APPLYING SENSOR WEB STRATEGIES TO BIG DATA EARTH OBSERVATIONS

Terence L. van Zyl and Graeme Mcferren

Earth Observation Science and Information Technology
CSIR, South Africa

1. INTRODUCTION

Earth observation data and meta-data are a central concern of the earth sciences. These data are generated by a myriad of both in-situ and remote sensors. Other sources of data include computational simulations, various ex-situ sources such as environmental sampling campaigns and emerging trends such as crowd sourcing.

The Big Data phenomenon is one that has always existed in the earth observation community due to the large swaths of homogeneous data grids produced by for instance earth observation satellites. These data grids are continuously increasing in volume as spatial resolution and number of observed variables increases. The Sensor Web adds to this volume challenge by adding two additional V's namely variety and velocity. Variety talks to the heterogeneity of the data available from the Sensor Web whilst velocity refers to the high temporal resolution of the data and the need for near-real time processing in order for the produced information to be relevant.

The Big Data phenomenon compounds the already existing technological challenges associated with these earth observation data sets. These challenges include data discovery, data access, data exploration, data pre-processing, data analytics and data presentation. What has not been explored by the research community[1] is how those Big-Data accessible to the end user through the Sensor Web are to be used effectively to achieve exploration, pre-processing and integration, processing and modelling and eventually presentation and visualisation. Added to this lack of know how is the additional uncertainty linked to the exact relationship between Big Data and Service oriented Architectures[2]. Some questions arise as a result [1]:

- Does the current set of Sensor Web service standards allow us to deal with the issues of data discovery effectively?
- Does the web service, remote method invocation paradigm allow for effective access to Big Data?
- How do we integrate and pre-process two or more big data sets via the Sensor Web?
- Given Big Data how can we explore it using our current Sensor Web standards?
- What is the most effective way of performing analytics and data mining on big data, and does Sensor Web rise to the challenge?

The answer to these questions and a number of others is a resounding "We don't know!". What is clear is that our current attempts are not a prefect fit. There are however some strategies that allow one to deal with the challenges of Big Data within the Sensor Web which provide some insight into where we should be focusing our research.

In the following section we explore these strategies as have been attempted in some of the projects we are currently working on, and then finally we draw some conclusions.

2. STRATEGIES FOR EARTH OBSERVATION BIG DATA AND THE SENSOR WEB

It is important when considering these strategies to consider Sensor Web in terms of its function rather than the form it takes with respect to some set of standards or architecture. In this respect we see the Sensor Web as an infrastructure that supports an integrated system of sensor systems and provides access to sensors, sensor networks, and the corresponding observational datasets and meta-data. The main function of the world wide web is to provide a mechanism to organise information and the means for people to access that information. The Sensor Web takes the same approach in that its intention is to provide a mechanism for organising sensor data and the means to access that sensor data [2] .

It is important to realise that one size does not fit all and that different strategies are required for dealing with each of the challenges of Volume, Variety and Velocity.

2.1. Data oriented web service standards

One strategy that has proved effective is a move to web service standards with a primary focus on the underlying data and the mechanism for sub-setting, filtering, aggregating and

[1] A search for the terms "Big Data" and "Sensor Web" on google scholar returns 63 results.

[2] A search for the terms "Big Data" and "Service Oriented Architecture" in the title on google scholar returns 0 results.

transporting large quantities of data over limited network resources. This data orientation is in contrast to other web service which have a primary focus on either functionality or on the sensor itself [3]. When dealing with large volumes of data web service interfaces such as OpenDAP allow for effective sub-setting and pre-specified aggregation. In addition the compact binary format provided by NetCDF [3] over OpenDAP [4] is efficient and completely compatible across heterogeneous platforms. NetCDF and OpenDAP are not without problems as the focus here is on data cubes and raster style data and does not lend itself well to hierarchical vector data. In the instance where the data is of the vector variety especially more complex geometries then the options are still predominantly OGC Sensor Observation Service or SQL. In fact SQL provides the richest interface to your data holdings over the web, with strong subs-setting, aggregation and transformation capabilities but is often not an option made available to users at large.

2.2. Mobile code

The current approach to web based data processing, preprocessing and exploration in the Sensor Web is primarily takes one of two forms. In the first instance the data is moved over the web to the location of the specific web processing service where the functionality is preformed before the results are made available to be moved to wherever they are required. This approach has one immediate drawback when considering Big Data in that even after filtering moving a large amount data is not practical.

The second mechanism overcomes the challenges of the first by placing said web processing services at the same physical location as the data. However this approach still leaves one other large challenge that of statically defined and predefined functionality. If a user wishes to develop a new algorithm or defined some non-standard aggregation, the web services approach leaves little room for this type of ad-hoc processing and querying of the data.

One solution that we have found to be highly effective in this regard is mobile code [4]. Probably the most broadly and effective known usage of mobile code is SQL in which ad-hoc queries may be submitted to a data store where they are processed and the results returned. However SQL has its limitations in that it is a declarative and as such does not allow for the easy expression of new algorithms and functionality. In this instance the use of mobile code such as python scripts through for instance RPyC can be highly effective in a similar vein to the approaches taken by the distributed map reduce architectures such as Hadoop [5] [5].

[3] http://www.unidata.ucar.edu/software/netcdf/

[4] http://www.opendap.org/

[5] http://hadoop.apache.org/

2.3. Message Oriented Middleware

Effectively dealing with generating information and events off high velocity data streams is better suited to an entirely different architecture than the web service oriented ones presented. An example of such an architecture that is primarily event based is actor oriented architecture or message oriented architecture with Advanced Message Queuing Protocol (AMQP) [6] as the prevailing open standard. Here the use of message queues and geospatial filtering over the web allows for high throughput systems that scale well and produce results in near real time [6].

3. CONCLUSIONS

Although the focus of the Sensor Web has been somewhat limited to a single architectural view in the form of web services and service oriented architectures our experience has shown in a number of projects that this is not always the most effective solution, especially when deal with Big Data. The correct approach is then to hold to the overall vision of the Sensor Web as a way of gaining access to and organising sensors and sensor data and to use the appropriate architectural patterns and strategies in overcoming the challenges presented. Using these other approaches is not necessarily conflict with an open systems view nor is it non web centric.

4. REFERENCES

[1] K Chiu, M Govindaraju, and R Bramley, "Investigating the limits of SOAP performance for scientific computing," *In Proceedings of HPDC-11*, vol. pages, pp. 246–254, 2002.

[2] T.L. van Zyl, Ingo Simonis, and Graeme McFerren, "The Sensor Web: systems of sensor systems," *International Journal of Digital Earth*, vol. 2, no. 1, pp. 16–30, 2009.

[3] L Di, K Moe, and TL van Zyl, "Earth observation sensor web: An overview," *Selected Topics in Applied Earth . . .* , vol. 3, no. 4, pp. 415–417, 2010.

[4] Peter Baumann, "Web-enabled raster gis services for large image and map databases," in *Database and Expert Systems Applications, 2001. Proceedings. 12th International Workshop on*. IEEE, 2001, pp. 870–874.

[5] TL Van Zyl, G McFerren, and A Vahed, "Earth observation scientific workflows in a distributed computing environment," in *FOSS4G*, 2011.

[6] Graeme McFerren and Derick Swanepoel, "Towards a wide area alerting and notification system," .

[6] http://www.amqp.org/