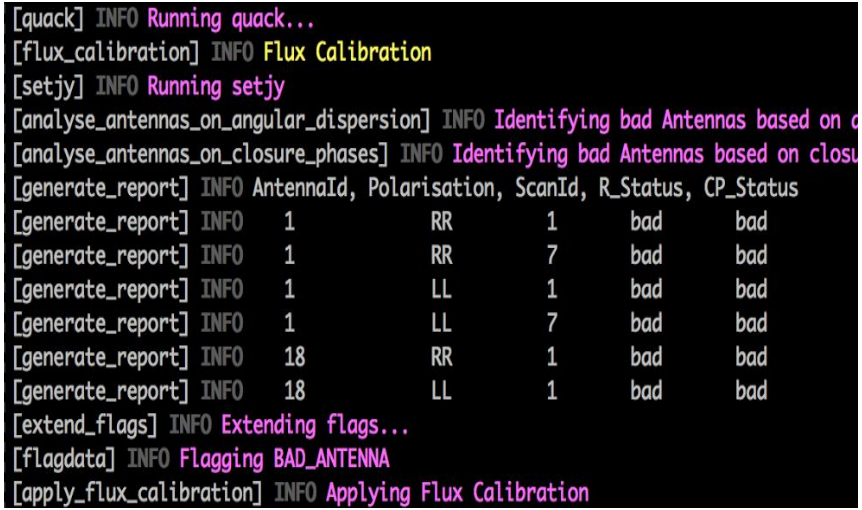


```
reason='BAD_ANTENNA' correlation='RR' mode='manual' antenna='1' scan='1'
reason='BAD_ANTENNA' correlation='RR' mode='manual' antenna='1' scan='7'
reason='BAD_ANTENNA' correlation='LL' mode='manual' antenna='1' scan='1'
reason='BAD_ANTENNA' correlation='LL' mode='manual' antenna='1' scan='7'
reason='BAD_ANTENNA' correlation='RR' mode='manual' antenna='18' scan='1'
reason='BAD_ANTENNA' correlation='RR' mode='manual' antenna='18' scan='7'
reason='BAD_ANTENNA' correlation='LL' mode='manual' antenna='18' scan='1'
reason='BAD_ANTENNA' correlation='LL' mode='manual' antenna='18' scan='7'
reason='BAD_ANTENNA' correlation='RR' mode='manual' antenna='1,18' scan='1,7,2,4,6,3,5'
reason='BAD_ANTENNA' correlation='LL' mode='manual' antenna='1,18' scan='1,7,2,4,6,3,5'
antenna='5&6' scan='2' timerange='2016/05/14/05:11:11~2016/05/14/05:13:38' reason='BAD_BASELINE_TIME'
antenna='5&8' scan='2' timerange='2016/05/14/05:12:31~2016/05/14/05:14:58' reason='BAD_BASELINE_TIME'
antenna='5&8' scan='2' timerange='2016/05/14/05:13:52~2016/05/14/05:16:19' reason='BAD_BASELINE_TIME'
antenna='5&8' scan='2' timerange='2016/05/14/05:15:12~2016/05/14/05:17:39' reason='BAD_BASELINE_TIME'
antenna='6&11' scan='2' timerange='2016/05/14/05:11:11~2016/05/14/05:13:38' reason='BAD_BASELINE_TIME'
antenna='6&11' scan='2' timerange='2016/05/14/05:12:31~2016/05/14/05:14:58' reason='BAD_BASELINE_TIME'
antenna='6&11' scan='2' timerange='2016/05/14/05:15:12~2016/05/14/05:17:39' reason='BAD_BASELINE_TIME'
antenna='2&8' scan='2' timerange='2016/05/14/05:11:11~2016/05/14/05:13:38' reason='BAD_BASELINE_TIME'
antenna='2&8' scan='2' timerange='2016/05/14/05:12:31~2016/05/14/05:14:58' reason='BAD_BASELINE_TIME'
antenna='2&8' scan='2' timerange='2016/05/14/05:13:52~2016/05/14/05:16:19' reason='BAD_BASELINE_TIME'
antenna='2&8' scan='2' timerange='2016/05/14/05:15:12~2016/05/14/05:17:39' reason='BAD_BASELINE_TIME'
antenna='7&11' scan='2' timerange='2016/05/14/05:13:52~2016/05/14/05:16:19' reason='BAD_BASELINE_TIME'
```

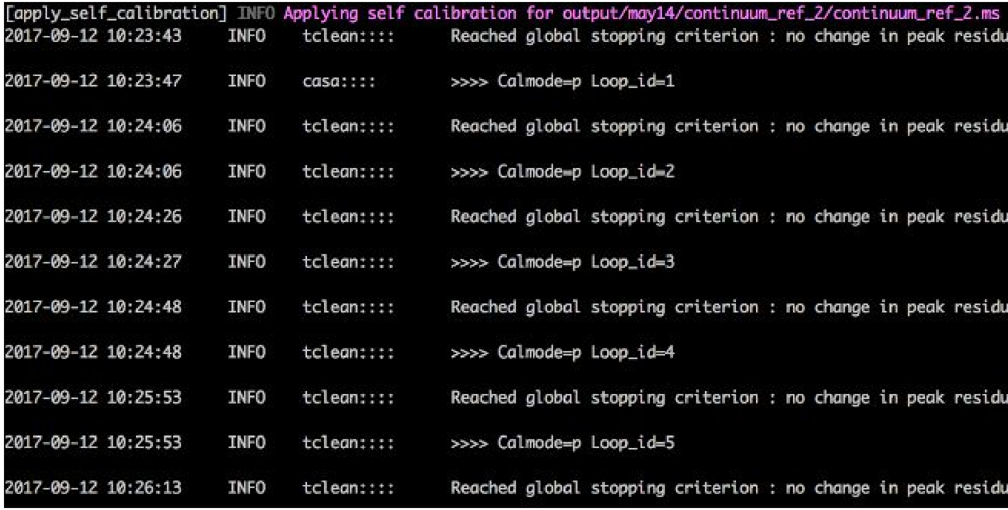
LOGGING



```
[analyse_baselines] INFO Started detailed flagging on all baselines
[_print_polarization_details] INFO Polarization =RR Scan Id=1
[_print_polarization_details] DEBUG Ideal values = { median:16.0585813522, sigma:1.08470663681 }
[is_bad] DEBUG matrix={3-17: [17.528724670410156, 15.715496063232422, 15.409339904785156, 20.676801681518555, 14.
3438, 12.664726257324219]}
[is_bad] DEBUG  median=15.6880111694, median sigma=3.60568202362, mean=16.2687013626, mean sigma=2.54059712542
[is_bad] DEBUG median deviated=False, amplitude scattered=True
[_flag_bad_time_window] DEBUG Baseline=3&17 was bad between2016/05/14/04:53:29[index=20] and 2016/05/14/04:55:54[
```



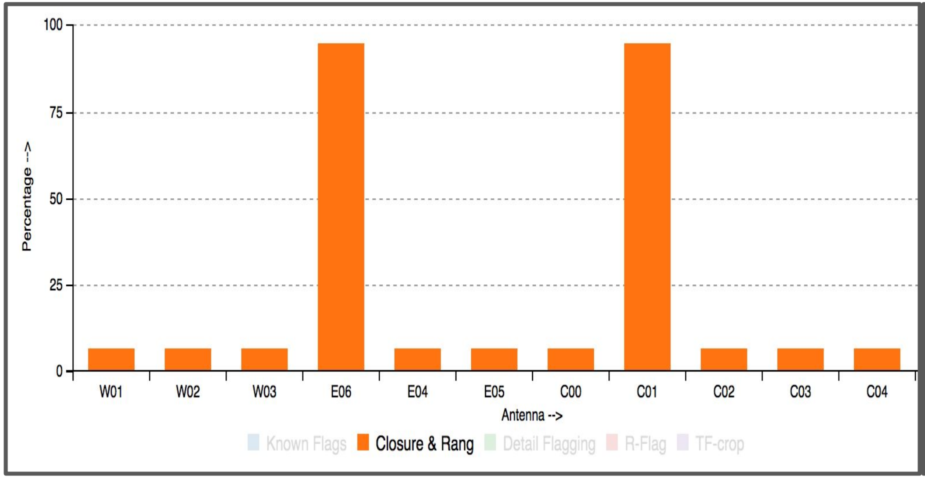
```
[quack] INFO Running quack...
[flux_calibration] INFO Flux Calibration
[setjy] INFO Running setjy
[analyse_antennas_on_angular_dispersion] INFO Identifying bad Antennas based on a
[analyse_antennas_on_closure_phases] INFO Identifying bad Antennas based on closu
[generate_report] INFO AntennaId, Polarisation, ScanId, R_Status, CP_Status
[generate_report] INFO   1            RR         1       bad       bad
[generate_report] INFO   1            RR         7       bad       bad
[generate_report] INFO   1            LL         1       bad       bad
[generate_report] INFO   1            LL         7       bad       bad
[generate_report] INFO   18           RR         1       bad       bad
[generate_report] INFO   18           LL         1       bad       bad
[extend_flags] INFO Extending flags...
[flagdata] INFO Flagging BAD_ANTENNA
[apply_flux_calibration] INFO Applying Flux Calibration
```



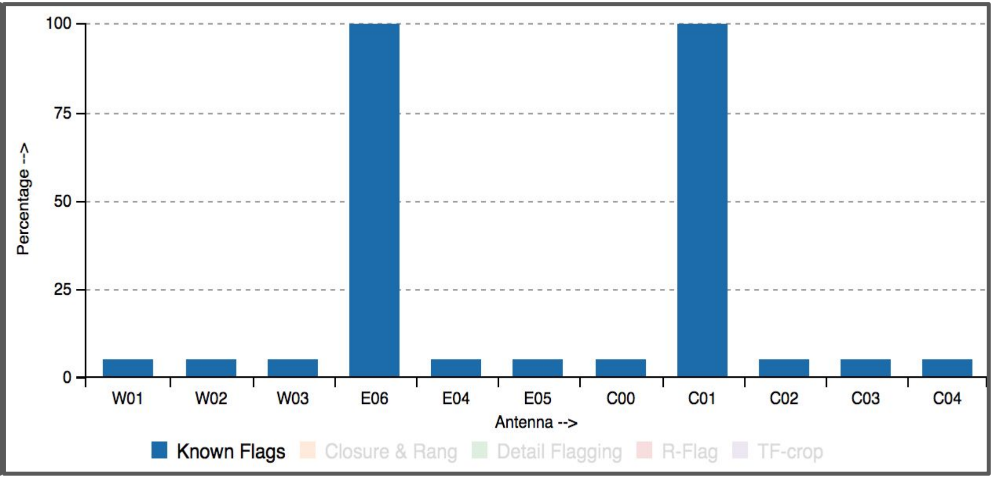
```
[apply_self_calibration] INFO Applying self calibration for output/may14/continuum_ref_2/continuum_ref_2.ms
2017-09-12 10:23:43     INFO    tclean::::      Reached global stopping criterion : no change in peak residu

2017-09-12 10:23:47     INFO    casa::::        >>>> Calmode=p Loop_id=1

2017-09-12 10:24:06     INFO    tclean::::      Reached global stopping criterion : no change in peak residu

2017-09-12 10:24:06     INFO    tclean::::      >>>> Calmode=p Loop_id=2

2017-09-12 10:24:26     INFO    tclean::::      Reached global stopping criterion : no change in peak residu

2017-09-12 10:24:27     INFO    tclean::::      >>>> Calmode=p Loop_id=3

2017-09-12 10:24:48     INFO    tclean::::      Reached global stopping criterion : no change in peak residu

2017-09-12 10:24:48     INFO    tclean::::      >>>> Calmode=p Loop_id=4

2017-09-12 10:25:53     INFO    tclean::::      Reached global stopping criterion : no change in peak residu

2017-09-12 10:25:53     INFO    tclean::::      >>>> Calmode=p Loop_id=5

2017-09-12 10:26:13     INFO    tclean::::      Reached global stopping criterion : no change in peak residu
```

FLAGGING GRAPHS

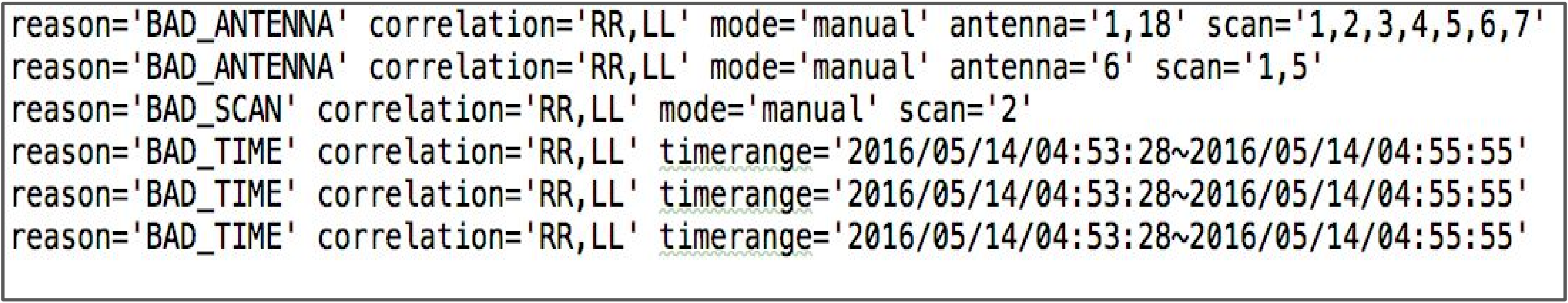
Calibrator flags



Target source extension



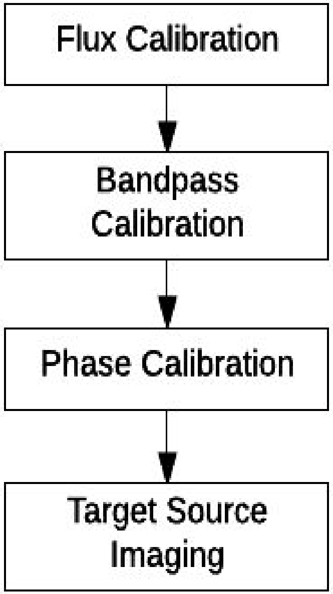
OBSERVATION FLAGS



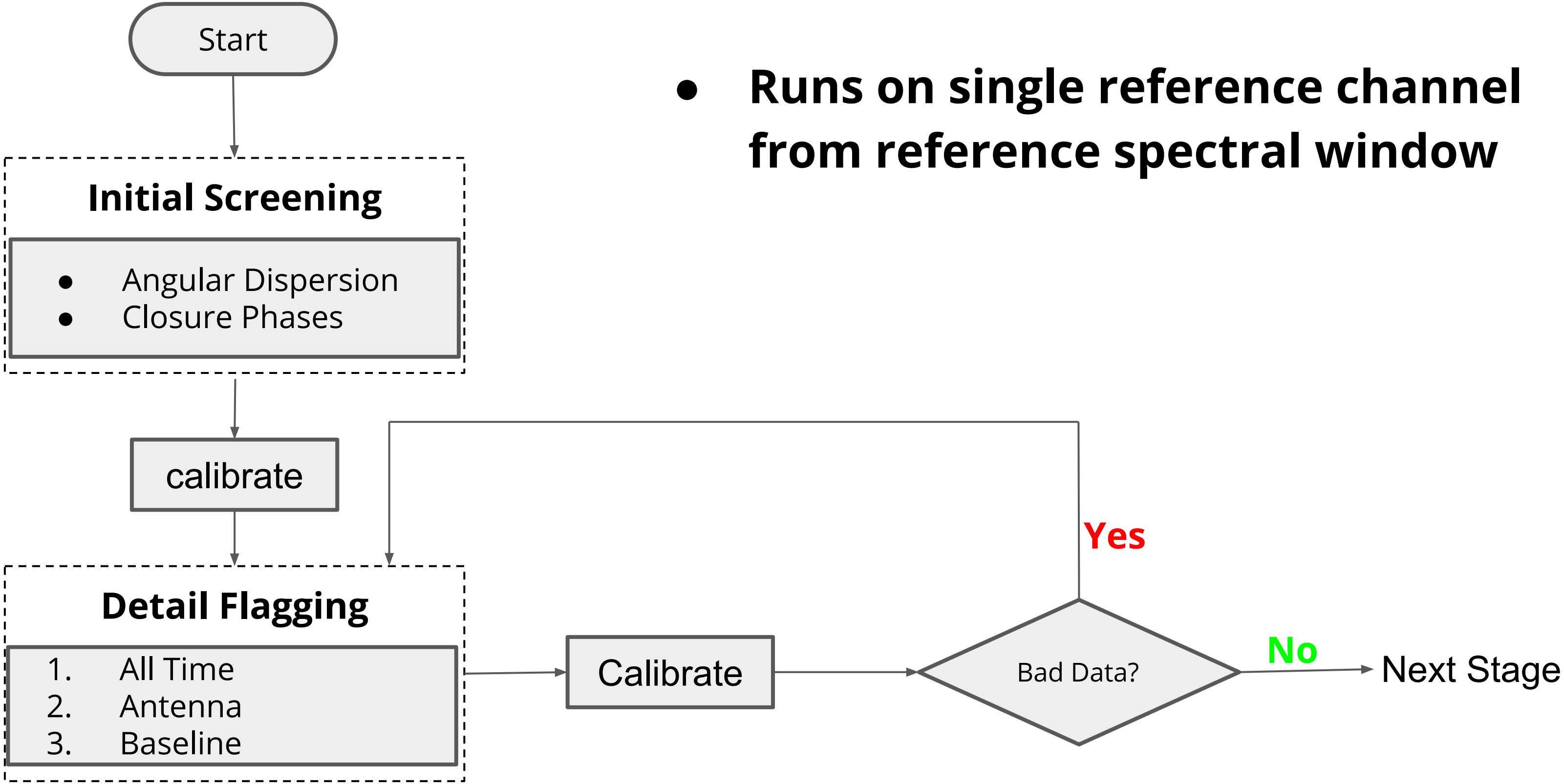
```
reason='BAD_ANTENNA' correlation='RR,LL' mode='manual' antenna='1,18' scan='1,2,3,4,5,6,7'
reason='BAD_ANTENNA' correlation='RR,LL' mode='manual' antenna='6' scan='1,5'
reason='BAD_SCAN' correlation='RR,LL' mode='manual' scan='2'
reason='BAD_TIME' correlation='RR,LL' timerange='2016/05/14/04:53:28~2016/05/14/04:55:55'
reason='BAD_TIME' correlation='RR,LL' timerange='2016/05/14/04:53:28~2016/05/14/04:55:55'
reason='BAD_TIME' correlation='RR,LL' timerange='2016/05/14/04:53:28~2016/05/14/04:55:55'
```

PIPELINE ARCHITECTURE

PIPELINE ARCHITECTURE: STAGES



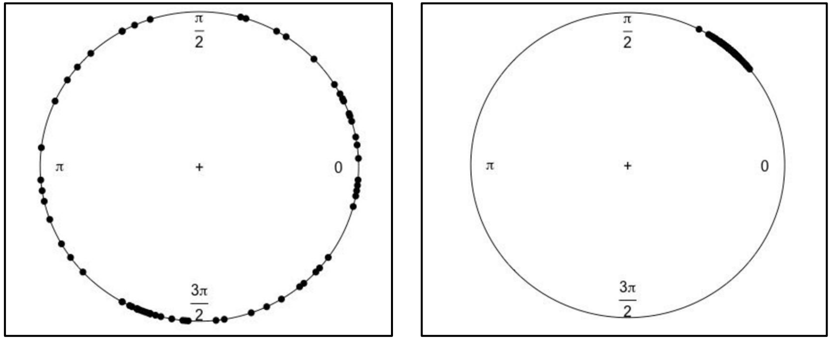
PIPELINE ARCHITECTURE: FLUX CALIBRATION



- **Runs on single reference channel from reference spectral window**

INITIAL SCREENING: PHASE DISPERSION

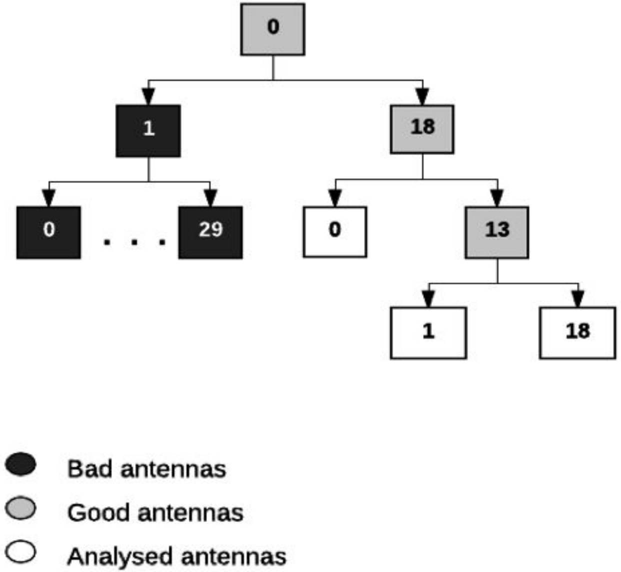
Angular Dispersion



R ≈ 0

R ≈ 1

Depth First Tree Traversal

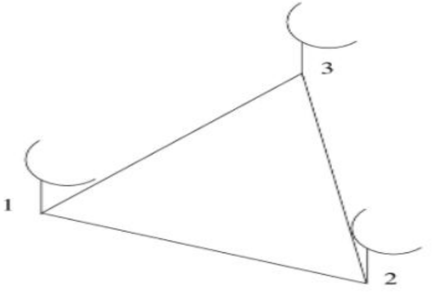


1. Percentage of good baselines for an antenna
2. Minimum percentage of doubt

INITIAL SCREENING: CLOSURE PHASES

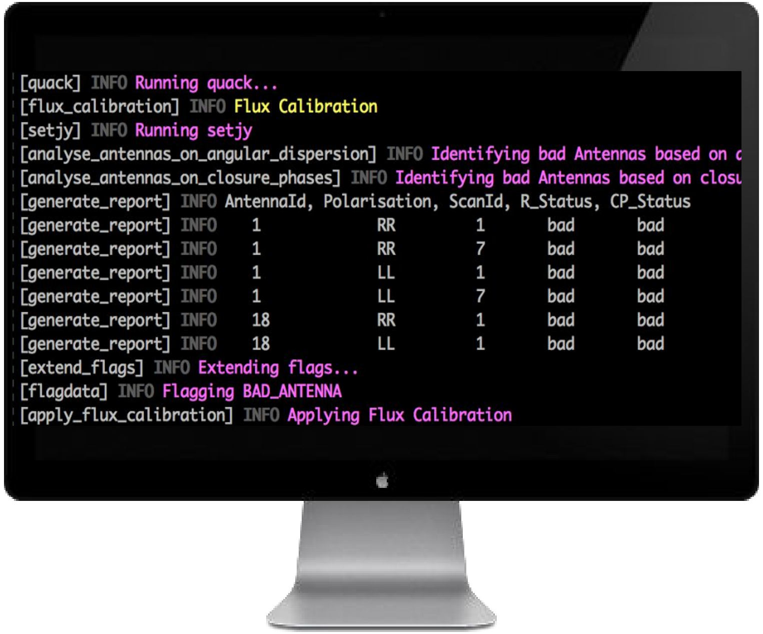
1. Works on triplets
2. Works only on compact sources

$$O = \Phi_{12} + \Phi_{23} - \Phi_{13}$$



In a triplet, the sum of phase differences between 2 baselines should be equal to the phase difference of the third baseline

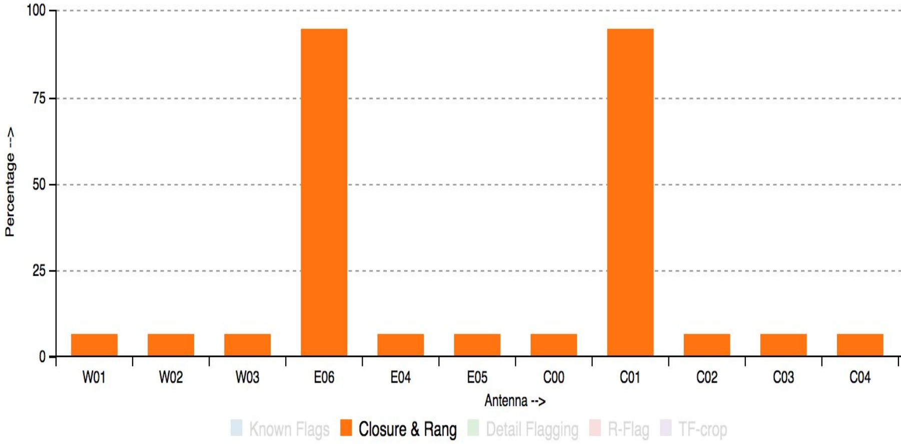
FLAGGING RESULTS



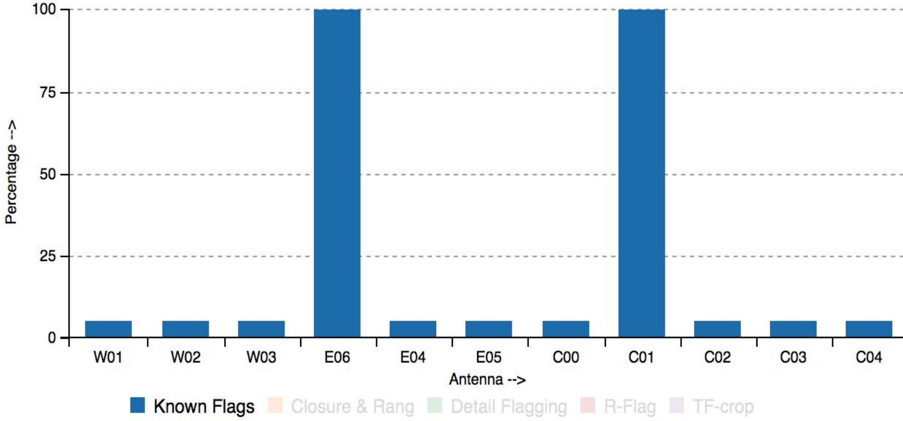
```
[quack] INFO Running quack...
[flux_calibration] INFO Flux Calibration
[setjy] INFO Running setjy
[analyse_antennas_on_angular_dispersion] INFO Identifying bad Antennas based on c
[analyse_antennas_on_closure_phases] INFO Identifying bad Antennas based on closu
[generate_report] INFO AntennaId, Polarisation, ScanId, R_Status, CP_Status
[generate_report] INFO   1           RR          1       bad        bad
[generate_report] INFO   1           RR          7       bad        bad
[generate_report] INFO   1           LL          1       bad        bad
[generate_report] INFO   1           LL          7       bad        bad
[generate_report] INFO   18          RR          1       bad        bad
[generate_report] INFO   18          LL          1       bad        bad
[extend_flags] INFO Extending flags...
[flagdata] INFO Flagging BAD_ANTENNA
[apply_flux_calibration] INFO Applying Flux Calibration
```

Antennas that are identified as bad in both the algorithms are flagged!

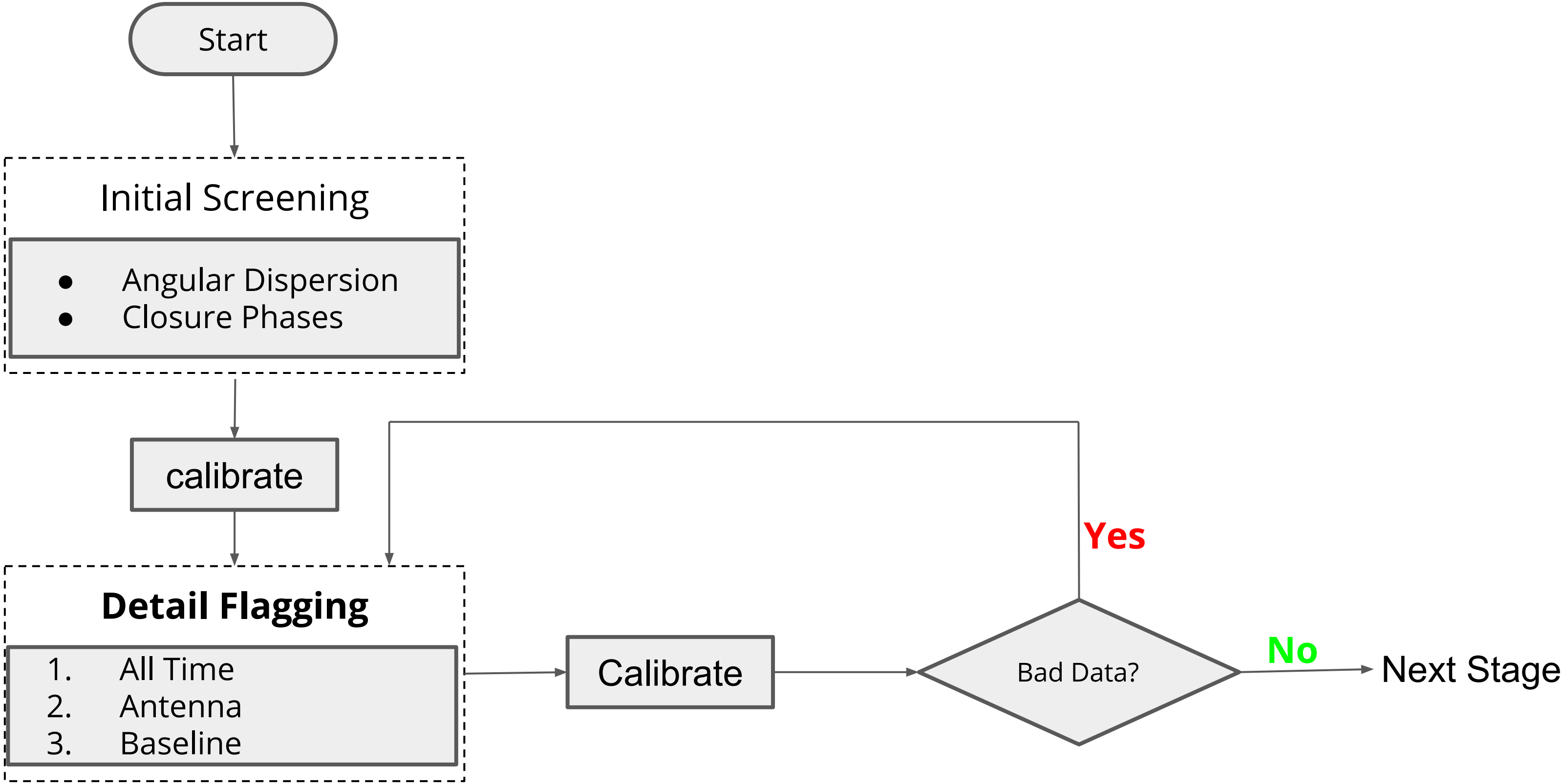
Calibrator flags



Target source extension



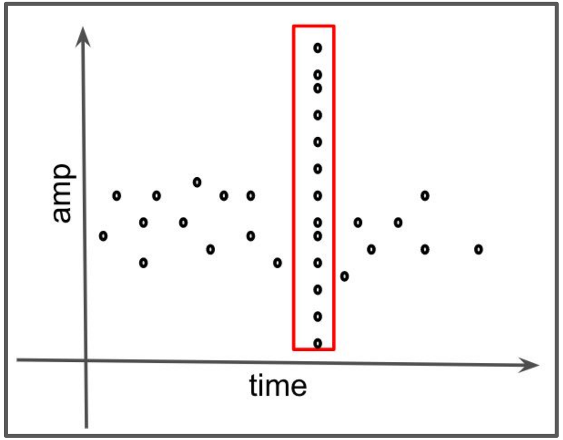
PIPELINE ARCHITECTURE: FLUX CALIBRATION



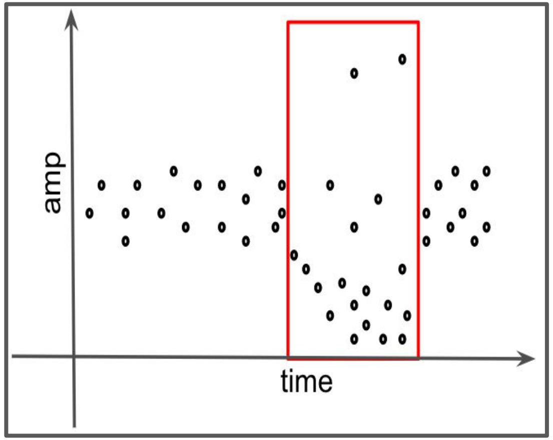
PIPELINE ARCHITECTURE: DETAIL FLAGGING

1. Works on amplitudes
2. Flags and calibrates iteratively till all the data looks good
3. Median and Median Absolute Deviation (MAD) statistics
4. Window size can be configured depending on the data quality
5. Windows with insufficient data points for statistics are not processed

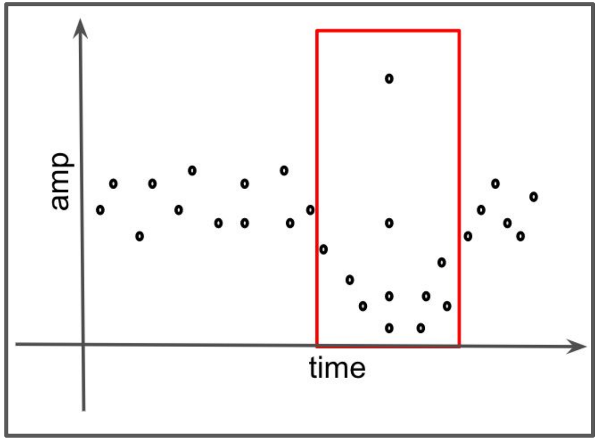
All Antennas



Each Antenna

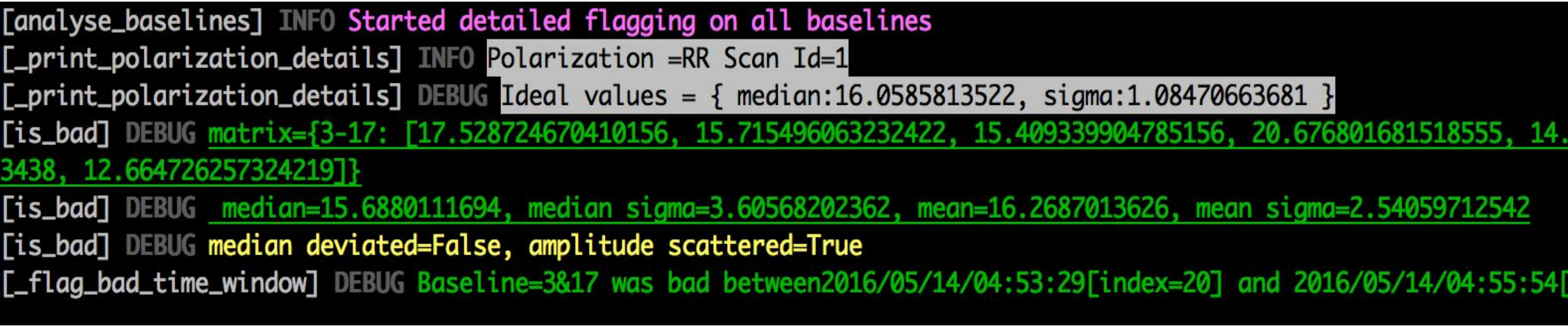


Each Baseline



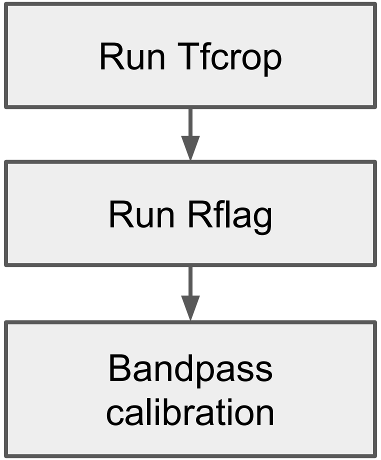
LOGGING FOR DEBUGGING

1. Window time
2. Mean and median and the deviations
3. Flagged due to deviated median or scatter



```
[analyse_baselines] INFO Started detailed flagging on all baselines
[_print_polarization_details] INFO Polarization =RR Scan Id=1
[_print_polarization_details] DEBUG Ideal values = { median:16.0585813522, sigma:1.08470663681 }
[is_bad] DEBUG matrix={3-17: [17.528724670410156, 15.715496063232422, 15.409339904785156, 20.676801681518555, 14.
3438, 12.664726257324219]}
[is_bad] DEBUG  median=15.6880111694, median sigma=3.60568202362, mean=16.2687013626, mean sigma=2.54059712542
[is_bad] DEBUG median deviated=False, amplitude scattered=True
[_flag_bad_time_window] DEBUG Baseline=3&17 was bad between2016/05/14/04:53:29[index=20] and 2016/05/14/04:55:54[
```

PIPELINE ARCHITECTURE: BANDPASS CALIBRATION



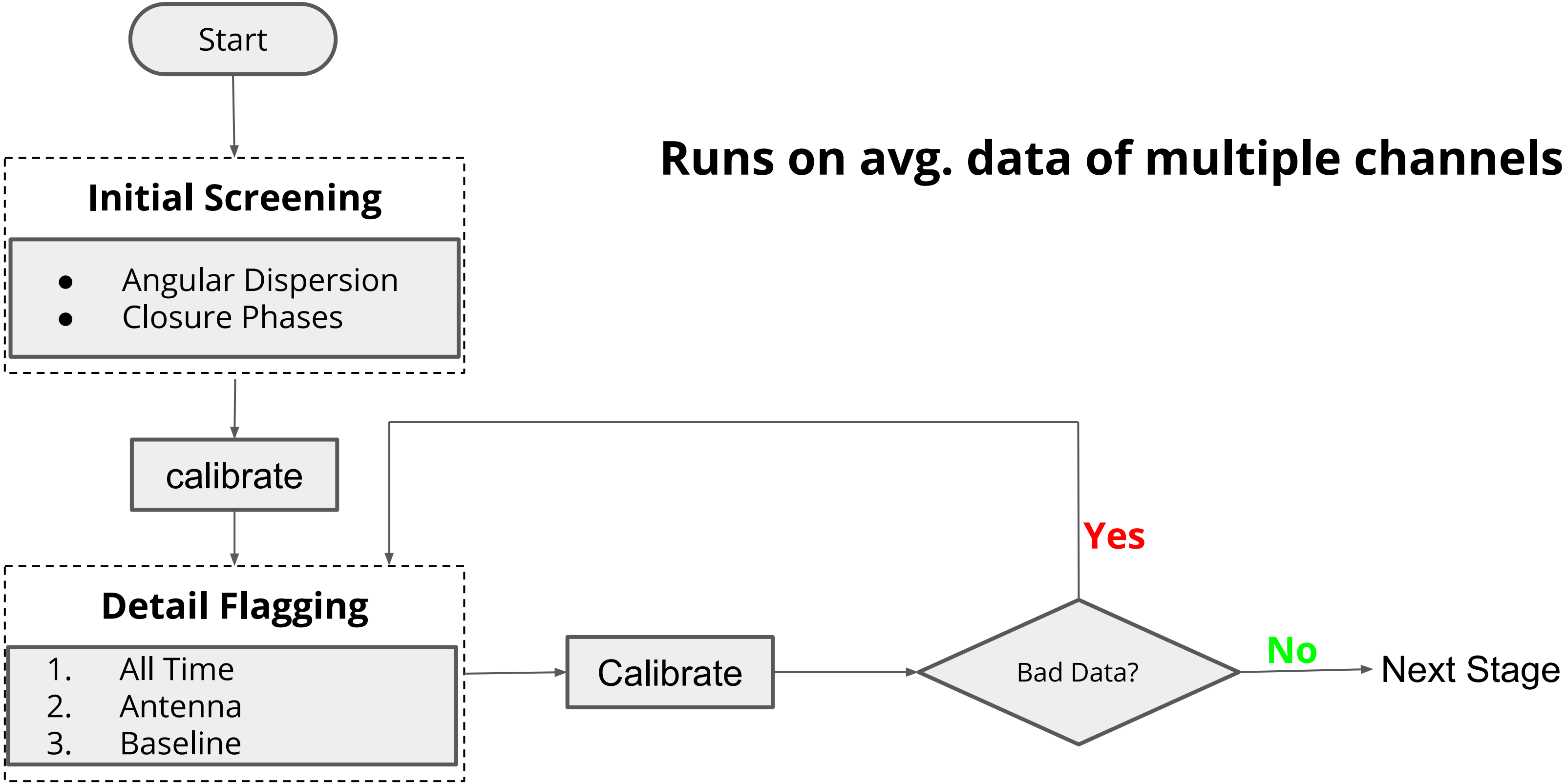
Run Tfcrop

Run Rflag

Bandpass calibration

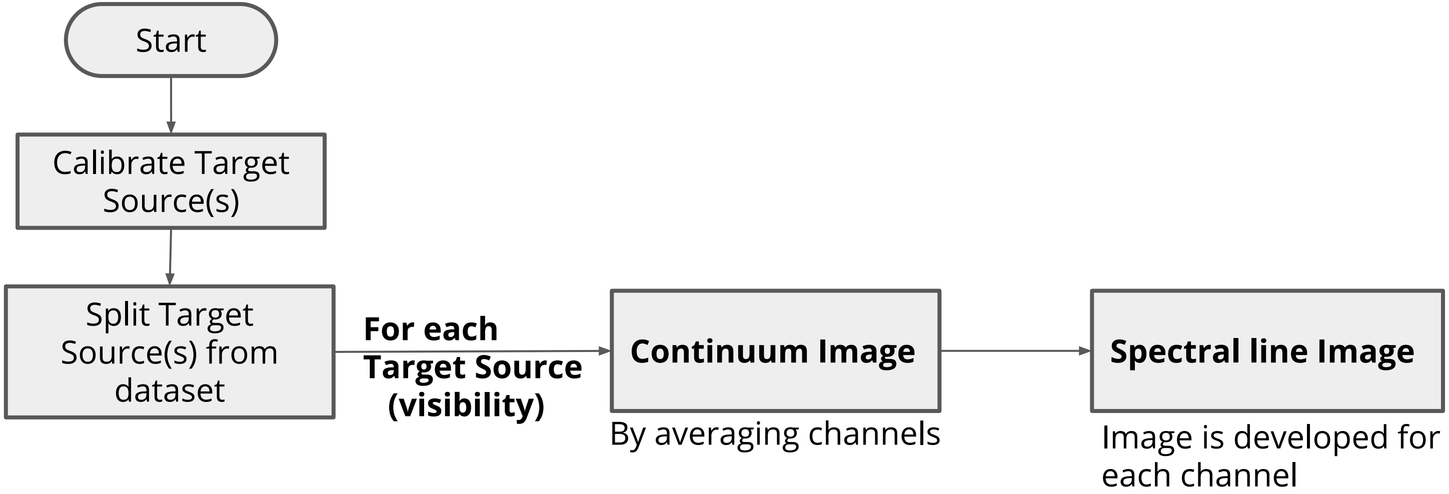
Different thresholds can be specified for each spectral window.

PIPELINE ARCHITECTURE: PHASE CALIBRATION



Runs on avg. data of multiple channels

PIPELINE ARCHITECTURE: IMAGING



Start

Calibrate Target Source(s)

Split Target Source(s) from dataset

For each Target Source (visibility)

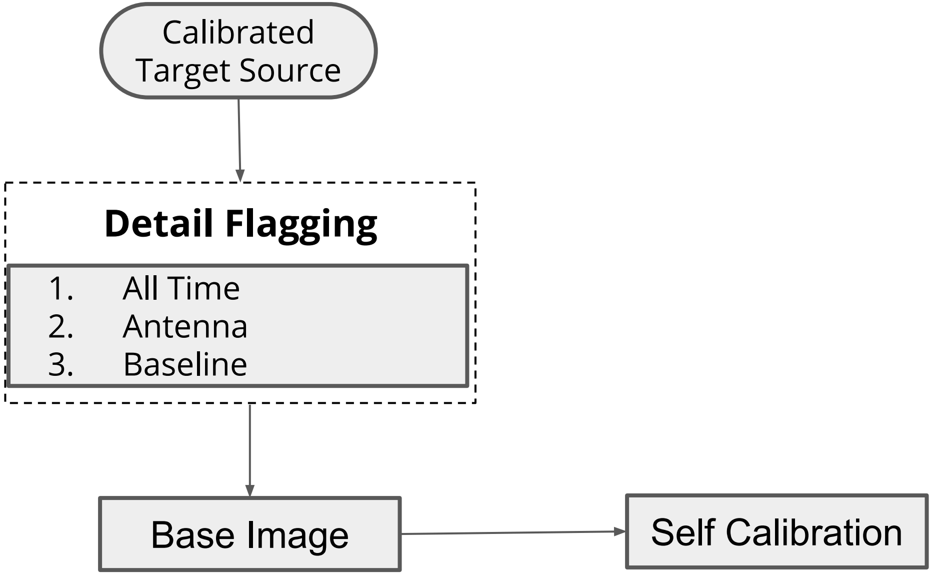
Continuum Image

By averaging channels

Spectral line Image

Image is developed for each channel

CONTINUUM IMAGING



Calibrated Target Source

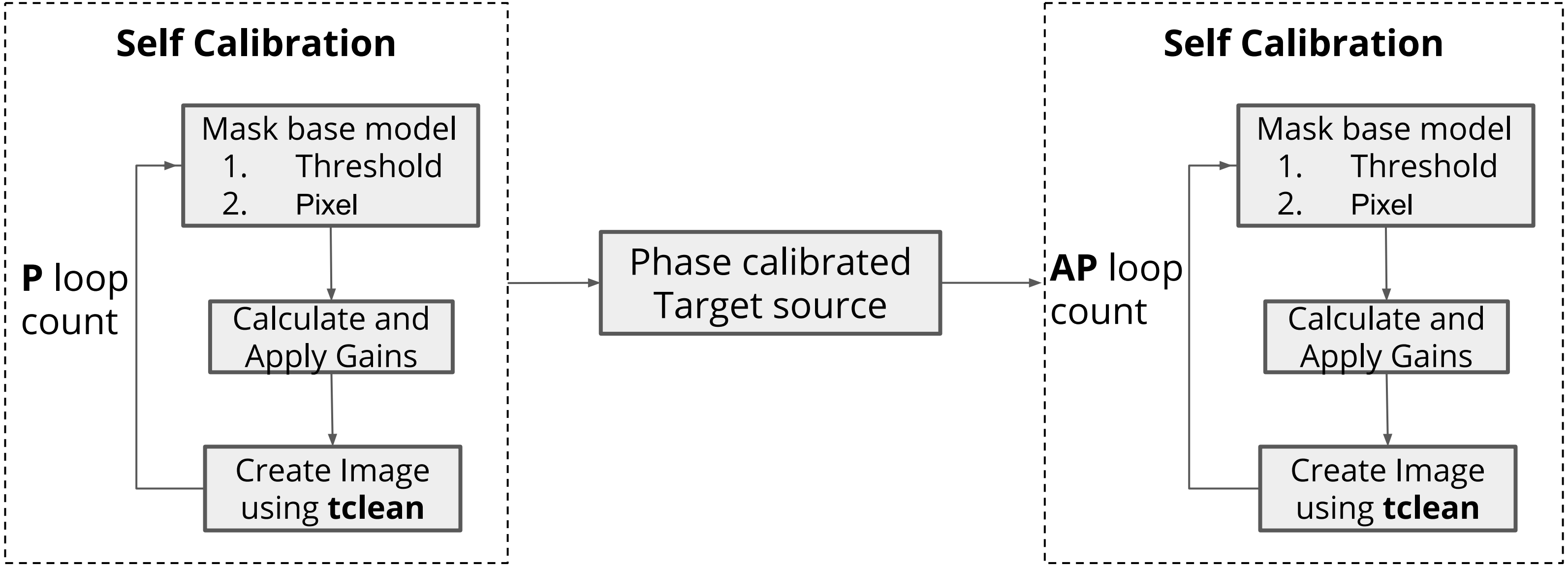
Detail Flagging

1. All Time
2. Antenna
3. Baseline

Base Image

Self Calibration

CONTINUUM IMAGING : SELF CALIBRATION



Self Calibration

Mask base model
1. Threshold
2. Pixel

Calculate and Apply Gains

Create Image using **tclean**

P loop count

Phase calibrated Target source

Self Calibration

Mask base model
1. Threshold
2. Pixel

Calculate and Apply Gains

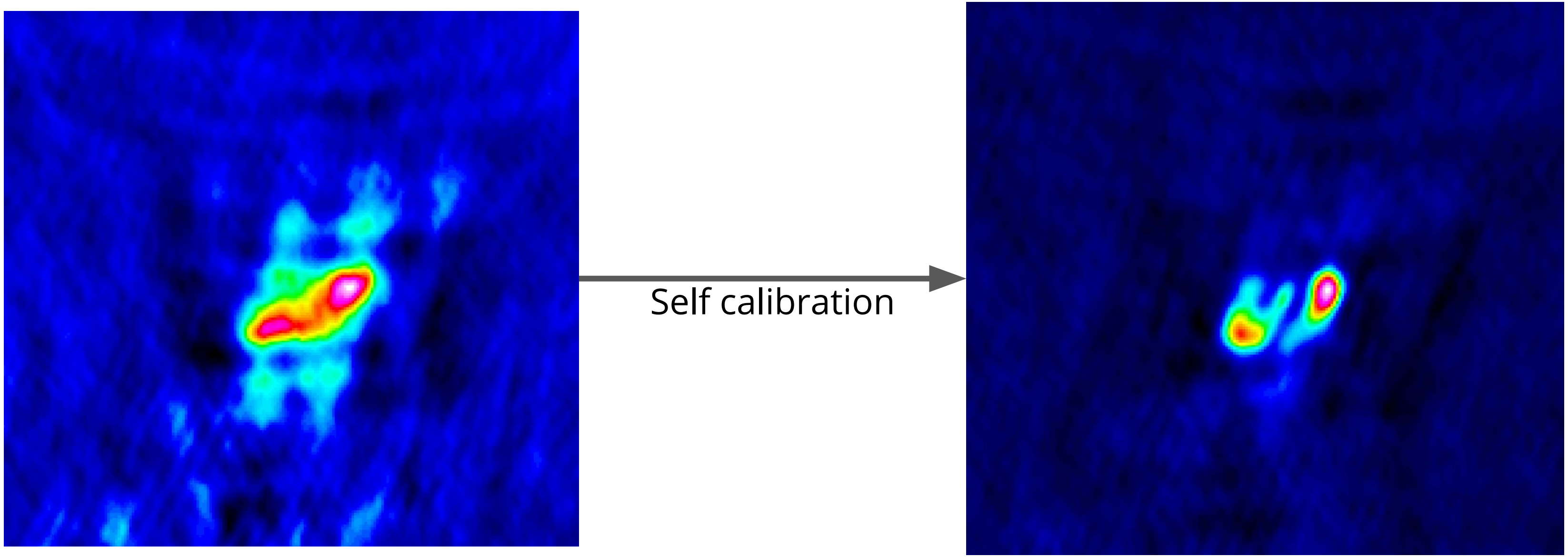
Create Image using **tclean**

AP loop count

Phase only Calibration

Amplitude Phase Calibration

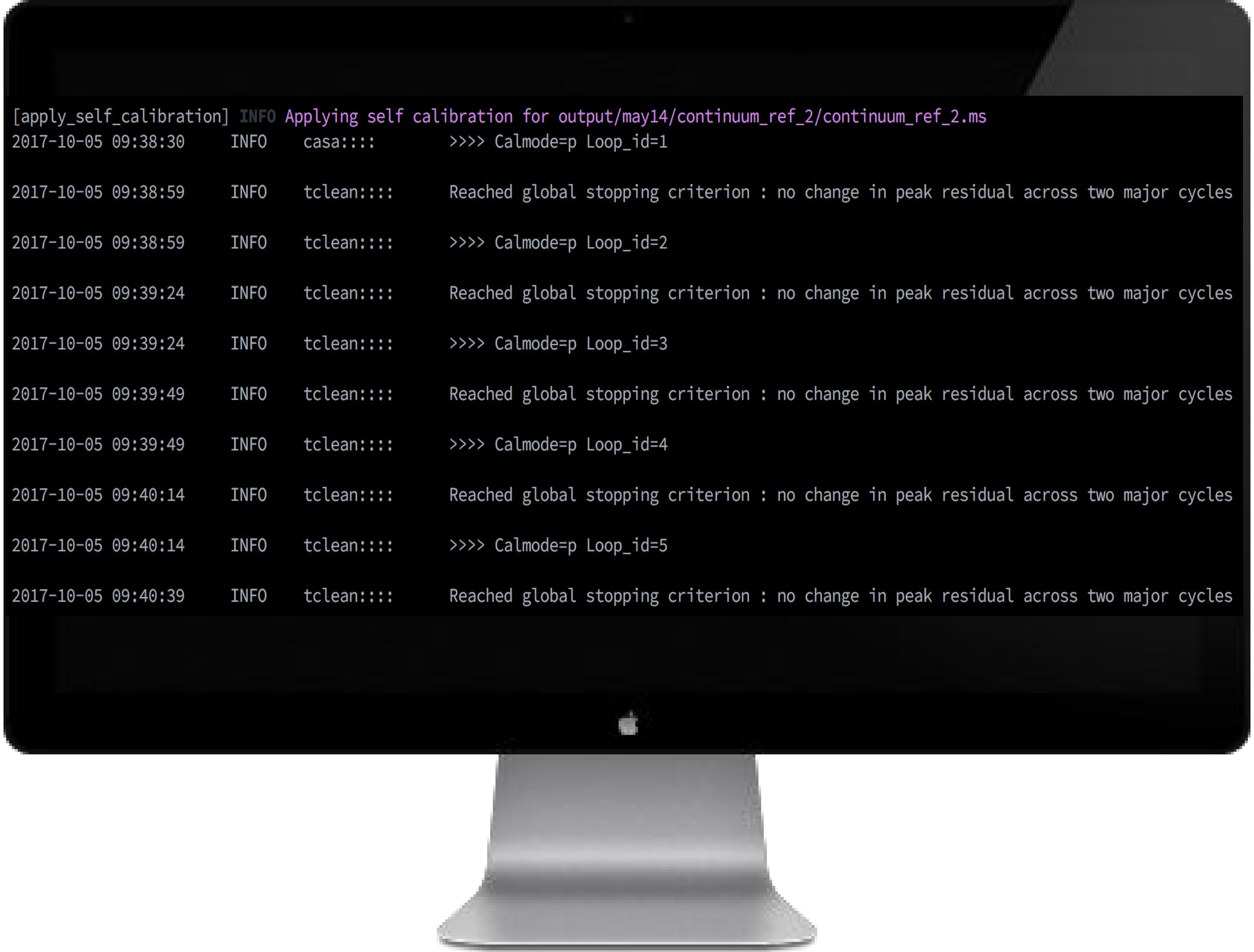
CONTINUUM IMAGE: SELF CALIBRATION



Self calibration

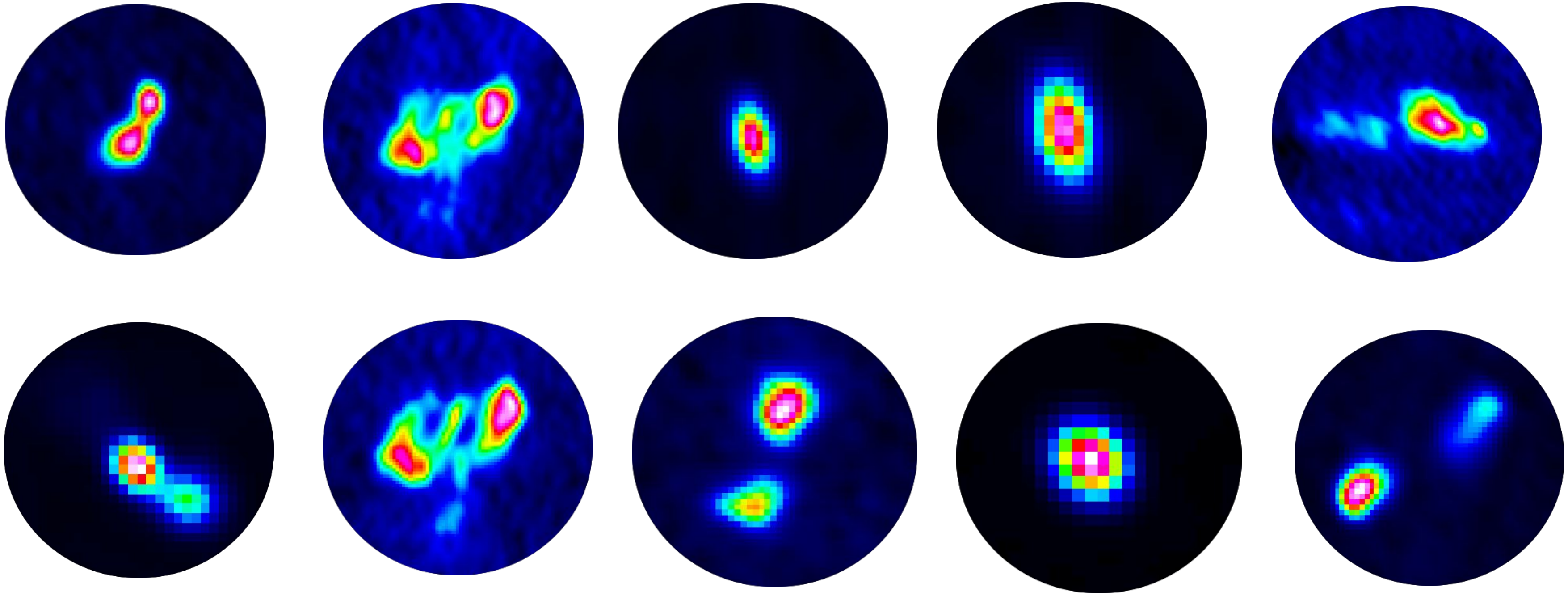
Before

After



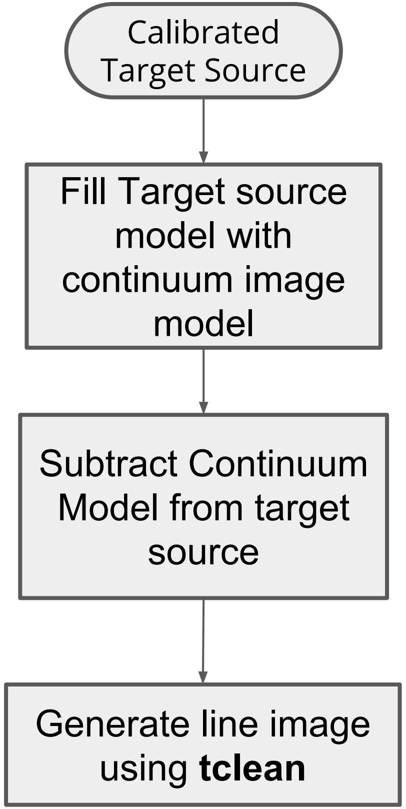
```
[apply_self_calibration] INFO Applying self calibration for output/may14/continuum_ref_2/continuum_ref_2.ms
2017-10-05 09:38:30     INFO    casa::::        >>>> Calmode=p Loop_id=1

2017-10-05 09:38:59     INFO    tclean::::      Reached global stopping criterion : no change in peak residual across two major cycles

2017-10-05 09:38:59     INFO    tclean::::      >>>> Calmode=p Loop_id=2

2017-10-05 09:39:24     INFO    tclean::::      Reached global stopping criterion : no change in peak residual across two major cycles

2017-10-05 09:39:24     INFO    tclean::::      >>>> Calmode=p Loop_id=3

2017-10-05 09:39:49     INFO    tclean::::      Reached global stopping criterion : no change in peak residual across two major cycles

2017-10-05 09:39:49     INFO    tclean::::      >>>> Calmode=p Loop_id=4

2017-10-05 09:40:14     INFO    tclean::::      Reached global stopping criterion : no change in peak residual across two major cycles

2017-10-05 09:40:14     INFO    tclean::::      >>>> Calmode=p Loop_id=5

2017-10-05 09:40:39     INFO    tclean::::      Reached global stopping criterion : no change in peak residual across two major cycles
```

IMAGES GENERATED BY THE PIPELINE



Data size = 10GB, Bandwidth = 4 MHz, Channels = 512;
Validated quality of data products and pipeline performance for standard GMRT modes.

LINE IMAGING

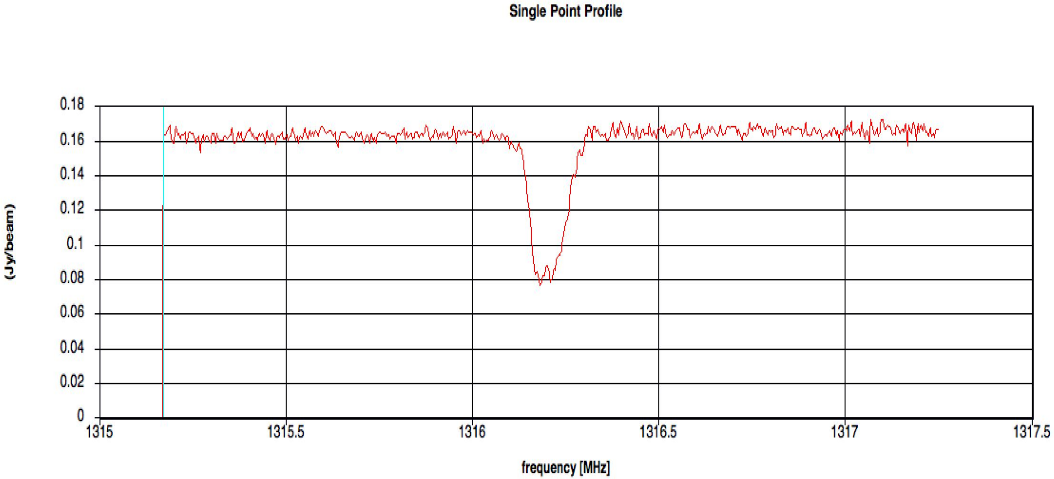


Calibrated Target Source

Fill Target source model with continuum image model

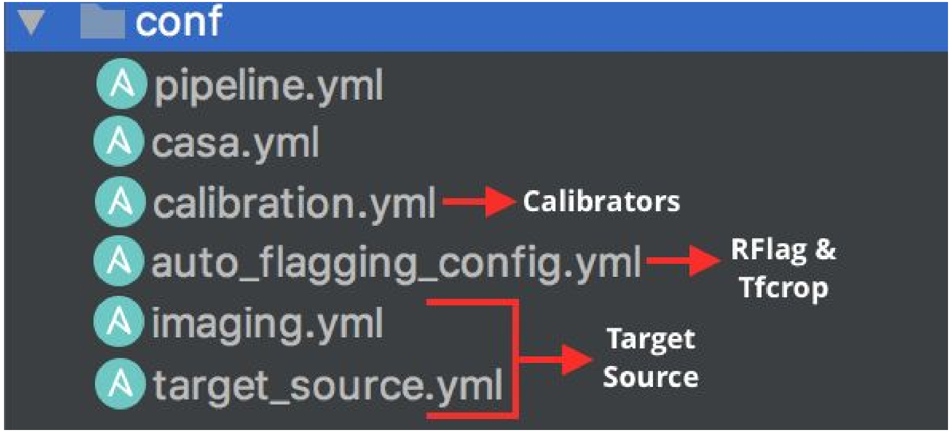
Subtract Continuum Model from target source

Generate line image using **tclean**

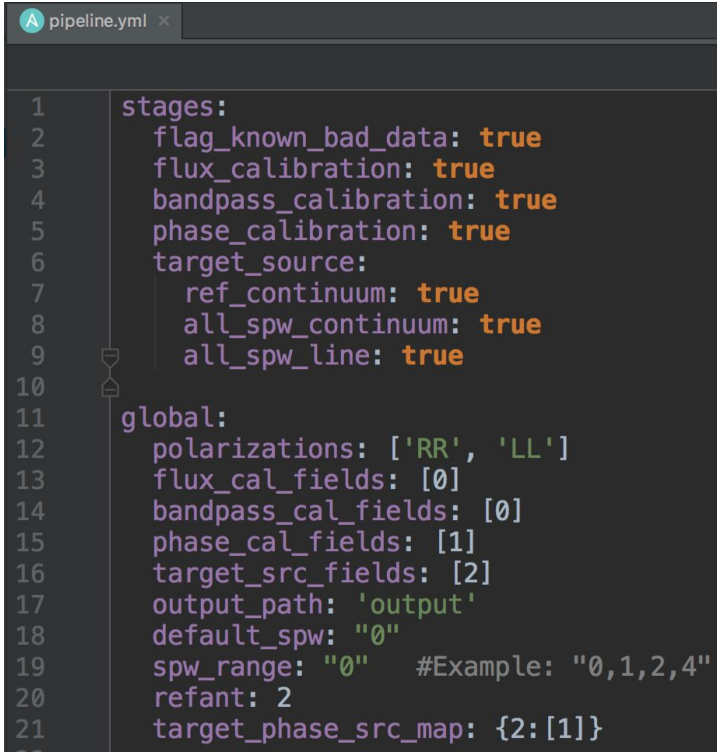


HI 21-cm absorption: signature of cold gas in galaxy (fuel for star formation)

PIPELINE ARCHITECTURE: CONFIGURATIONS



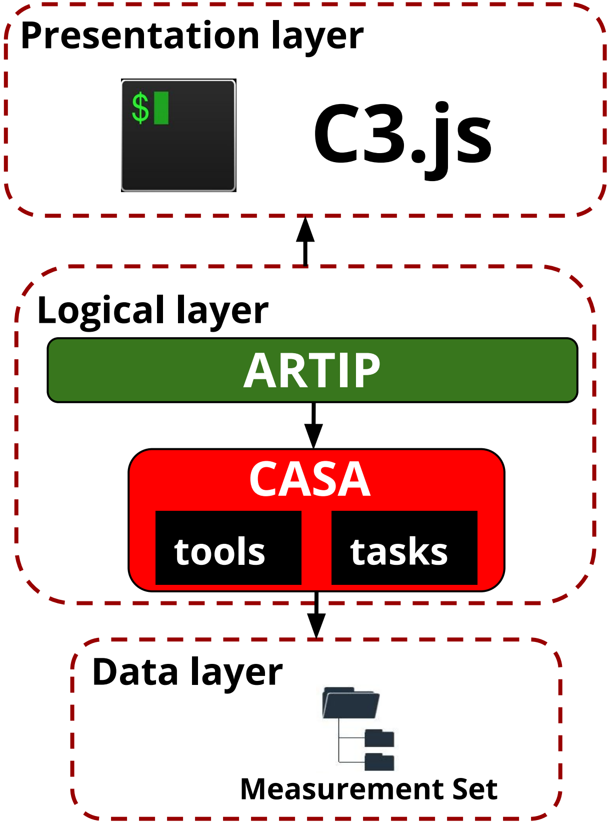
```
conf
  pipeline.yml
  casa.yml
  calibration.yml
  auto_flagging_config.yml
  imaging.yml
  target_source.yml
```



```yaml
stages:
  flag_known_bad_data: true
  flux_calibration: true
  bandpass_calibration: true
  phase_calibration: true
  target_source:
    ref_continuum: true
    all_spw_continuum: true
    all_spw_line: true

global:
  polarizations: ['RR', 'LL']
  flux_cal_fields: [0]
  bandpass_cal_fields: [0]
  phase_cal_fields: [1]
  target_src_fields: [2]
  output_path: 'output'
  default_spw: "0"
  spw_range: "0"   #Example: "0,1,2,4"
  refant: 2
  target_phase_src_map: {2:[1]}
```

Pipeline configuration

PIPELINE ARCHITECTURE: TECH STACK



Presentation layer

$

C3.js

Logical layer

ARTIP

CASA

tools

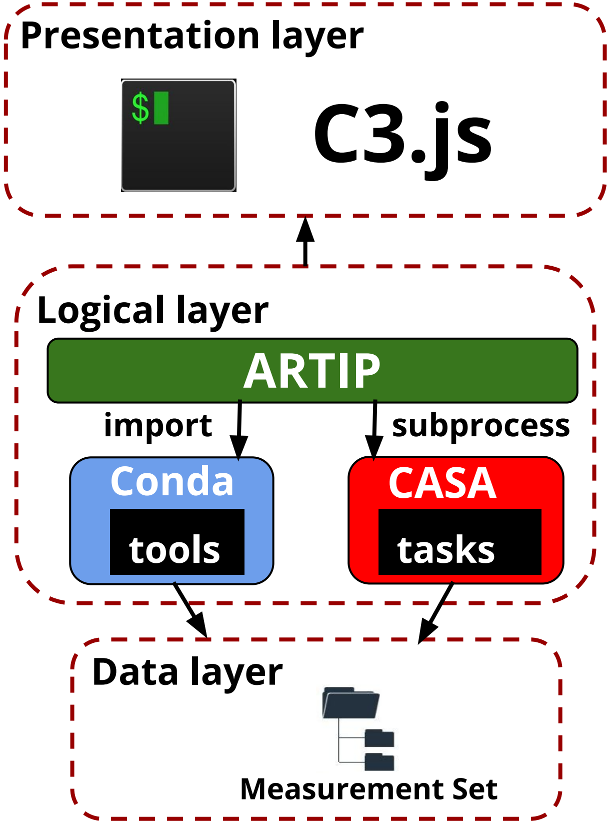
tasks

Data layer

Measurement Set

https://github.com/pkgw/conda-recipes

PIPELINE ARCHITECTURE: TECH STACK

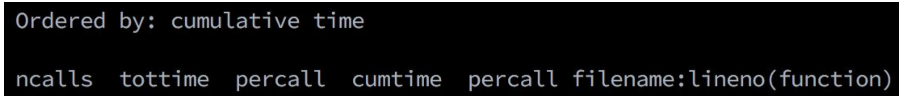


https://github.com/pkgw/conda-recipes

PIPELINE ARCHITECTURE: TECH STACK

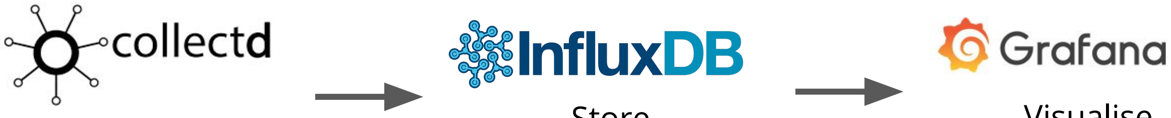
Code Profiling

Python cProfile :



```
Ordered by: cumulative time

ncalls  tottime  percall  cumtime  percall filename:lineno(function)
```

System Monitoring



Collect System Metrics

Store Metrics

Visualise Metrics

Profiling Tools

PIPELINE ARCHITECTURE: SETUP

- Fully Automated
- All pipeline dependencies/libraries are installed in a separate conda environment
- Tested on OS X and Linux platform

Prerequisites :

- Anaconda Python 2.7
- CASA 4.7.2

Setup Time : **~ 35 minutes**
Disk Space : **~ 2.5 GB**

PIPELINE PERFORMANCE

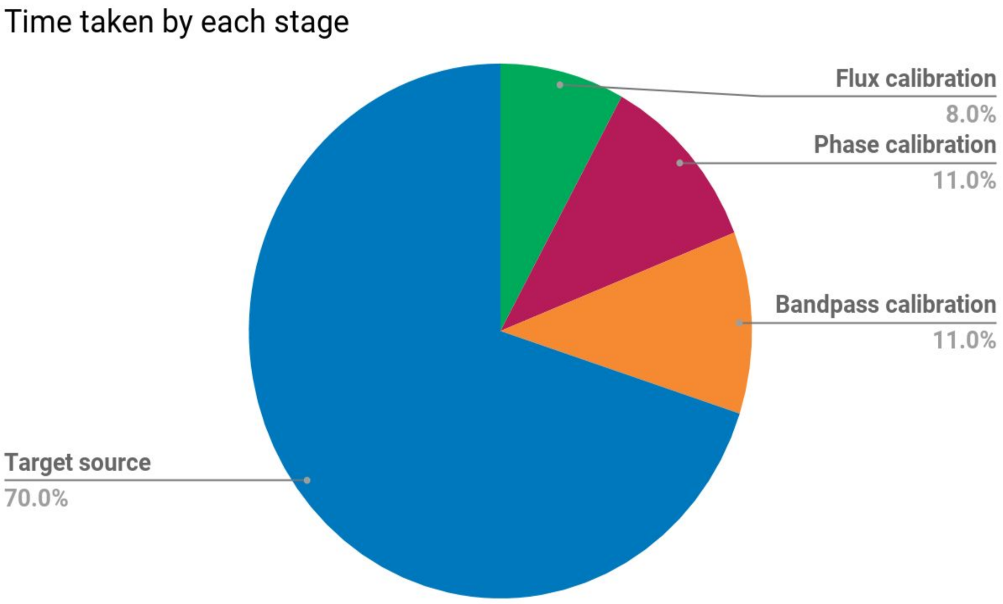
Specs:

RAM - 256 GB

Cores - 40

Storage - 18 TB

Data volume: 8 GB



Time taken: 20 minutes (Sequential)

EFFORTS AND CONTRIBUTORS

Dr. Neeraj Gupta

Dolly Gyanchandani

Unmesh Joshi

Sarang Kulkarni

Santosh Mahale

Arti Pande

Vineet Pathak

Ravi Sharma

Gunjan Shukla

Chhaya Yadav

Thank you !