International Journal of Spatial Data Infrastructures Research, 2016, Vol.11, 09-19

Big Data – a step change for SDI?*

Chrisa Tsinaraki and Sven Schade

European Commission, Joint Research Centre, Institute for Environment and Sustainability
{chrysi.tsinaraki; sven.schade}@jrc.ec.europa.eu

Abstract

The globally hyped notion of Big Data has increasingly influenced scientific and technical debates about the handling and management of geospatial information. Accordingly, we see a need to recall what has happened over the past years, to present the recent Big Data landscape from an infrastructural perspective and to outline the major implications for the SDI community. We primarily conclude that it would be too simple and naïve to consider only the technological aspects that are underpinning geospatial (web) services. Instead, we request SDI researchers, engineers, providers and consumers to develop new methodologies and capacities for dealing with (geo)spatial information as part of broader knowledge infrastructures.

Keywords: Big Data, SDI, infrastructures, technology, databases, hype

INTRODUCTION

Spatial Data Infrastructures (SDIs) have been recently challenged by Big Data. The notion emerged when volume, velocity and variety were mentioned for the first time as the three dimensional challenges in data management by Doug Laney (2001). This was followed by Tim O'Reilly's (2005) publication on Web 2.0 and more intensively by James P. Collins's (2010) postulation of data-intensive

DOI: 10.2902/1725-0463.2016.11.art2

scientific discovery. As a response to these milestones, the term “Big Data” has been featured in hundreds of SDI and geodata-handling conference papers and journal articles; and discussions on the impacts of Big Data on SDIs, Digital Earth and Geographic Information Systems (GIS) have been held on each continent. Having been the subject of many heated debates, we considered that now is the right time to dedicate an Editorial of the International Journal of Spatial Data Infrastructure Research (IJSDIR) to a brief stocktaking at the technical and infrastructure level of Big Data; and to reflect on its actual impact on the SDI community.

GENERAL OBSERVATIONS

While Big Data now seems a widely accepted term, it is worthwhile mentioning that – in spite of the growing volumes of data made available every day from sensors, social media sources, Web logs, medical histories, etc. – Big Data was completely removed from Gartner’s hype cycle for emerging technologies in 2015 (Vorhies, 2015). This occurred at the same time as the introduction of terms that describe the Big Data landscape in both generic and abstract ways. They include, for example, the **data lake**: “If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples” (Dixon, 2010).

Nonetheless, as a notion, Big Data has become popular as it satisfies the requirements that led to its initial appearance – including, for example, the need of the scientific research community to unify the traditionally separated disciplines of high-performance computing and analytics (Reed and Dongarra, 2015). It also became clear that the Big Data challenges have to be resolved at the service level, leading to a “Big Service” ecosystem (Xu et al., 2015). Where the “five Vs” (**volume**, **variety**, **velocity**, **veracity** and **value**) are frequently used to characterise Big Data, seven main features represent Big Services: **massiveness, heterogeneity, complexity, convergence, customer focus, credibility and value** (BDVA, 2016).

Simultaneously, the way that humans are involved in the data life-cycle has changed drastically during the past ten years, where roles such as developers, business analysts and end-users, used to be clear. Nowadays, however, the same person may play multiple roles and some roles may be served by multiple actors, as seen in crowdsourcing (Howe, 2006). According to the Beckman report (Abadi et al., 2016), for example, humans can be assigned different roles in the Big Data era and they can be classified as **data producers, data curators, data consumers, and community members**. The socio-organisational dimension to Big Data should, as with any data infrastructure, not be overlooked.

THE CURRENT BIG DATA LANDSCAPE

The Big Data landscape, including both the *Big Data Technologies* (i.e. data lakes) and *Big Data Infrastructures* (i.e. databases and analytical tools) is evolving quickly. An overview of the main players and disciplines is, for example, maintained at the Big Data Landscape website (Feinleib, 2016). Due to the scope of this journal, we focus on its infrastructural aspects.

The Hadoop ecosystem

The Big Data technology level is undoubtedly dominated by the **Hadoop** ecosystem (ApacheTM, 2016), which is currently the only viable example of a data lake. The heart of the Hadoop ecosystem is **Apache Hadoop**, a Java based open source software library that supports the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to the thousands of machines that offer local computation and storage. A whole ecosystem of infrastructures for Big Data analysis has been developed around Apache Hadoop, including infrastructures for cluster and data management (e.g. Ambari, Avro, Chukwa, Pig, ZooKeeper, Mesos, etc.), database and data warehousing infrastructures (e.g. Cassandra, HBase, Hive, Drill, etc.), and analytical infrastructures (e.g. MapReduce, Tez, Spark, Flink, Mahout, S4, Samza, Storm, Kafka, etc.). Notably, in order to join the Hadoop ecosystem, some commercial tools have become open source and are undergoing the incubation process in the Apache Software Foundation (ASF), such as the In-Memory Data Fabric which is now included in the ecosystem as Apache Ignite.

Databases

As far as databases and data management are concerned, research and practice indicate that the traditional architecture of relational Structured Query Language (SQL)-based Data Base Management System (DBMS) hardly satisfies Big Data requirements (Pokorny, 2013). The various approaches that have been adopted in order to satisfy different subsets of these requirements include:

- **NoSQL databases** to provide a mechanism for data storage and retrieval not based on the tabular relations of relational databases. NoSQL is interpreted as ***N***ot ***o***nly ***SQL***, in order to emphasize that a NoSQL database may also support an SQL-like query language. The performance of NoSQL databases scales with increasing read-write workloads, but they compromise part of the guarantees in respect to Atomicity, Consistency, Isolation and Durability (ACID) (Rafique et al., 2015). Different types of NoSQL databases have been developed in order

to address competing requirements (NoSQL, 2016). The most common NoSQL database types include *key-value stores*, *column databases*, *document stores, XML databases*, *graph databases* and *array databases*. Some of these are equipped with particular spatial-processing capabilities. Examples include the column store *MonetDB* (used for point clouds), the document store *MongoDB* (used for spatial objects), and the array databases *rasdaman* and *SciDB* (both used for spatial fields).

- **NewSQL databases** to modernize relational database management systems in order to reach the scalable performance of NoSQL approaches (especially for On-Line Transaction Processing – OLTP), while preserving the ACID guarantees of a traditional SQL-based DBMS (Aslett, 2011). NewSQL systems vary greatly in their internal architectures but they all support the relational data model and use SQL as their primary interface.
- **NoDB approaches** to introduce a new generation of data management systems (Alagiannis et al., 2015). The NoDB approach eliminates major bottlenecks of current state-of-the-art technology to make database systems more accessible to the user, while still maintaining the features of modern DBMSs. To achieve this, traditional database architectures are extended to perform query processing in situ on the data, depending on the demands of a particular request.

Data analytics

The analytics supported by Big Data infrastructures can be divided into three types (Bertolucci, 2013): (a) ***Descriptive Analytics***, which apply descriptive statistics in order to gain insight from historical data, such as time series; (b) ***Predictive Analytics****,* a form of data mining which involves extracting information from data and using it to predict trends and behaviour patterns; and (c) ***Prescriptive Analytics***, which is a current trend involving the automatic synthesis of Big Data with multiple disciplines of mathematical sciences and computational sciences, and business rules, to make predictions and then suggest decision options taking advantage of these predictions.

This typology is accompanied by three major strategic approaches (Shneiderman and Plaisant, 2015): (i) **Extraction Strategies**, which use only the relevant records, event types, or events from a large dataset, thereby making detection of meaningful patterns easier; (ii) **Folding Strategies**, which replace a single long data sequence with many shorter ones so that cyclic patterns, such as weekend days in a week, are easier to recognize; and (iii) **Pattern Simplification Strategies**, which re-label similar structural types, in order to cope with the

variety of (event) types and to reduce the volume of structures that often makes it difficult to see meaningful patterns.

Not surprisingly, supporting tools are (again) primarily situated within the Hadoop ecosystem. Initial solutions for Big Data analytics used the **MapReduce** programming model (Dean and Ghemawat, 2008), which essentially applies batch processing. A MapReduce ***job*** usually splits an input data-set into independent chunks that are processed by the ***map tasks*** in a completely parallel manner. The MapReduce framework sorts the outputs of the maps, which are then inputs to the ***reduce tasks***. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executing the failed tasks. Several extensions of MapReduce have been developed, including ***SpatialHadoop*** (2013), a fully-fledged MapReduce framework with native support for spatial data (Eldawy and Mokbel, 2013), and ***Hadoop-GIS*** (Aji et al., 2013), a scalable and high performance spatial data warehousing system for running large-scale spatial queries on Hadoop.

Specialised real-time analytics provide the capacity to handle continuously generated **data streams** from application logs, events or a large pool of devices that deliver data in high velocity (Thaploo, 2014). Such **stream processing** applies ***streaming analytics***, i.e. the continuous calculation of mathematical or statistical analytics on-the-fly within the data stream. Stream processing solutions are designed to handle high volumes in real-time with a scalable, highly available and fault tolerant architecture (Wähner, 2014). The new requirements imposed by the need for stream processing and stream analytics have overthrown MapReduce in favour of **Spark** (2016), since MapReduce supports only batch parallel processing, while Spark is an up-to-100-times-faster generic engine for large-scale parallel data processing that also allows the integration of data streams (Goth, 2015).

An important development is the availability of open-source machine learning software, including **MLlib scalable machine learning library** of Apache Spark, **Apache Samoa** (2014) (initially developed by Yahoo!) and Google **TensorFlow** (2015).

Considering the other side of the spectrum: programming languages for the Big Data ecosystem, the **Scalable Language (Scala)** (2016) has become highly popular. Scala is object-functional, and its functional nature makes it easier to write safe and performant multi-threaded code.

IMPLICATIONS

The implications from these generic observations and evolving landscape on the SDI community are twofold. On the one hand, one might argue that many of the technical evolutions reside in back-end technologies; and that existing (web) service solutions already cover parts of the challenges related to data volumes, real-time access and the integration of heterogeneous sources. For example, (1) the Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) suite of standards (SWE, 2016) is equipped to deliver real-time information and trigger events; (2) the OGC Web Coverage Processing Services (WCPS) (2016) provides the means to extract value-added information out of massive raster or point cloud databases, moving processing capacities close to the data store; and (3) the INSPIRE geospatial interoperability standards (INSPIRE, 2016) provide a flagship for integrating public sector information from a variety of domains, administrative levels and cultural borders. In this way – while continuously working on technological improvements of the infrastructures and platforms that underlie geospatial services by including parts of the Hadoop ecosystem or developing specialised geospatial solutions – SDIs continue to ease data use and integration from remote sensing images (especially from the Sentinel satellites of Copernicus), point clouds, cartographic databases, in situ sensor networks (now also including contributions from Citizen Scientists) and much more.

To this respect, the (simplified and naïve) technological challenge of the SDI-community would be the addition of semi- and unstructured data sets and data streams into the existing data-handling structures and serving them as part of the established architectures, already including cloud storage and processing. This would particularly address *ad hoc* integration of sources with formats that are not known *a priori*. Such activities address new and much more dynamic work areas, i.e. go far beyond the current business cases. New methodological approaches have to be researched in order to cover the extended suite of possible analytics and the diverse implementation strategies. These methodologies also have to take into account the new roles of humans as part of the data life-cycle, thereby supporting arising notions such as “deep-learning” (Bengio et al., 2013) and “edge intelligence” (Serrano et al., 2015) for the SDI community in itself.

On the other hand, however, the growing user base for spatial data that is not well served with the existing technologies, and the overall request paradigm is changing. Many of the traditional SDI users know relatively well what they are looking for, and can deal with the typical standards of, for example, OGC, ISO TC211, and CEN TC287. This will no longer be applicable to the new generation of potential customers. They demand “all data that is available for an area X”, “all that is available about an object Y” or “everywhere that is related to this Z”. In addition, they rarely have a background in geomatics or in any particular field of

space, earth, environmental or geographical information science. In other words: WE NEED NEW INTERFACES!

Several years ago, it became clear that the Linked Data paradigm (Bizer et al. 2009) would not lead to the next-generation of SDIs (Schade and Smits, 2012) – as a silo community in itself. While, for example, INSPIRE already breaks the barriers between the SDI community and eGovernment (ISA 2015; ISA 2016), the Big Data era now urges us even more to simplify our service models and adopt existing solutions to more conventional approaches. Spatial cannot be treated as special any more. We are all challenged to contribute to the capacity building of new analytical approaches and in the sense of building geospatial capabilities into the newly emerging “Big Services” of mainstream IT. The required solution has to build on interconnected technical infrastructures. In addition, it should equally rely on the appropriate training and education of a new generation of service providers and consumers.

Simultaneously, while updating to the latest technology and continuing to connect to the mainstream, we must be mindful of the evolving trajectories, where the future of Big Data is difficult to predict. If Big Data is a transient problem and data grows slower than the hardware that allows us to cope with it, then the Big Data of today will be in our pockets tomorrow. If, however, we see a data oversupply from an infrastructure point of view, we could expect research to prosper or disruptive forces completely reshaping the computing landscape (Lin, 2015). How likely is it that the concept of an SDI will find a prominent place in this possible future?

CONCLUSION

So, is Big Data a step change for the SDI community? At least, it forces us to reflect on its impact and our role. While affecting underlying technical infrastructure and system connectivity, the new data providers, data extractors and knowledge sharers should not be underestimated, along with an opportunity to re-think our overall approach to SDI.

Researchers have to develop new integrative methodological approaches. Standardisation communities and geospatial software engineers should ease the use of the current services of SDIs for mainstream application developers and to enable new user experiences. Industry and academia have to adopt their educational approaches accordingly. And the research community is additionally challenged to anticipate future developments and explore promising new scientific frontiers.

We hope, therefore, to see some of these reflections and experiments on a new way of dealing with (geo)spatial information as part of knowledge infrastructures in the future issues of the journal.

ACKNOWLEDGEMENTS

The content presented in this editorial was largely influenced by discussions that the authors had over the past years with many colleagues working on SDI, data management, knowledge extraction, Open Science and many other fields. We are particular grateful for the numerous inspiring talks with our colleages from the Digital Earth and Reference Data Unit of the European Commission's Joint Research Centre in Ispra, Italy. Special thanks go to our colleage Robin Smith who helped polishing the final version of this editorial.

REFERENCES

Abadi, D., Agrawal, R., Ailamaki, A. et al. (2016). The Beckman report on database research, Communications of the ACM 59, 2 (January 2016), pp: 92-99. DOI: dx.doi.org/10.1145/2845915

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J. (2013). Hadoop GIS: a high performance spatial data warehousing system over mapreduce, in: Proceedings of VLDB Endow. 6, 11 (August 2013), pp: 1009-1020.

Alagiannis, I., Borovica-Gajic, R., Branco, M., et al. (2015). NoDB: Efficient Query Execution on Raw Data Files, Communications of the ACM, Research Highlights, 2015.

Apache Samoa (2014) Official Web page of Apache SAMOA, at https://samoa.incubator.apache.org/ [accessed 7 March 2016]

ApacheTM (2016) Official Web page of Apache™ Hadoop®, at http://hadoop.apache.org/ [accessed 7 March 2016]

Aslett, M. (2011). How will the database incumbents respond to NoSQL and NewSQL? 451 TechDealmaker, at http://cs.brown.edu/courses/cs227/archives/2012/papers/newsql/aslett-newsql.pdf [accessed 7 March 2016]

BDVA (2016). Official Web page of the Big Data Value Association, at http://www.bdva.eu/ [accessed 7 March 2016]

Bengio, Y., Courville, A., Vincent, P. (2013). Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8), pp:1798–1828. DOI: 10.1109/tpami.2013.50

Bertolucci, J. (2013). Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive, InformationWeek, at http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279 [accessed 7 March 2016]

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data – The Story So Far, International Journal on Semantic Web and Information Systems 5(3), pp1-22. DOI: 10.4018/jswis.2009081901

Collins, J.P. (2010). Sailing on an Ocean of 0s and 1s, Science 327(5972), pp: 1455-1456. DOI: 10.1126/science.1186123

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters, Communications of the ACM 51, 1 (January 2008), pp: 107-113.

Dixon, J. (2010). Pentaho, Hadoop, and Data Lakes. Post at James Dixon's Blog, at https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-anddata-lakes/ [accessed 7 March 2016]

Eldawy, A. and Mokbel, M. (2013). A demonstration of SpatialHadoop: an efficient mapreduce framework for spatial data, in: Proceedings of VLDB Endow. 6, 12 (August 2013), pp: 1230-1233.

Feinleib, D. (2016). Web page of the Big Data Landscape, at http://www.bigdatalandscape.com/ [accessed 7 March 2016]

Goth, G. (2015). Bringing big data to the big tent, Communications of the ACM 58(7), pp. 17-19. DOI: 10.1145/2771299

Howe, J. (2006). The Rise of Crowdsourcing, Wired magazine article, June 2006, at http://www.wired.com/2006/06/crowds/ [accessed 7 April 2016]

INSPIRE (2016). Official Web page of the Infrastructure for Spatial Information in the European Community, at http://inspire.ec.europa.eu/ [accessed 7 March 2016]

ISA (2015). Official Web page of ARE3NA - A Re-usable INSPIRE Reference Platform. Interoperability Solutions for European Public Administrations

(ISA) Action 1.17, at http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-17action_en.htm [accessed 11 March 2016]

ISA (2016). Official Web page of EULF - Establishment of a European Union Location Framework. Interoperability Solutions for European Public Administrations (ISA) Action 2.13, at http://ec.europa.eu/isa/actions/02-interoperability-architecture/2-13action_en.htm [accessed 11 March 2016]

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group File 949, at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [accessed 7 March 2016]

Lin, J. (2015). Is Big Data a Transient Problem? IEEE Internet Computing 19(5), pp. 86-90. DOI: 10.1109/MIC.2015.97

NoSQL (2016). Web page listing NoSQL data bases, at http://nosql-database.org/ [accessed 7 March 2016]

O'Reilly, T. (2005). What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software, at http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1 [accessed 7 March 2016]

Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. IJWIS 9(1), pp: 69-82.

Rafique, A., Van Landuyt, D., Lagaisse, B., Joosen, W. (2015). On the Performance Impact of Data Access Middleware for NoSQL Data Stores, IEEE Transactions on Cloud Computing PP(99), pp: 1-14. DOI: 10.1109/TCC.2015.2511756

Reed, D.A. and Dongarra, J. (2015). Exascale computing and big data, Communications of the ACM 58(7), pp. 56-68. DOI: 10.1145/2699414

Scala (2016). Official Web page of the Scala programming language, at http://www.scala-lang.org/ [accessed 7 March 2016]

Schade, S. and P. Smits (2012). Why Linked Data Should Not Lead to Next Generation SDI, in: Proceedings of IGARSS 2012, Munich, Germany.

Serrano, M., Hoan Nguyen Mau, Q.; Le Phuoc, D., et al. (2015). Defining the stack for service delivery models and interoperability in the Internet of Things: A practical case with OpenIoT-VDK, IEEE Journal of selected

areas in communications 33(4), pp: 676-689. DOI: 10.1109/JSAC.2015.2393491

Shneiderman, B. and Plaisant, C. (2015). Sharpening Analytic Focus to Cope with Big Data Volume and Variet,. IEEE Computer Graphics and Applications 35(3), pp.10-14. DOI: 10.1109/MCG.2015.64

Spark (2016) Official Web page of Apache Spark, at http://spark.apache.org/ [accessed 7 March 2016]

SpatialHadoop (2013) Official Web page of SpatialHadoop, at http://spatialhadoop.cs.umn.edu/ [accessed 7 March 2016]

SWE (2015) Official Web page of the OGC Sensor Web Enablement suite of standards, at http://www.opengeospatial.org/ogc/markets-technologies/swe [accessed 7 March 2016]

TensorFlow (2015) Google blog post on TensorFlow, 9 November 2015, at http://googleresearch.blogspot.it/2015/11/tensorflow-googles-latest-machine_9.html [accessed 7 March 2016]

Thaploo, V. (2014). Hadoop Gives way to Real Time Big Data Stream Processing – The 3 Key Attributes – Collect | Process | Analyze, Blazeclan blog, at http://blog.blazeclan.com/hadoop-real-time-big-data-stream-processing-3-key-attributes-collect-process-analyze/ [accessed 7 March 2016]

Wähner, K. (2014). Real-Time Stream Processing as Game Changer in a Big Data World with Hadoop and Data Warehouse, InfoQ, 10 September 2014

Vorhies, W. (2015). Big Data Falls Off the Hype Cycle, Data Science Central blog post, at http://www.datasciencecentral.com/profiles/blogs/big-data-falls-off-the-hype-cycle [accessed 7 March 2016]

WCPS (2016). Official Web page of the OGC Web Coverage Processing Service standard, at http://www.opengeospatial.org/standards/wcps [accessed 7 March 2016]

Xu, X., Sheng, Q.Z., Zhang, L., et al. (2015). From Big Data to Big Service, IEEE Computer 48(7), pp. 80-83. DOI: 10.1109/MC.2015.182