Remote Sensing of Environment xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

The Australian Geoscience Data Cube — Foundations and lessons learned

Adam Lewis [a,*], Simon Oliver [a], Leo Lymburner [a], Ben Evans [b], Lesley Wyborn [b], Norman Mueller [a], Gregory Raevksi [a], Jeremy Hooke [a], Rob Woodcock [c], Joshua Sixsmith [a], Wenjun Wu [a], Peter Tan [a], Fuqin Li [a], Brian Killough [d], Stuart Minchin [a], Dale Roberts [a,b], Damien Ayers [a], Biswajit Bala [a], John Dwyer [e], Arnold Dekker [c], Trevor Dhu [a], Andrew Hicks [a], Alex Ip [a], Matt Purss [a], Clare Richards [b], Stephen Sagar [a], Claire Trenham [b], Peter Wang [c], Lan-Wei Wang [a]

[a] *Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia*
[b] *Australian National University, Canberra, ACT 2601, Australia*
[c] *Commonwealth Scientific Industrial Research Organisation (CSIRO), GPO Box 1700, Canberra, ACT 2600, Australia*
[d] *NASA Langley Research Center, 1 Nasa Dr, Hampton, VA 23666, United States*
[e] *U.S. Geological Survey Earth Resources Observation and Science (EROS) Center, 47914 252nd Street, Sioux Falls, SD 57198, United States*

ARTICLE INFO

Article history:
Received 27 July 2016
Received in revised form 7 February 2017
Accepted 12 March 2017
Available online xxxx

Keywords:
Landsat
Time-series
Big data
Data cube
High performance computing
High performance data
Collection management
Geometric correction
Pixel quality
Australian Geoscience Data Cube

ABSTRACT

The Australian Geoscience Data Cube (AGDC) aims to realise the full potential of Earth observation data holdings by addressing the Big Data challenges of volume, velocity, and variety that otherwise limit the usefulness of Earth observation data. There have been several iterations and AGDC version 2 is a major advance on previous work. The foundations and core components of the AGDC are: (1) data preparation, including geometric and radiometric corrections to Earth observation data to produce standardised surface reflectance measurements that support time-series analysis, and collection management systems which track the provenance of each Data Cube product and formalise re-processing decisions; (2) the software environment used to manage and interact with the data; and (3) the supporting high performance computing environment provided by the Australian National Computational Infrastructure (NCI).

A growing number of examples demonstrate that our data cube approach allows analysts to extract rich new information from Earth observation time series, including through new methods that draw on the full spatial and temporal coverage of the Earth observation archives. To enable easy-uptake of the AGDC, and to facilitate future cooperative development, our code is developed under an open-source, Apache License, Version 2.0. This open-source approach is enabling other organisations, including the Committee on Earth Observing Satellites (CEOS), to explore the use of similar data cubes in developing countries.

1. Introduction: A vision for the Australian Geoscience Data Cube

In this paper we describe the Australian Geoscience Data Cube (AGDC), a 'Big Data' infrastructure that aims to realise the full potential of Earth observation data holdings for Australia. The AGDC is a collaborative initiative of Geoscience Australia, the National Computational Infrastructure (NCI), and the Australian Commonwealth Scientific Industrial Research Organisation (CSIRO). The AGDC was developed over several years as we sought to maximise the impact of Land surface image archives that dated from Australia's first participation in the Landsat program in 1979. The AGDC is demonstrating, through a growing number of example applications, an unprecedented ability to leverage Earth observations to support research and operational users. The AGDC is also demonstrating an architecture that can allow the remote sensing community to improve access to, manage and take full advantage of increasing volumes of Earth observation data.

The characteristics of the AGDC are high quality calibration of satellite observations, the use of basic measurements (notably standardised surface reflectance for optical data), accurate geo-location, quality assessment and pixel level quality flags, structuring of data to facilitate analysis including time-series analysis, and the use of scientific file formats to facilitate efficient computation and exploratory data analysis. These characteristics allow automated workflows to be developed which produce continental-scale products drawing on the full time-series of available data.

The AGDC vision is of a 'Digital Earth' (Craglia et al., 2012), composed of observations of the Earth's oceans, surface and subsurface taken through space and time and stored in a high performance computing environment. A fully developed AGDC would allow governments, scientists and the public to monitor, analyse and project the state of the Earth, and will realise the full value of large Earth observation datasets by

* Corresponding author.
E-mail address: Adam.Lewis@ga.gov.au (A. Lewis).

http://dx.doi.org/10.1016/j.rse.2017.03.015
0034-4257/© 2017 Published by Elsevier Inc.

allowing rapid and repeatable continental-scale analyses of Earth properties through space and time.

This paper builds upon the initial concept and results that were developed as AGDC 'version one' (AGDCv1) and described in Lewis et al. (2016). It next presents the fundamental changes that we have made since then, responding to user feedback, computational challenges and lessons learnt, to build AGDC 'version two' (AGDCv2). In this paper we:

1. Describe the challenges that called for radical changes to traditional models of processing and analysis of Earth observation data, leading us to build the AGDC;
2. Define the foundations and core elements of the AGDC, including our approach to data preparation and collection management, the software environment used to manage and interact with the data, and the supporting high performance computing/high performance data environment;
3. Outline some examples that, in our view, demonstrate the benefits of our approach through the ability to rapidly complete analyses and produce products that draw on tens or hundreds of thousands of images without manual intervention;
4. Discuss the international uptake and potential relevance of the AGDC; and
5. Summarise general learnings from our AGDC journey and outline some important future directions.

2. The challenge

The Earth and its systems are complex and interconnected and gaining a comprehensive understanding of these systems and how we interact with them is of critical national and international interest (Boyd and Crawford, 2012; Frankel and Reid, 2008; Hart and Saunders, 1997; Lynch, 2008).

Although many sensors are used to acquire data about the Earth and its systems, constellations of Earth observation satellites (EOS) may be the single most important source thereof (Australian Academy of Science, 2009). Satellites capture observations of diverse physical phenomena such as crustal deformation, soil mineralogy, vegetation and surface water conditions, sea surface temperature, and magnetic and gravimetric field intensities, over the entire globe and for over long periods of time (decades). Of these constellations, the Landsat Program, operated by the United States Geological Survey (USGS) and NASA, represents the longest, most continuous, openly available Earth observation program in the world (Arvidson et al., 2001). Landsat missions have acquired moderate resolution multispectral data for over 40 years. Systematic analysis of these archives provides the capability to compare the changes that are now occurring on the Earth's surface with those that have occurred in the past.

However, extracting information from EOS data holdings poses an enormous technological challenge due to the volume and variety of the data. Satellite Earth observation data collections are 'Big Data'. The Oxford Dictionary defines 'Big Data' as "Datasets that are too large and complex to manipulate or interrogate with standard methods or tools". The term was first described by Laney (2001) as "the three V's: Volume, Velocity and Variety". Others have extended this list to include validity, veracity, value, and visibility. Historical sensor archives represent a 'volume and variety' challenge, however the next generation of meteorological sensors now available (such as the GOES-R Advanced Baseline Imager (ABI), EUMETSAT (Met-10), and Himawari-8 Advanced Himawari Imager (AHI)) mean that Earth observation data is now facing a 'velocity' challenge too. Satellite Earth observation data are often highly structured and stored as large to very large binary data files, each of which may contain gigabytes or even terabytes of data. This 'Big Earth Data' is massive, diverse and growing at an increasing rate (Overpeck et al., 2011).

Users of satellite Earth observation data therefore face challenges of volume, as archives of data grow; of variety, as instruments produce finer resolution observations that must be related to existing archives to produce an on-going and consistent record; and of velocity, as the intervals between observations reduce from weeks to days, or from hours to minutes in the case of Geostationary satellites. According to the international Committee on Earth Observation Satellites (CEOS) the number of Earth observation satellites grew from 12 in 1980 to over 69 operational EOS missions in 2014, with 137 missions planned for launch between 2015 and 2030 (CEOS, 2014). Considering only a handful of these missions, e.g., Landsat-8, Sentinel-1, -2, -3, and Himawari-8/9, we should expect at least a 20-fold increase in the global volumes of raw data acquired over a 15 year period. Our calculations of data growth for Australia are shown in Fig. 1. Adding processed data products derived from this raw data will lead to three to five times greater data volumes, consistent with the assessment of Overpeck et al. (2011).

New approaches are required for the management, analysis and distribution of Earth observation data and products in order to meet these big data challenges, which are compounded by growing expectations that Earth observation data streams will support the delivery of operational products as well as the on-going development of scientific knowledge (Mattmann, 2013). This challenge has been recognised for some time (Jensen, 1986; Schott, 2007; Schowengerdt, 2006). Furthermore new Earth observation data analysis methods are emerging which demand real-time access to full archives of moderate resolution Earth observation data, these include:

- The application of time-series analysis techniques to accurately detect change (Griffiths et al., 2014; Kennedy et al., 2010; Masek et al., 2013; Zhu and Woodcock, 2014);
- The systematic characterisation of a particular cover type across multiple decades (Masek et al., 2008; Mueller et al., 2016; Sexton et al., 2013);
- The use of 'best available pixel' composites to overcome the challenges posed by heavily cloud affected settings (Hermosilla et al., 2015; Thompson et al., 2015; Zald et al., 2016); and
- The use of time series of Earth observation data and derived products as Essential Climate Variables (ECV's) (Hollmann et al., 2013) and Climate Data Records (Kim et al., 2015).

The challenge of manipulating continental/global archives of moderate resolution Earth observation data to enable these analyses is

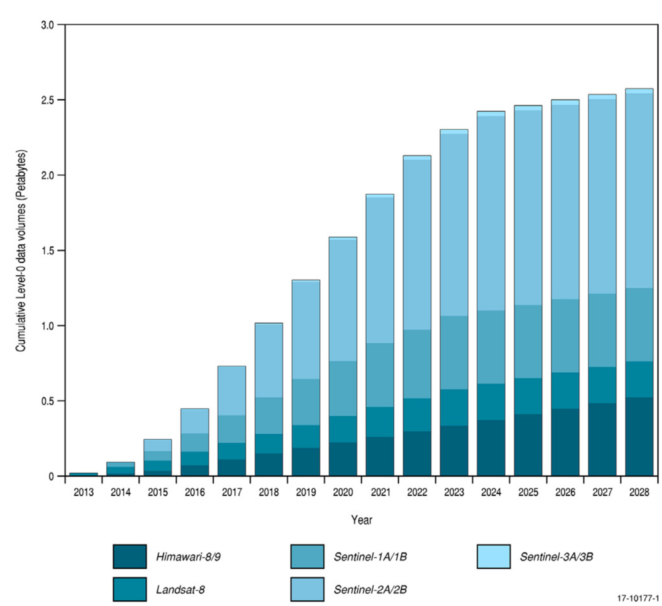

Fig. 1. The estimated volumes EOS data produced by the Landsat-8, Sentinel-1,-2,-3 and Himawari-8/9 missions from 2014 and 2029 for Australia. Only 'raw' data are considered. Data volume estimates are based on CEOS (2014).

Please cite this article as: Lewis, A., et al., The Australian Geoscience Data Cube — Foundations and lessons learned, Remote Sensing of Environment (2017), http://dx.doi.org/10.1016/j.rse.2017.03.015