# Adding Big Data's Velocity to Oceanographic Data Systems

**Adam Leadbetter,** Marine Institute (Ireland), adam.leadbetter@marine.ie
**Damian Smyth**, Marine Institute (Ireland), damian.smyth@marine.ie
**Rob Fuller**, Marine Institute (Ireland), rob.fuller@marine.ie
**Eoin O'Grady**, Marine Institute (Ireland), eoin.ogrady@marine.ie

## Introduction

In August 2015, a new seafloor observatory was depolyed in Galway Bay, Ireland. This sensors on the observatory platform are connected by fibre optic cable to a shore station, and from there a broadband connection allows data to be delivered to the Marine Institute's data centre.

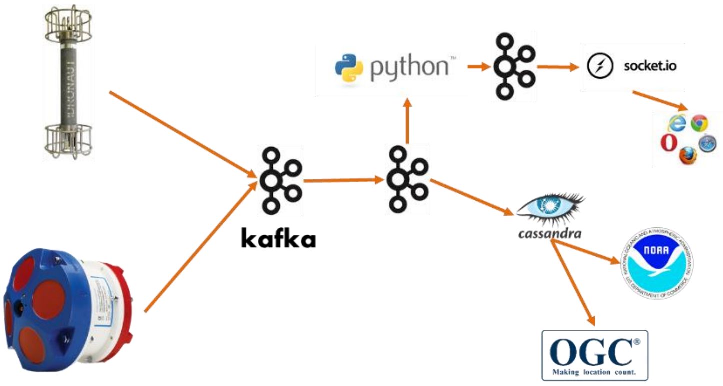

Fig. 1: Streaming data system **architecture overview**

This setup allowed the development of a new data acquisition system which takes advantage of the open source streaming data solutions which have been developed in response to the Big Data paradigm, in particular the "Velocity" aspect. Big Data is commonly defined in the following terms and we also expand on their application to marine data:

- "*Volume*": Big Data deals with data volumes which cannot be easily managed on a personal computer, stretching away from the megabyte and gigabyte scales towards the terabyte and petabyte scales,
- "*Velocity*": The speed at which data are acquired, processed and made available is key to Big Data and for oceanographic data represents a move from batch, delayed mode processing through near real-time processing into genuine real-time data availablilty,
- "*Variety*": There is a range of data types and sources in a Big Data system, as is typical in an oceanographic data system, and incorporates data types which move beyond those easily incorporated into database tables such as images and video,
- "*Veracity*": The accuracy of the data is important in Big Data applications, as it is in oceanographic data systems as can be seen by the various flagging schemes available (IOC 2013),
- "*Variability*": The context of a data point is important in interpreting it within the Big Data paradigm. This is also important in an oceanographic context when integrating data from models, remote sensing systems and *in-situ* observations.

## The Streaming Data System for Galway Bay

The streaming data system is an analytic computing platform that is focused on speed. This is because these applications require a continuous stream of often unstructured data to be processed. Therefore, data is continuously analysed and transformed in memory before it is stored. Streaming data applications typically manage a lot of data and process it at a high rate of speed. Because of the amount of data, it is typically managed in a highly distributed clustered environment. This satisfies the real-time data delivery requirement of the Big Data paradigm.

There are many open source software components which allow the deployment of such streaming data systems which have been developed by web companies such as LinkedIn. The Galway Bay cable observatory data system makes extensive use of the Apache Kafka messaging queue to relay data messages in a managed way from shore station to data centre and between processing scripts and data store (see Figure 1). Stream processing software, in particular the Apache Storm software, was investigated but is not mature enough to deploy in a production environemnt. Finally, a time-series optimised Big Data datastore, Apache Cassandra, has been deployed for archiving data.

### Evolving Data Standards

Access to the outputs of the Galway Bay observatry data system is provided visually via dashboards; via a file store archive; through an ERDDAP interface for downloading subsets of the data; and using standardised interfaces for programmatic access – namely the Open Geospatial Consortium's Sensor Observation Service and the ISO standard MQTT messaging protocol.

The Sensor Observation Service has been developed as a component of the streaming data system and interacts directly with the Cassandra database, which also provides a backend to the ERDDAP interface. Mindful of the "Born Connected" approach (Leadbetter *et al.* 2016), the fields of the Observations & Measurements records delivered by the Sensor Observation Service use URLs from the NERC Vocabulary Server where possible. Similarly, the emergence of new JSON-encodings for the Observations & Measurements data model (Cox & Taylor 2015) has also allowed a more consumer-friendly serialisation of the results to be produced.

In order to allow true streaming of the data into client software, a MQTT broker has been deployed. MQTT is an important piece of the Internet of Things architecture as it is currently the only accepted standard for machine-to-machine communications in that domain. The Galway Bay observatory implementation required development of an Apache Kafka to MQTT bridge which has been incorporated into the main software branch and exposes the message queue using a standard protocol and in more a secure manner than simply exposing the Apache Kafka server. Future developments within the MQTT broker will define standardised representations of the data within the messages exposed, and further the Born Connected concept.

### Conclusions

Big Data, streaming data and the Internet of Things are all currently highly important topics in software engineering at enterprise scale. However, they have not been tranlsated well to the Earth and Space Science Informatics realm through the lack of ability to connect them with the pre-existing, well-defined and highly-adopted data standards. In this paper, we have shown that the connection between the two paradigms is possible and that there is further work to do in the standardisation of data messages in the Internet of Things domain. This final challenge may be the precursor to connecting many more instruments much more quickly to the well established data systems of the oceanographic data management world.

### References

Cox, S., Taylor, P. (2015). OM-JSON - a JSON implementation of O&M. Presentation to the OGC Sensor Web Enablemenr Domain Working Group, Nottingham, UK, September.

Intergovernmental Oceanographic Commission of UNESCO. 2013. Ocean Data Standards, Vol.3: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. (IOC Manuals and Guides, 54, Vol. 3.) 12 pp. (English.) (IOC/2013/MG/54-3)

Leadbetter, A., Cheatham, M., Shepherd, A., Thomas, R. (2016) Linked Ocean Data 2.0, in Diviacco,P., Leadbetter, A., Glaves, H., Oceanographic and Marine Cross-Domain Data Management for Sustainable Development, Hershey, PA: IGI Global.