Astronomical Data Analysis Software and Systems XXV
ASP Conference Series, Vol. 512
Nuria P. F. Lorente, Keith Shortridge, and Randall Wayth, eds.

CASDA: The CSIRO ASKAP Science Data Archive

J. M. Chapman,[1] J. Dempsey,[2], D. Miller,[2] I. Heywood,[1] J. Pritchard,[2]
E. Sangster,[2] M. Whiting,[1] and M. Dart[3]

[1]*CSIRO Astronomy & Space Science, Marsfield, NSW, Australia*

[2]*CSIRO Information Management & Technology, Yarralumla, ACT, Australia*

[3]*Pawsey Supercomputing Centre, Kensington, Perth, WA, Australia*

Abstract. ASKAP is an array of 36 radio antennas, located at the Murchison Radio Observatory, in Western Australia. Radio astronomy signals collected with the ASKAP antennas are transferred to the Pawsey Supercomputing Centre in Perth where they are processed, archived, and made available to astronomers. Astronomers interact with ASKAP data products through the CSIRO ASKAP Science Data Archive (CASDA). CASDA provides search and discovery tools using the CSIRO Data Access Portal (DAP) and international Virtual Observatory (VO) protocols. The first CASDA production release took place on 5 November 2015.

1. CASDA Overview

ASKAP is an array of thirty-six 12-m diameter radio antennas, located at the Murchison Radio Observatory (MRO) in Western Australia. The ASKAP antennas are equipped with innovative phased array feed (PAF) receivers. These provide up to 36 separate beams on the sky providing a field-of-view of 30 square degrees. This technology will enable ASKAP to carry out sensitive large-scale surveys of the Southern Sky for frequencies in the range from 700 to 1,800 MHz.

In 2014, six ASKAP antennas equipped with first-generation PAFs and nine beams per antenna were commissioned for use as the Beta Engineering Test Array (BETA). Observations taken with BETA for engineering and science commissioning purposes have since provided excellent demonstrations of ASKAP capabilities. Second-generation (MkII) PAFs that provide the full set of 36 beams, are now being installed on the ASKAP antennas. The ASKAP Early Science program, carried out with a minimum of 12 antennas and MkII PAFs, is expected to begin around mid-2016.

The CSIRO ASKAP Science Archive (CASDA) will provide astronomers with data products produced from ASKAP observations. CASDA is a critical component of the end-to-end ASKAP systems. All ASKAP data products from major surveys will be made openly available to the global community, after a process of data quality validation.

Figure 1 illustrates the ASKAP data flow for the MRO and Pawsey Supercomputing Centre (Pawsey). Radio astronomy signals collected with the ASKAP antennas are formed into beams and correlated at the MRO to produce uncalibrated (raw) visibili-

ties. These are transferred to Pawsey in Perth, Australia over four 10 Gigabit per second links provided by the Australian Academic and Research Network (AARNET).

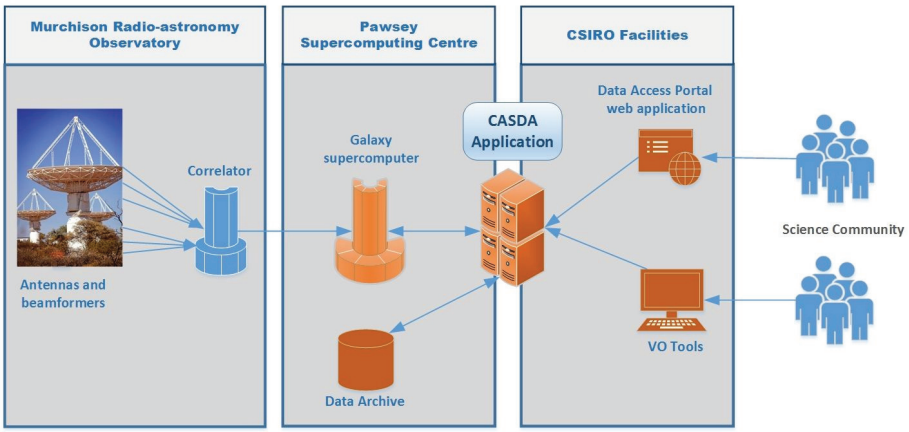

Figure 1. A simplified overview of the ASKAP data flow, showing the transfer of data from the Murchison Radio Observatory to Pawsey and the science archive.

For full operations the ingest data rate of ASKAP will be equivalent to about 75 Petabytes (PB) per year. Given this extremely high data rate, the uncalibrated visibilities are not stored. Instead, quasi real-time data processing is carried out using high performance computing algorithms and pipelines running on the Pawsey *Galaxy* supercomputer. The output data products are archived on tape and disk storage in Pawsey and made available to astronomers through the CASDA application. Storage is managed using the Next-Generation Archive System (NGAS) (Wu et al. 2013).

The pipeline data processing produces three types of data products: calibrated visibilities; images and image cubes; and source detection catalogues (see Whiting et al., 2015, these proceedings). CASDA is designed to handle an ingest rate of 16 TB per day. The application supports:

- Long term data tape and disk storage at Pawsey (up to 5 PB per year).
- Access to data products using web services provided through the CSIRO Data Access Portal (DAP). The DAP is an enterprise-wide system that archives and provides access to data across many areas of CSIRO research.
- Data access through Virtual Observatory services: CASDA supports VO protocols including the Table Access Protocol, the Cone Search Protocol and the Simple Image Access Protocol (version 2). Researchers are encouraged to use the CASDA VO services together with external applications such as TOPCAT (Taylor 2014) and Aladin Sky Atlas (Bonnarel et al. 2000).
- Tools for setting data validation flags and information.
- Uploads of science catalogues provided by science teams from analysis of ASKAP data products.
- User authentication – CASDA requires authentication for transfer of data files and for some tasks such as data validation.

- Digital Object Identifiers and persistent identifiers; these provide a permanent record of the data set at the time of access.

- Administration tasks to support system users, and to monitor the system usage and performance.

2. CASDA Implementation and Data Deposit

CASDA has been implemented across two data centres, at Pawsey for operations that need to be close to the data and at the CSIRO data centre in Canberra for interactive services. The Pawsey infrastructure resources that support the operations of CASDA are shown in Table 1. The division of modules is shown in Figure 2. All modules are written in Java, use the Spring application framework and communicate using RESTful web services. The CASDA metadata server is a PostgreSQL database running PGSphere.

Table 1. CASDA infrastructure at the Pawsey Supercomputing Centre

	Components	**Storage/notes**
Development, test and acceptance environments	Five physical servers	256 TB on SGI CXFS disks
Production environment	Two physical servers	256 TB (initial allocation)
Long term data storage	Tape archives	2 x 10 PB (initial allocation)
Internal networks	SGI Infiniband, Cray Infiniband	56 Gbps (storage), 10 Gbps (other)

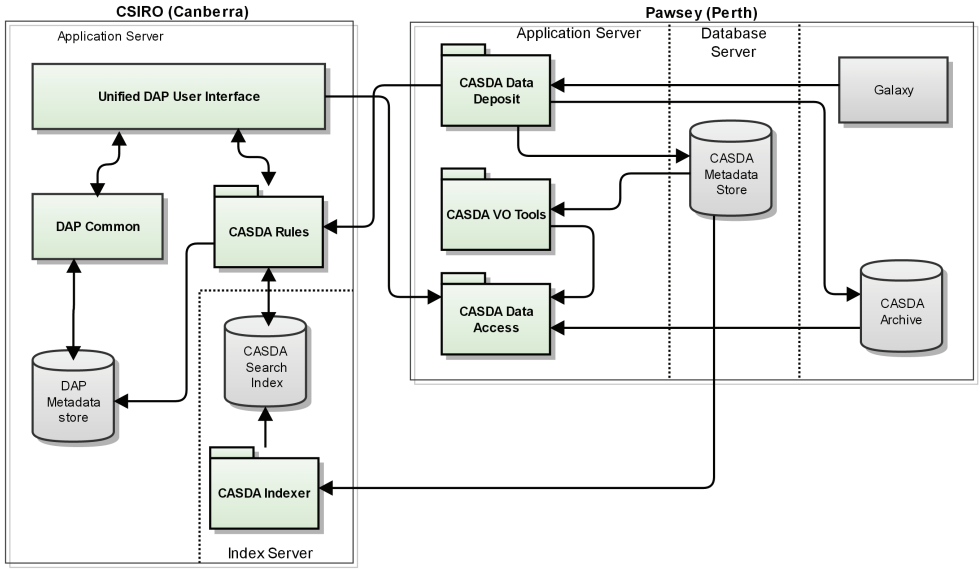

Figure 2. Major modules in the CASDA Software Architecture. Green boxes are software modules and grey are databases and file systems used by CASDA.

We now discuss the data deposit module as an example of CASDA implementation. The process of depositing data products is broken down into discrete steps such as reading the FITS metadata, importing catalogue data into the database, or registering a data product with NGAS. Each of these is implemented as a command line process, either in Java or as native commands. Depositing all the data products associated with a scheduling block is orchestrated by a Java process control application, which includes an administration web interface. The steps are then scheduled using SLURM, a standard job manager tool. This configuration allows both control of concurrent jobs and the ability to scale horizontally for use with additional servers.

The development of the CASDA deposit process has used two techniques which we consider important to its success. Each major process was prototyped in advance to test out implementation strategies and allow a fast fail for unsuitable approaches and the refinement of the adopted approaches. Regular performance testing ensures we can meet the requirement of processing 16 TB within 24 hours.

2.1. CASDA Virtual Observatory Tools

The implementation of Virtual Observatory (VO) standards is a major part of the CASDA application. To provide these, we have made use of the open source Astronomical Data Query Language (ADQL) (`http://cdsportal.u-strasbg.fr/uwstuto/`) and Universal Worker Service (UWS) (`http://cdsportal.u-strasbg.fr/adqltuto`) libraries obtained from CDS. We have built a continuous delivery pipeline that includes automated testing of every change against the VO standards using the TAPLint and VOTLint tools from the STILTS (`http://www.starlink.ac.uk/stilts/`) package. The CASDA VO library is available for use by other data centres.

3. CASDA Public Release

CASDA was first released on 5 November 2015. The initial release includes around 3 TB of data products produced from several pilot science studies carried out with BETA. These include searches for highly red-shifted neutral hydrogen (HI) towards quasars (Allison et al. 2015), radio continuum observations of an intermittent pulsar (Hobbs et al. 2016), HI imaging of a group of galaxies (Serra et al. 2015) and widefield mosaic radio continuum imaging of the Tucana region (Heywood et al. 2016).

To get started with CASDA see the CASDA Users Guide (`http://www.atnf.csiro.au/observers/data/casdaguide.html`).

Acknowledgments. This paper is written on behalf of the full CASDA team. We thank the many individuals who have contributed to the project.

References

Allison, J. R., et al. 2015, MNRAS, 453, 1249
Bonnarel, F., et al. 2000, A&AS, 143, 33
Heywood, I., et al. 2016, MNRAS, 457, 4160. 1601.05857
Hobbs, G., et al. 2016, MNRAS, 456, 3948. 1512.02702
Serra, P., et al. 2015, MNRAS, 452, 2680
Taylor, M. B. 2014, in ADASS XXIII, edited by N. Manset, & P. Forshay, vol. 485 of ASP Conf. Ser., 257
Wu, C., et al. 2013, Experimental Astronomy, 36, 679