DALiuGE/CASA based processing for the extragalactic HI observations with FAST.

V. Kitaeff,[1] M. Zhu, [2] L. Staveley-Smith,[3] R. Tobar,[4] K. Vinsen,[5] A. Wicenec,[6] C. Wu[7]

[1]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* slava.kitaeff@uwa.edu.au

[2]*National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China;* mz@nao.cas.cn

[3]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* lister.staveley-smith@uwa.edu.au

[4]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* rtobar@icrar.org

[5]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* kevin.vinsen@uwa.edu.au

[6]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* andreas.wicenec@uwa.edu.au

[7]*The International Centre for Radio Astronomy Research, The University of Western Australia, Australia;* chen.wu@uwa.edu.au

Abstract.

We present a prototype for the spectral-line data reduction pipeline based on the graph-based execution framework DALiuGE, and the CASA single-dish spectral-line package. The pipeline has been designed for the drift-scan mode of FAST multi-beam telescope targeting extra-galactic HI observations.

1. Introduction

FAST is planning a multi-beam multi-purpose survey that includes an extragalactic HI survey (Li et al. 2018).

We have performed a design study for the software options to built a pipeline and design a framework that will be flexible, scalable, and overall future proof option to reduce the data for FAST-HI extragalactic survey. We've considered a number of options such as LiveData/Gridzila, ASAP, proprietary code development, CASA. At this stage we have selected the CASA (McMullin et al. 2007) single dish spectral line software option that has been just refurbished into a new more compact interface started in version 5.0. At the time of writing the paper v 5.1 is already available however all the tests were performed using v.5.0.

It had been thought that a robust way of data management should be a part of the development. ICRAR has many years of experience of using and developing the NGAS software. NGAS is reliable and user friendly software used to manage such astronomy archives as MWA, that contains over 25PB of observations (Wu et al. 2013). NGAS was an obvious choice for us to include in the framework.

DALiuGE is another software package and framework developed by ICRAR to support astronomy data reduction pipelines (Wu et al. 2017). DALiuGE is agnostic to the software used in the modules, however, it does require a high level of modularity of the software that comprises a pipeline. Modules using different software and different languages can be mixed together in the same pipeline as long as the inputs and outputs of each module are defined. DALiuGE is data driven and take care of these inputs and outputs, while NGAS takes care of archiving the final data products. The developed prototype pipeline is a mix of bash scripts, Python code, with underlying C/C++ code as part of CASA.

For the quality assessment purposes, the pipeline requires to output some intermediate data products, which can be inspected later or at the time of the continues batch processing. CASA does satisfy this requirement providing an extensive set of tools for inspecting the intermediate data products.

2. Software packages

The Data Activated Liu Graph Engine (DALiuGE) developed by ICRAR is an execution framework for processing large astronomical datasets at a scale required by the SKA1 (Wu et al. 2017), (`https://github.com/ICRAR/daliuge`). It includes an interface for expressing complex data reduction pipelines consisting of both data sets and algorithmic components and an implementation run-time to execute such pipelines on distributed resources. By mapping the logical view of a pipeline to its physical realisation, DALiuGE separates the concerns of multiple stakeholders, allowing them to collectively optimise large-scale data processing solutions in a coherent manner. The execution in DALiuGE is data-activated, where each individual data item autonomously triggers the processing on itself. Such decentralisation also makes the execution framework scalable and flexible.

The Next Generation Archive System (NGAS) is a very feature rich, archive handling and management system (`https://github.com/ICRAR/ngas`). In its core it is a HTTP based object storage system. It can be deployed on single small servers, or in globally distributed clusters. It is possible to run more than one server on a single host and it is possible to run many servers across hundreds of nodes as well as across various sites. The more advanced features allow mirroring of sites running independent NGAS clusters, but it is also possible to run multiple clusters against a central database.

Some important, for this study, features include the archiving and retrieving data programmatically, data integrity checking via various checksum methods, server-side data compression and filtering, automatic mirroring of data, clustering and swarming, disk tracking and off-line data transfer, and high customisation via user-provided plug-ins.

CASA, the Common Astronomy Software Applications package, is being developed with the primary goal of supporting the data post-processing needs of the next generation of radio astronomical telescopes such as ALMA and VLA (McMullin et al. 2007), (`https://casaguides.nrao.edu/index.php/Main_Page`). The package

can process both interferometric and single dish data, and is developed by an international consortium of scientists based at NRAO, ESO, NAOJ, ASIAA, CSIRO, and ASTRON. The CASA infrastructure consists of a set of C++ tools bundled together under an iPython interface as a set of data reduction tasks. This structure provides flexibility to process the data via task interface or as a python script. In addition to the data reduction tasks, many post-processing tools are available for even more flexibility and special purpose reduction needs.

The Single Dish tool was initially developed by CSIRO based on the ASAP software package. Beginning from release 5.0.0 the development is driven by ALMA. *Sdcal* function contains calibration modes that make CASA suitable tool for the planned drift-scan HI survey with FAST.

3. FAST-HI pipeline

The software currently provides six modules:

1. *FASTcal* – based on *sdcal* that implements a single-dish data calibration scheme similar to that of interferometry, i.e., generate calibration tables (caltables) and apply them.

2. *FASTimaging* – mapping Tsys and Tsky calibrated data onto an image grid.

3. *FASTflagging* – based on *dataflag* that flags an MS or a calibration table.

4. *FASTMLflagging* – convolutional neural network inference automatic RFI flagging (in works).

5. *FASTbaseline* – based on *sdbaseline* that performs baseline fitting/subtraction for single-dish spectra.

6. *FASTexportfits* – converts a CASA image to a FITS file in accordance with FITS 3.0 standard.

There are three processing modes currently available and prototyped in DALiuGE logical graphs: real-time calibration, imaging, and reprocessing.

The real-time calibration pipeline assumes that the observation dataset is copied into NGAS archive as soon as it becomes available at the telescope. This will trigger deployment of calibrate_pipeline that will calibrate the observation and copy the resulting datasets into an archive.

The imaging pipeline provides gridding and imaging for selected observations. The configuration file contains a list of observations to be imaged. The image is produced as a measurement set and then exported as FITS cube.

The reprocessing pipeline combines flagging, calibration and imaging step from the archive. This is computationally extensive mode that normally requires a compute cluster.

CASA dataflag provides the algorithms to flag RFI. */conf/RFIflagging.conf* configuration file demonstrates how to do it. However, a part of this design study we've made an attempt to begin the development of new approach for the single dish spectral-line observation RFI flagging based convolutional neural networks algorithm using PyTorch. This technique is promising in characterisation of the RFIs that are specific to

the telescope in a location, therefore can be more accurate than signal processing based techniques in flagging difficult cases of RFI that have a broadband continues nature of the signal. At the time of writing this documentation there's no multi-feed data available from FAST yet. While the training code already exists , and have been tested with a small subset of Arecibo ALFALFA data, FAST spectral-line data will be required first before such flagging can be enabled.

CASA uses the data format known as a Measurement Set (MS). All data must be imported into this format before CASA can be used. The current version of MS is V2.0 (`https://casa.nrao.edu/Memos/229.html`) but version 3 is currently been developing. CASA from version V5.0 is no longer supporting any other import except ASDM into MS, therefore such conversion must be done outside CASA using casacore library (`https://github.com/casacore/casacore`) and a custom code.

The current plan for FAST is to produce SDFITS during the observations. Due to the complications with the calibrator signal SDFITS files may not contain all the necessary data for the calibration. This means that all the data would need to be recombined once downloaded from the archive (or archives) before processing. This seemed to be unavoidable step can be used to convert the recombined data into measurement sets, and all the calibrated data, intermediate data products, and intermediate images then can be stored in this format.

Default MS directory/file structure is not ideal for archiving the data as a single file if the metadata needs to be quickly accessed in archive. This can be solved by storing MS data model in a different format, such as HDF5 or ADIOS.

Each processing module can have any number of configuration files. If a module is executed without specifying one, it will try using a default configuration. If it cannot find the default configuration, it will create one. The *conf* directory contains some configuration files. These files are not the default configurations, but rather, those that were used during testing on ICRAR test system, and likely would need to be modified when the tests are done on a different system using different datasets.

At the time of this design study no FAST multi-beam data was available. To test the framework and demonstrate the pipelines we have used the CASA spectral line single dish reference dataset M100. FAST-HI demo logical graphs reproduce the results of CASA tutorial `https://casaguides.nrao.edu/index.php/M100_Band3_SingleDish_5.1`. While the reference dataset and tutorial are not drift-scan (they are ON-OFF pointed observations), all the steps are useful for testing purposes. More work will be needed to make sure that drift-scan calibration is done correctly once FAST data becomes available.

4. Conclusion

We have studied various options for the software to build data reduction tools for the planned FAST HI extragalactic survey. We have developed a prototype that is scalable, extendible and simple to use and develop further. Commissioning the FAST telescope with 19 beam receiver will allow testing the software on real data, and training machine learning based RFI flagging in near future.

References

Li, D., Wang, P., Qian, L., Krco, M., Dunning, A., Jiang, P., Yue, Y., Jin, C., Zhu, Y., Pan, Z., & Nan, R. 2018, IEEE Microwave Magazine, 19, 112

McMullin, J. P., Waters, B., Schiebel, D., Young, W., & Golap, K. 2007, Astronomical Data Analysis Software and Systems XVI (ASP Conf. Ser. 376), ed. R. A. Shaw, F. Hill, and D. J. Bell (San Francisco, CA: ASP), 127

Wu, C., Tobar, R., Vinsen, K., Wicenec, A., Pallot, D., Lao, B., Wang, R., An, T., Boulton, M., Cooper, I., Dodson, R., Dolensky, M., Mei, Y., & Wang, F. 2017, Astronomy and Computing, 20, 1. URL http://www.sciencedirect.com/science/article/pii/S2213133716301214

Wu, C., Wicenec, A., Pallot, D., & Checcucci, A. 2013, Experimental Astronomy, 36, 679. 1308.6083