Dipping in Data Lakes

written by Jeremy Huggett | 15/07/2019

We're becoming increasingly accustomed to talk of Big Data in archaeology and at the same time beginning to see the resurgence of Artificial Intelligence in the shape of machine learning. And we've spent the last 20 years or so assembling mountains of data in digital repositories which are becoming big data resources for mining in the pursuit of machine learning training data. At the same time we are increasingly aware of the restrictions that those same repositories impose upon us – the use of pre-cooked 'what/where/when' queries, the need to (re)structure data in order to integrate different data sources and suppliers, and their largely siloed nature which limits cross-repository connections, for example. More generally, we are accustomed to the need to organise our data in specific ways in order to fit the structures imposed by database management systems, or indeed, to fit our data into the structures predefined by archaeological recording systems, both of which shape subsequent analysis. But what if it doesn't need to be this way?

This is the promise offered by proponents of data lakes. These are characterised as massive stores of unstructured, semi-structured, and structured data stored in their original native format, as 'raw' data, which make few or no assumptions about the kinds of questions or analyses which will be required, or indeed whether the data are useful, if not now then maybe sometime in the future. Effectively the data lake model shifts the decision-making surrounding data to the point at which they are used rather than stored; they consist of multiple data sources, accessible to many users for many undefined purposes, where the type of queries are not known in advance, where no particular data schema are defined or imposed (e.g. O'Leary 2014, table 1). The data lake employs a variety of tools largely unfamiliar to archaeologists: Hadoop, cloud services such as Amazon Web Services, Microsoft Azure or Google Cloud Platform, and NoSQL or MongoDB databases, for example (e.g. Mikhailouskaya 2018).

> "The data lake dream is of a place with data-centered architecture, where silos are minimized, and processing happens with little friction in a scalable, distributed environment. Applications are no longer islands, and exist within the data cloud, taking advantage of high bandwidth access to data and scalable computing resource. Data itself is no longer restrained by initial schema decisions, and can be exploited more freely by the enterprise." (Wilder-James 2014).

So a data lake is seen as addressing many of the perceived challenges of more traditional data warehousing, in particular through de-siloing data and connecting disparate data resources, and deferring data modelling and remodelling (e.g. Fang 2014; Kim 2018; Lave 2018) until the data are eventually combined and processed using big data methods (e.g. Fang 2014; Maroto nd). The concept behind a data lake recognises that we often don't know the value of the data we have, or the questions we might want to ask of it in the future, and capitalises on the cheap mass storage available rather than incurring up-front the high costs of ingest associated with structured data repositories (e.g. Fang 2014; Woods 2011). It's only fair to say that many data lake proponents represent vendors or consultancies promoting the benefits of data lakes.

Posey (2019) suggests three reasons for the data lake trend. First, that data lakes overcome the requirement that data are structured according to a specific schema and consequently can handle data that might not otherwise be retained or be usable. Secondly, that data lakes have to potential to contain as-yet unrealised information simply because they contain so much more data that previously might not be analysed or analysable. Thirdly, that data lakes handle data completely differently: rather than requiring data to be structured and organised prior to storage, the schema is created (or imposed) at the time the data are used in what is called 'late binding' or 'schema on read'. 'Early binding' is what we are used to, in which data are evaluated, their structure defined, and the data collected within a specific data model, whereas in 'late binding' the data are collected schema-free and loaded into the data lake leaving the data end-user responsible for defining the schema to be used at need depending on their research questions (e.g. Fang 2015, 821). These subsequent data transformations, updates and aggregations applied preparatory to analysis can then be captured as part of the data lifecycle within the data lake (Stein and Morrison 2014, 5).

Not for nothing are data lakes described as transformative, revolutionary, providing freedom from the imposition of rigid data models and the consequent loss of potentially valuable information. From an archaeological perspective, much of this sounds very attractive, offering the possibility of overcoming the shortcomings of data schemas, the loss of data that has no immediately identifiable analytical destiny, and recognising the importance of historical data, for example. Inevitably, however, all is not serene below the surface of the data lake.

Data lakes can quickly become massively chaotic with no easy way to making sense of the data and requiring considerable sophistication to navigate them: just accumulating data doesn't somehow reveal useful information in and of itself, nor does increasing the quantity of data improve its quality, despite some of the arguments surrounding big data. Further, amassing so-called 'raw' data strips away the interrelationships between data and their contexts. It also presumes that such a concept as 'raw' data can exist in the first place: that archaeological data exist 'out there', waiting to be discovered, rather than being created by an interpretive act of recognition. In such an environment, data are perceived as unprocessed, distinct from the subjectivities that created them and independent of their contexts of creation. This in part may lay behind the growing recognition that data lakes can quickly become data swamps (Gorelik 2019, 12) or data graveyards (Stein and Morrison 2014, 6), or data garbage dumps (Inmon 2014), storing useless, unused, and unusable data.

For data lakes to facilitate both the discoverability of data and the possibility of new analyses, their data – structured, semi-structured, and unstructured – has to be associated with metadata about the data and their contexts of creation, as well as paradata or metaprocess data which provide

information about the processing that the data have undergone, since no data are truly 'raw'. So the concept of 'late binding' to create datasets tailored to new enquiries is actually predicated upon a degree of 'early binding' as the data are ingested into the data lake. This doesn't mean that all the data conform to the same schema, but simply reflects that all data have been collected according to some kind of schema (recognised or not) and this needs to be captured alongside the data themselves. Nor is it the same as suggesting that all data have to be cleaned, harmonised, aligned and transformed upon ingest: this remains a task deferred to their point of use (with such subsequent processing also captured within the data lake). It does mean, however, that the kind of frictionless ingest associated with data lakes seems an increasingly improbable dream. Data lakes encapsulate a fundamental philosophical dislocation over the nature of data itself: the idea that data preparation, data cleansing, and data transformation tasks are eliminated in a data lake so as to store data in its rawest form (e.g. Pasupuleti and Purra 2015, 8) is predicated upon a specifically empiricist scientistic perspective.

Gorelik (2019, 13) talks of a 'logical data lake', where a virtual data lake layer connects multiple heterogenous systems: a hybrid approach which links a data lake with data warehouses and other traditional data suppliers. In some respects we have the potential to do this already through the Linked Open Data and OAI-PMH endpoints provided by archaeological repositories such as the Archaeology Data Service and Open Context, although these are perhaps closer to what Gorelik characterises as 'data puddles' rather than data lakes (2019, 5). So maybe we should see the opportunities for archaeology as dipping our toes into the data lake rather than plunging in headlong, since the idealised images associated with data lakes do not match up to the archaeological realities?

References

Fang, H. 2015 'Managing Data Lakes in Big Data Era', *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, 2015*, pp. 820-824. https://doi.org/10.1109/CYBER.2015.7288049

Gorelik, A. 2019 *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science* (O'Reilly: Sebastopol CA)

Inmon, B. 2016 *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump* (Technics Publications: Basking Ridge NJ).

Kim, D. 2018 'What's the Difference between Hadoop and a Data Lake', *Arcadia Data* July 10 2018 https://www.arcadiadata.com/blog/whats-the-difference-between-hadoop-and-a-data-lake/

Lave, M. 2018 'Data lakes in business intelligence: reporting from the trenches', *Procedia Computer Science* 138, 516-524 https://doi.org/10.1016/j.procs.2018.10.071

Maroto, C. nd 'A Data Lake Architecture with Hadoop and Open Source Search Engines: Using Enterprise Data Lakes for Modern Analytics and Business Intelligence', *Search & Analytics Insights* https://www.searchtechnologies.com/blog/search-data-lake-with-big-data

Mikhailouskaya, I. 2018 'Alternative approaches to implementing your data lake', *ScienceSoft* May

21 2018 https://www.scnsoft.com/blog/data-lake-implementation-approaches

O'Leary, D. 2014 'Embedding AI and Crowdsourcing in the Big Data Lake', *IEEE Intelligent Systems* 29 (5), 70-73. https://doi.org/10.1109/MIS.2014.82

Posey, B. 2019 'Use Data Lakes to Bet on the Future of Artificial Intelligence', *IT Pro Today* June 26 2019 https://www.itprotoday.com/storage/use-data-lakes-bet-future-artificial-intelligence

Pasupuleti, P. and Purra, B. 2015 *Data Lake Development with Big Data* (Packt: Birmingham).

Stein, B. and Morrison, A. 2014 'The enterprise data lake: Better integration and deeper analytics', *PwC Technology Forecast: Rethinking integration* 1, 1-9.

Wilder-James, E. 2014 'The Data Lake Dream', *Forbes Magazine* January 14 2014 https://www.forbes.com/sites/edddumbill/2014/01/14/the-data-lake-dream/

Woods, D. 2011 'Big Data Requires a Big, New Architecture', *Forbes Magazine* July 21 2011 https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/