Electronic Government and Electronic Participation
E. Tambouris et al. (Eds.)

doi:10.3233/978-1-61499-570-8-132

Open Skies with Cloud Computing

Mauricio SOLAR[a,1] and Mauricio ARAYA[a]
[a] *Universidad Técnica Federico Santa María, Chile*

Abstract. This article shows that high costs to obtain data in astronomy justifies the use of cloud computing. In such a field, like others in science, there are several benefits when this data is freely available to the whole scientific community. The availability of these data will foster innovative applications to exploit and explore data. These applications could be offered as a service in the cloud. A National Data Center for e-Science could take advantage of these high costs data and offer open data to the community to maximize data accessibility.

Keywords. open data, data science in government, cloud computing in science.

1. Introduction

The concept of open access to scientific data was institutionally established with the formation of the *World Data Center* system, in preparation for the International Geophysical Year of 1957–1958 [1]. The *International Council of Scientific Unions* (now the International Council for Science) established several World Data Centers to minimize the risk of data loss and to maximize data accessibility, further recommending in 1955 that data be made available in machine-readable form (http://www.icsu-wds.org).

More than 50 years later, in 2007, the Science Ministers of all nations of the *Organisation for Economic Co-operation and Development* (OECD), which includes 34 most developed countries of the world, published the "OECD Principles and Guidelines for Access to Research Data from Public Funding" as a soft-law recommendation [2].

In recent years, the flood of data problem is visible in both, science and business environments, becoming more relevant the proper use and management of what has been called Big Data. This concept encompasses the management research of large amounts of information, and that these are not easy to process using traditional tools and procedures. When the data volume reaches TeraBytes (TB) to ZetaBytes (ZB) ranges, algorithms and procedures must be adapted for its use in new high-performance computing platforms, with Cloud tools, in a distributed manner, and on-line. Additionally, we are not only dealing with large stationary volumes of data, but also with the data-generation frequency, and the heterogeneous nature of data sets; this creates new challenges in developing solutions, as are the storage, variability in the format, and response time.

Big Data is not a technology in itself, but rather a work approach to obtain value and benefits of these large volumes of data generated nowadays. These are some of the features to consider:

[1] Corresponding author, Informatique Department, UTFSM. E-mail: mauricio.solar@usm.cl

- How to grasp, manage and make use of them
- How to ensure, check authenticity and reliability
- How to share and obtain improvements and benefits
- How to communicate, simplifying the decision-making and subsequent analysis.

One of the domains where the Big Data problem is approaching its turning point is astronomy. Its state-of-the-art facilities in operations, such as the *Atacama Large Millimeter/submillimeter Array* (ALMA) will generate over 1 TB of data per observation day [3], and those under construction, as the *Large Synoptic Survey Telescope* (LSST), and the *Square Kilometer Array* (SKA) will produce large-scale data [4]. The plan for the year 2020 is to have more than 60 PetaBytes (PB) of accessible information for the astronomical community. The LSST initial computer requirements are estimated at 100 TFLOPS of computing power and 15 PB of storage, rising as the project collects data (full operations for a ten-year survey commencing in January 2022), (http://lsst.org/lsst/science/overview).

The cost to design, build and operate such mega-observatories is very high, so the cost of on-site observation time and the end-to-end cost to obtain astronomical data is also very high. Therefore, it would be very profitable if the obtained data could be freely available to the whole scientific community.

In Section 2 we show the costs involved in the construction of observatories. In Section 3, we show the *International Virtual Observatory Alliance* (IVOA) [5] as a good example to follow in other fields of science in the world. We also show the development of the Chilean Virtual Observatory (ChiVO) based on the standards and protocols defined by IVOA. In Section 4 we show some challenges to install a data center to offer e-science data and cloud services to the scientific community. Finally, we conclude in Section 5.

2. Big Data, Big Costs

Computationally, the concern is focused on the rapid growth of the generated data volumes, which had passed from the GigaBytes (GB) to the TB level in the past decade, and will pass from TB to PB in the near future. For example:

- *Galaxy Evolution Explorer* (GALEX): the first orbiting telescope in the space generated 30 TB of data in the first 3 years of operation [6].
- *Sloan Digital Sky Survey* (SDSS): in its 12th version gathers 116 TB of data published as images, catalogs, and other products (www.sdss.org/dr12/data_access/volume/).
- *Panoramic Survey Telescope & Rapid Response System* (Pan-STARRS): its goal is to characterize objects that will come closer to the earth, as asteroids or comets. 10 TB of data is expected per night of observation [7].

We could continue listing more observatories, but in summary, there is a growing *avalanche* of astronomical data, changing the way that astronomy is done nowadays. From the computational point of view, it is necessary to implement standard services that allow to access, process, and modeling the data produced by each observatory (which generates public data).

2.1. Cost of Observing

The privileged atmospheric conditions make Chilean skies one of the most favorable places for astronomical scientific research: over 330 clear sky nights per year. Chile hosts also the world's closest ground observatories to space, at an altitude of 5,000 meters above the sea level (masl). Clear skies, easy access, communication infrastructure, isolation from urban settlements, and effective protection of sighting sites from light/luminal contamination, preserve the area for astronomical purposes. For these reasons, many observatories have settled in the Regions of Antofagasta, Atacama and Coquimbo.

There are more than a dozen wide sweeping astronomical facilities throughout Chile; for example, the already mentioned ALMA, the "*Very Large Telescope*" (VLT), and in the coming years, the "*European Extremely Large Telescope*" (E-ELT). As a result of Chilean clear skies and appropriate weather conditions, by 2018 about 68% of the global astronomical infrastructure will be settled in Chile [8]. Table 1 shows the investment (over USD 5 billion) in new astronomical observatories installed or to be installed in Chile [8].

Table 1. Investments in observatories

State	Telescope	costs US$
Operating	Magellan Telescopes	100M
	Gemini	300M
	VLT	700M
	ACT	40M
	ALMA	1.500M
In construction	E-ELT	1.500M
Projected	GMT	800M
	TAO	100M
	LSST	500M
	CCAT	200M
	Total	**5.740M**

One of the conditions established as country is that 10% of the observing time should belong to the Chilean astronomical community; thus, it is justified the need to develop a national astro-informatic platform, for an intelligent management, and analysis.

According to the investment indicated in **Table 1**, the estimated cost of the 10% of the observation time of astronomical observatories in Chile is (considering operational costs as 5% of total investment):

- About US$30.000K in the year 2011.
- About US$87.000K in the near future (2018-2020);
- Between US$ 25.000 and US$ 30.000 is valued one hour of observation time in ALMA.

2.2. Other (Big) Costs

The current and future large-scale data generated by the astronomical observatories placed in Chile, have created new needs and requirements, which serve as an opportunity for the development of a Data Center (DC) to offer public data to the scientific community. To have a general idea of the amount of data that will need to be stored and processed, we can consider the ALMA project: the fully operational observatory (with the complete array) will generate over 1 TB of data per observation day. The handling of large-data volumes can generate complications and high costs in the following issues:

- **Storage**: it is necessary to have a DC capable of storage according the data-consumption needs, without ignoring the physical space of the equipment and the architecture behind the storage procedure.
- **Access**: the open data requirements of astronomical data sets stress the necessity of a reliable and fast access from anywhere and for anyone. A web-based system under a high-speed networking environment is an appropriate mechanism to supply this requirement.
- **Processing**: data processing is an important area of applied computer science, which is concern on understanding the nature and the structure of the data, as well as the development of tools and techniques that can be used to carry it out. When the processing to accomplish deals with large volumes of data --- whether it is correcting, calibrating, analyzing, etc.---, it requires more time than usual, and of course, more computational power, resources that a user usually does not have.

Thus, the need of a system with these characteristics, as a solution to public access and advanced manipulation of the large-scale astronomic data, is an excellent good to maximize the availability of public high-cost data. The idea of a Virtual Observatory (VO) was born in the year 2002, and the standards and protocols of how to inter-operate between different resources are in charge of IVOA [5], [9], [10].

The VO paradigm is an international initiative that enables universal access to astronomical data, under the responsibility of specialized centers for its storage and processing, at which both astronomers and regular people may access. With the standardization of methods and information is possible to study the astronomic registries without requesting new observations, reducing the physical requirements of instruments and locations, and minimizing data duplication.

2.3. Benefits of VO

The cost of acquiring new astronomical data is too high. Moreover, astronomers use their data to proof, test or discard their theories once, and after that they usually don't need more those data. This is why observatories publish this kind of data as a public data after one year (or more depending on the observatory's policy).

ALMA is an astronomical interferometer of radio telescopes using two or more radio antennas. When combining these signals to analyze them, it is possible to obtain detailed information of the source of emission with unprecedented resolutions. From the standpoint of data production, this process generates 3D cubes given by two position shafts in the celestial vault, and one shaft of frequency spectrum (Figure 1). The uniqueness of these data cubes is their size, since the spatial and spectral high resolution that this observatory provides generates large-scale data cubes (GB and TB).

An astronomer will receive a big file containing more information than he needs, because he wants to proof, test or reject his theory with a very small portion of the data cube he received. Usually, only a small part of the image is actually analyzed (see the square on the left side of Figure 1), or only a slice of the cube is used. Even though this is correct from the astronomer point of view, all the rest of information could be very helpful to other astronomers.

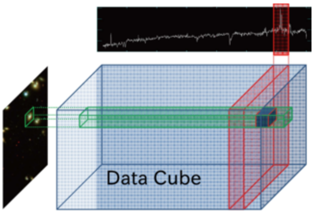

Figure 1. ALMA data cube

Open data allows reusing these very expensive data, and making it available to other scientists. In fact, the use of this open data increases the number of publications in astronomy: about 3 per researcher per year [11]. The GAVO Millennium database service is been referred to in more than 10% of the "Millennium" publications [12].

In the case of the observatories installed in Chile, it is common that big data is transferred from the southern hemisphere to the northern hemisphere. This obviously takes long time: in most cases moving big files will take hours if not days.

3. Virtual Observatory as an example of e-science

The first important concept to understand is that the VO is not a software package that allows user access to data like a web repository, despite it is exactly that what is expected from those two words. Technically, a VO can be described as an integral architecture, which formalizes in each application level, necessary protocols and standards to interoperate between VO partners worldwide. Therefore, when an observatory wants to publish data, it will not be the need to invent a new architecture, as they can use the existing one.

IVOA's basic mission is focused on coordination and collaboration between facilities, which is the main entanglement to enable global and integrated access to data collected by the international astronomical community. In IVOA, 21 VOs are currently engaged, and is composed by several working groups that discuss about the creation and versioning of standards and protocols, including its architecture.

The Data Access Protocol (DAP) defines a family of access services interfaces to the astronomical data available through VO. The DAP describes how the data providers share the information to users, and how users retrieve information. There are several standards involved, some of these are:

- **Simple Image Access** (SIA) [13]: A query defining a rectangular region on the sky is used to query for candidate images.
- **Simple Cone Search** (SCS) [14]: The query describes sky position and an angular distance, defining a cone on the sky.
- **Simple Spectral Access** (SSA) [15]: defines an interface to remotely discover and access one-dimensional spectra.
- **Table Access Protocol** (TAP) [16]: defines a service protocol for accessing general table data. The access is provided for both database and table metadata as well as for actual table data.

Each service returns a list of candidate images, astronomical sources, data or metadata formatted as a *VOTable* [17].

3.1. Chilean Virtual Observatory (ChiVO)

Although Chile is one of the most dynamic countries in its astronomic activity in the world, it did not have a VO until a year ago. For the moment, nor has ALMA services to support the protocols and standards of other VOs; therefore, it was a challenge to pose the needs and requirements of this type of platform.

Though the international community has been working and refining VOs and their standards, every new telescope and instrument imposes new challenges and opportunities of development. This is particularly true for the ALMA data case, which introduces the Big Data problem as a current problem. This means to equip the VO that is hosting its data, with last generation technology and frontier research in its tools.

The development of the ChiVO aims the development of several VO services following IVOA standards. The ChiVO will host the data of the observatories located in Chile, adhered to the interoperability standards of IVOA.

Broadly, the architecture of IVOA has three layers:

- **Users**: astronomers and scientists in general, interested in the data published by the observatories.
- **Resources**: observatories and centers producing astronomical data.
- **Intermediate layer**: defines what a VO is; that is, how users and resources communicate using protocols and standards to search and access data.

3.2. Requirements

The creation of the ChiVO required the identification of the current needs of the national astronomy community, which can be summarized in:

- **Discover:** Find astronomical data of an object or instrument on a high dimension specific region of the space, based on the spatial, temporary, spectral shafts, red shift, polarization parameters, etc., either by search or exploration.
- **Obtain**: A download link of the required data in different formats, either through the VO or through an external service.
- **Compare**: Information crossing of the data obtained between the different sources of information.

A multidisciplinary team participated in this process (astronomers, engineers, scientists, experts in ALMA data, etc.), while the astronomy community was defining its requirements and use cases, the IT team contrasted it with international standards and designed the following architecture and development model (Figure 2).

Abstraction layers: **Clients.** This layer represents the final user, and how communication between the user and data can be simplified. In this layer the user conducts queries through the access protocols offered by ChiVO, or through an advanced form, using compatible applications with VO and its web portal. Once the query is performed, the system returns to the user a list describing the objects or observations found (metadata), and provides access to them through a downloading link associated to each result.

Abstraction Layer: Applications. In this layer, we find the programs that process the queries between users and data. The server-side analysis tools are essential for ChiVO efficiency, because data are usually large and transfer it is expensive. Bringing the analysis and processing tools closer to the place where the data is stored solves this problem as cloud computing does. Moving the algorithms to the data is faster than moving the data to the algorithms.

Abstraction Layer: Data. This layer contains resources that describe the data and metadata. This is basically a relational database that stores metadata linked to the data model recommended by IVOA (ObsCore DM).

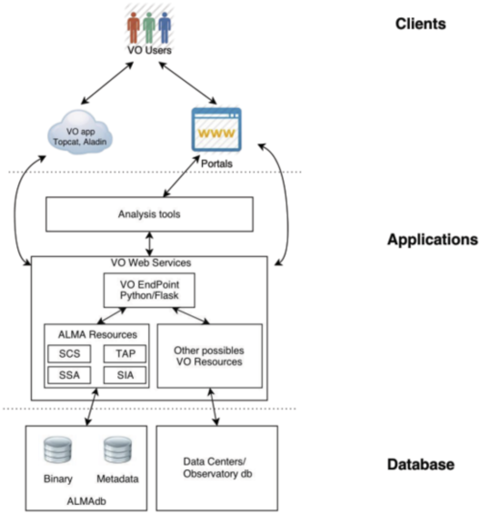

Figure 2. ChiVO architecture

4. A Data Center (DC) or Cloud Service for Open e-Science Data

Observational sciences need to secure their data because experiments are usually expensive and difficult to reproduce. In astronomy is even more important because each observation have a unique timestamp, so each experiment is irreproducible from the time domain point of view. Therefore, experimental sciences need to store their data in the order of PB of data indefinitely (ideally). Unfortunately, the cost of this storage requirement is high. In this line, there are different challenges that should be faced in Technological, Organizational, and Economical terms.

Besides those high-level challenges, Big Data science also imposes new challenges as: (a) data are being produced at an exponentially growing rate, which is currently exceeding the capabilities of local storage and processing capabilities; (b) data production and collection is expensive and it must be accessible and preserved for a very long time during operations (many decades), and even beyond the end of the specific project that provided the funding; and (c) historical data could be more important than the actualized data, so there is no clear archiving mechanism that can ease the previous challenges.

A DC for Science and Industry shall provide the capacity to store 1 EB (1000 PB) worth of data. The ChiVO DC is starting with 1 PB, will be connected to a 10 GB network, and will have a reduced environmental impact, a Power Usage Effectiveness (PUE) less than 1.3.

5. Conclusions

It is a fact that the cost of obtaining astronomical data is too high, so it does not seem reasonable to use just a portion of the data and discard it without sharing it with other scientists. With these datasets we can make further researches and even produce more and higher-level knowledge. A cloud service for storing data should provide search capabilities to the scientists that allow finding and downloading, if necessary, the required files containing the data. The offer of cloud services allows scientists to access data, and to take advantage of computing capabilities required to process big data.

Acknowledgements

This work was partially supported by the grants DGIP 241142, FONDEF D11I1060 and ICHAA 79130008.

References

[1] National Academy of Science. *Earth Observations from Space: The First 50 Years of Scientific Achievements Committee on Scientific Accomplishments of Earth Observations from Space*, National Research Council. ISBN: 0-309-11096-3, 142 pages. (2008)

[2] OECD. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. http://www.oecd.org/science/sci-tech/38500813.pdf, (2007)

[3] Atacama Large Milimeter/sub milimeter Array (ALMA). ALMA Basics - Welcome to the Science Portal at NRAO. https://almascience.nrao.edu/about-alma/alma-basics, (2013).

[4] Ball N.M., Brunner R. J. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, **19**:1049–1106, (2010).

[5] International Virtual Observatory Alliance. IVOA.net. http://www.ivoa.net/, (2014).

[6] Martin et al. The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission. *The Astrophysical Journal*, **619**:L1–L6, January, (2005)

[7] Denneau L. et al. The Pan-STARRS Moving Object Processing System. http://arxiv.org/pdf/1302.7281v1.pdf (2013)

[8] Addere Consultores. *Capacities and Opportunities for Industry and Academy in Activities related to or derived from Astronomy and Large Astronomical Observatories in Chile*. Chilean Ministry of Economy. http://www.economia.gob.cl/wp-content/uploads/2012/06/OPORTUNIDADES-ASTRONOMIA-EN-CHILE-INFORME-FINAL.pdf (in Spanish) (2012)

[9] Borne K. Virtual Observatories, Data Mining, and Astroinformatics. In H.E. Bond (ed.), *Planets, Stars and Stellar Systems*, volume 2, pages 409-443. Springer, Dordrecht, (2013).

[10] Hanisch R., Quinn P. The International Virtual Observatory. Retrieved from http://www.ivoa.net/about/TheIVOA.pdf, (2003).

[11] European Southern Observatory (ESO)/Government of Chile Joint Committee for the development of Astronomy. 10 Years Exploring the Universe. http://www.eso.org/public/archives/books/pdfsm/book_0040.pdf, (2005)

[12] GAVO. German Astrophysical Virtual Observatory. http://www.g-vo.org/pmwiki/ Documentation/Documentation (2013).

[13] Tody, D., Plante, R., and Harrison, P., “Simple image access specification version 1.0,” IVOA Recommendation 20091111 (2004).

[14] Williams, R., Hanisch, R., Szalay, A., and Plante, R., “Simple cone search version 1.03,” IVOA Data Access Layer WG Reccommendation (2008).

[15] Tody, D. et al., “Simple spectral access protocol,” http://www.ivoa.net/Documents/REC/DAL/SSA-20080201.pdf. IVOA Standards (2008).

[16] Dowler, P. et al., “Table access protocol version 1.0,” IVOA Recommendation, March (2010).

[17] Ochsenbein, F. et al., “IVOA recommendation: VoTable format definition version 1.2,” arXiv preprint arXiv:1110.0524 (2011).