Highlights of Astronomy, Volume 16
XXVIIIth IAU General Assembly, August 2012
T. Montmerle, ed.

doi:10.1017/S1743921314012848

Taming the ALMA Data Avalanche

Felix Stoehr

ESO, Karl-Schwarzschild-Str 2, 85748 Garching, Germany
email: fstoehr@eso.org

Abstract. The Atacama Large Millimeter/submillimeter Array (ALMA) is nearing its phase of full operations. ALMA will collect about 200TB/year of astronomical data which will be reduced by an automatic pipeline and turned into fully calibrated science-ready data products. We present design choices, challenges and solutions from data capturing over data reduction and data distribution to archival research, that allow to deal with the large amounts of data and, hopefully, achieve the maximum amount of science return.

Keywords. Data intensive astronomy, ALMA, data management, archive, pipeline

1. Introduction

With its 66 antennas in the Atacama Desert in Chile and with baselines up to 16 km, ALMA will allow an unprecedented view of the the mm/sub-mm sky. Currently Early Science observations are underway, while the array continues to be completed. The start of the phase of full operations is expected to be in the second half of 2013. Out of the many challenges that had to be faced, we will present here those related to the large amounts of data that ALMA will produce.

2. Data Challenges

2.1. *Data rate*

Based on the predicted data-rate requirements of the expected type of science observations, the official ALMA data rate has been set to 200TB/yr (6MB/s) with peak values reaching 10 times these values for extended periods. These values have been recently revised using experience from the first Early Science observations (Lacy & Halstead, 2012) concluding that the expected average will be larger, up to 700TB/yr (23MB/s). The ALMA data flow system was designed with scalability in mind and was build without hard bottlenecks out of commodity hardware, so that this new data rate and future data rate increases can be coped with.

2.2. *Data storage*

Once written out by the ALMA correlator, the data is stored following an “All data taken have to be stored forever” policy. The storage solution is required to be cost-effective, scalable, future-proof, PB-scale, allow for high read/write performance despite large variation in file size, contain a built-in file-management layer, do automatic consistency checks and support globally distributed archives. NGAS, the Next Generation Archive System (Wicenec & Knudstrup, 2007) developed at ESO was chosen. It is is a combined hardware-software solution, is portable, provides a powerful plugin architecture, uses commodity hardware and allows for online processing capabilities.

2.3. *Archive copies*

The main ALMA archive is located in Santiago in Chile. Three full copies of that archive exist, one at each of the three ALMA Regional Centres (ARC): in Charlottesville at NRAO in USA, in Mitaka at NAOJ in Japan and in Garching at ESO in Germany. The copies serve as backup of the main archive and allow the ARCs to provide user-support, reprocessing, quality control, as well as phase 3 work. The data transfer happens over the network through VPN tunnels.

2.4. *Operation*

ALMA is a general purpose telescope with standard calls for proposals once a year. Also, ALMA will support a large variety of different types of observations (continuum, line, spectral sweep, mosaic, solar) with different observing settings (12m array/ACA/TP) in different configurations). The choice was to create a fully-integrated data-flow system with the Archive in the centre. A sophisticated observing tool and the ALMA Science Data Model (ASDM) have been developed in order to cope with this complexity.

2.5. *Pipeline*

The amount of data of a single project can easily exceed the data-reduction resources available to a typical user today. A fully automatic pipeline is therefore required for ALMA delivering science-grade products (a first in radio astronomy). The policy is that all ALMA science data will be reduced by the project. Indeed, as telescopes will become more complex and deliver more data, the pressure on data providers will rise to deliver science-grade products. An analysis software package CASA (Common Astronomy Software Applications) was written (nearly) from scratch. A pipeline (heuristics + CASA) was developed. Commissioning of the pipeline is currently ongoing.

2.6. *Support*

If users want to re-analyze the data e.g. an optimized reduction, for their particular science case, they need help. ALMA is meant to be a telescope that can be used by all astronomers, not only those that have a radio background. The choice was therefore, to provide extensive support to users including face-to-face support. This support is provided by the three ARCs and is already now in the early stages of the project highly appreciated.

3. Summary

High data rates and the long-term nature of ALMA are a real challenge and a huge effort was deployed at all ends to build a system that can handle the data intake. Hard and software solutions have been built to be scalable and flexible to allow to adapt to change. The goal of ALMA is to help the scientists wherever possible from proposal preparation over the science-grade data products and data-reduction to archival research and to work towards a great end-to-end user-experience.

Acknowledgements

The ALMA construction team deserves all the credit for the work presented here.

References

Lacy M. & Halstead D. 2012, *NAASC Memo* 110
Wicenec A. & Knudstrup J. 2007, *The Messenger* 129, 27

https://doi.org/10.1017/S1743921314012848 Published online by Cambridge University Press