Geophysical Research Abstracts
Vol. 17, EGU2015-14909-2, 2015
EGU General Assembly 2015

Something old, something new: data warehousing in the digital age

Rob Maguire and Andrew Woolf
Bureau of Meteorology, Canberra ACT, Australia ({r.maguire,a.woolf}@bom.gov.au)

The implications of digital transformation for Earth science data managers are significant: big data, internet of things, new sources of third-party observations. This at a time when many are struggling to deal with half a century of legacy data infrastructure since the International Geophysical Year. While data management best practice has evolved over this time, large-scale migration activities are rare, with processes and applications instead built up around a plethora of different technologies and approaches. It is perhaps more important than ever, before embarking on major investments in new technologies, to consider the benefits first of 'catching up' with mature best-practice.

Data warehousing, as an architectural formalism, was developed in the 1990s as a response to the growing challenges in corporate environments of assembling, integrating, and quality controlling large amounts of data from multiple sources and for multiple purposes. A layered architecture separates transactional data, integration and staging areas, the warehouse itself, and analytical 'data marts', with optimised ETL (Extract, Transform, Load) processes used to promote data through the layers. The data warehouse, together with associated techniques of 'master data management' and 'business intelligence', provides a classic foundation for 'enterprise information management' ("an integrative discipline for structuring, describing and governing information assets across organizational and technological boundaries to improve efficiency, promote transparency and enable business insight", Gartner).

The Australian Bureau of Meteorology, like most Earth-science agencies, maintains a large amount of observation data in a variety of systems and architectures. These data assets evolve over decades, usually for operational, rather than information management, reasons. Consequently there can be inconsistency in architectures and technologies. We describe our experience with two major data assets: the Australian Water Resource Information System (AWRIS) and the Australian Data Archive for Meteorology (ADAM). These maintain the national archive of hydrological and climate data. We are undertaking a migration of AWRIS from a 'software-centric' system to a 'data-centric' warehouse, with significant benefits in performance, scalability, and maintainability. As well, the architecture supports the use of conventional BI tools for product development and visualisation. We have also experimented with a warehouse ETL replacement for custom tsunameter ingest code in ADAM, with considerable success.

Our experience suggests that there is benefit to be gained through adoption by science agencies of professional IT best practice that is mature in industry but may have been overlooked by scientific information practitioners. In the case of data warehousing, the practice requires a change of perspective from a focus on code development to a focus on data. It will continue to be relevant in the 'digital age' as vendors increasingly support integrated warehousing and 'big data' platforms.