Life and Death of Data in Data Lakes: Preserving Data Usability and Responsible Governance

Marzieh Derakhshannia, Carmen Gervet, Hicham Hajj-Hassan, Anne Laurent, Arnaud Martin

Life and Death of Data in Data Lakes: Preserving Data Usability and Responsible Governance

Marzieh Derakhshannia[1], Carmen Gervet[2], Hicham Hajj-Hassan[3][0000−0001−9917−4606], Anne Laurent[1][0000−0003−3708−6429], and Arnaud Martin[4]

[1] LIRMM, Univ Montpellier, CNRS, Montpellier, France
anne.laurent@umontpellier.fr,dm.derakhshannia@gmail.com
[2] Espace Dev, Univ Montpellier, IRD, Univ. La Runion, Univ de Guyane, Univ des Antilles, Montpellier, France
carmen.gervet@umontpellier.fr
[3] CNRS-L, Beirut, Lebanon
hishamhh@cnrs.edu.lb
[4] CEFE, Univ Montpellier, CNRS, Montpellier, France
arnaud.martin@umontpellier.fr

Abstract. Data crossing seeks the extraction of novel knowledge through correlations and dependencies among heterogeneous data, and is considered a key process in sustainable science to push back the current frontiers of knowledge, especially to address challenges such as the socio-economic impacts of climate change. To tackle such complex challenges, interdisciplinary approaches and data sharing methodologies are ubiquitous, with a strong focus on data openness and ensuring that the fair principles hold. Data lakes are data repositories, recently developed to store such big heterogeneous data that are then available for crossing and be exploited without a priori objectives regarding their usage (unlike data warehouses). Such data lakes can then be used to populate Open and Linked Open Data in a central location regardless of its source or format. In this context of no prior knowledge regarding its usage, it may be tempting to store and share all the available data. However, this comes with two main disadvantages: 1) overwhelming amount of data that could prevent end users from exploiting the data, 2) and environmental reasons (energy consumption of data storage). Moreover, data of poor quality may deserve the lake usability and be deleted. We thus claim in this position paper that a data life cycle must be designed so as to integrate data death for some of the data. The choice of the data to be stored regarding the ones to forget is then of crucial importance in data lakes. We propose here some first positions for this aspect of data governance.

Keywords: Data Lakes · Web of Data · Data Life Cycle and Data Governance · Sustainability.

1 Context

Data crossing is often considered as a key for discovering new knowledge and developing advanced models for phenomena understanding, as for instance climate change and its impacts on societal impacts For this reason, data sharing, especially data opening through Internet and the Web of data is a crucial challenge.

Data lakes [9] have been designed for allowing data retention before their usage. In such data repositories, data are made available from various data sources and can be exploited by data analysts to serve end users' needs with the idea that *the larger the data, the better the results.*

However in this paper, we state the conjecture that data perenity is no longer an option for a sustainable science despite the availability of massive data storage capacities. This raises the challenge of the lifespan of stored data, while maintaining its ability to generate the knowledge necessary for the users over time (consumers). How to make decisions regarding the lifetime and death of data when you only have the present knowledge of its usage, without further planning? How to ensure that the decisions taken regarding each data lifespan, with not entail a potential future loss of knowledge?

We consider mainly the case where the laws regarding the life cycle of the data is not imposed (ex: GDPR that forces the destruction or confidentiality of personal data).

We might think that a strong element to guide such decisions is to account for the data producer and consumers profiles. However, there is no guarantee that current behavior might dictate properly the 'forgetting function' definition. Thus in this position paper, we wish to distinguish both aspects, and not assume that they are necessarily interdependent over a certain time interval. This brings about the question of determining the criteria underlying the formal definition of the forgetting function (most probably time dependent), and thus the death concept.

2 Background

Data bases have been studied for the last 50 years from the very beginning of computer science. When dealing with data, their management is a key topic has been intensively addressed. The way they are collected, stored and exploited is managed through the so-called data governance process [7]. Regarding scientific data in particular, the concept of data life cycle[5], described by Figure 1 is often referred to as a key principle. In such a process, the question of deleting data and purging data bases has been discussed.

In particular, the question of big data management imposes to consider the infrastructures to be deployed. Some works have pointed out the necessity to build large data repositories in order to keep them sustainable [6].

[5] https://www.dataone.org/data-life-cycle

In the context of big data, data lakes are one of the emerging models to manage such huge repositories. In [9], data lakes have been defined as *a logical view of all data sources or data set, in their raw format, available and accessible by data scientist or statistician to find new insight.*

- *A data lake is governed by a metadata sources index to guarantee the data quality.*
- *A data lake is controlled by rules, tools and processes to guarantee the data governance.*
- *A data lake is limited to data scientist or data statistician access to guarantee data security, data privacy and compliance.*
- *A data lake access all type of data.*
- *A data lake has a logical and physical design.*

A data lake is a key element of the information architecture and a new step on its evolution.

Data lakes must be distinguished from data warehouses as data warehouses are rather dedicated to answer a priori known key performance indicators regarding several analysis dimensions, which impose a fixed multidimensional model, data sources and data preprocessing while data lakes are repositories for raw data whose further usages are not a priori known.

Indeed, [2] and [3] consider data lakes as a data storage management placement populated by all data sources in a raw format or as-is data. Data governance is pointed as being crucial and the information architecture is discussed. In the IBM redbook [5], the notion of data reservoir is associated to the data lake term and used as same meaning. The concept a metadata catalogue guarantees the respect of data governance to prevent the data lake to transform into a data swamp.

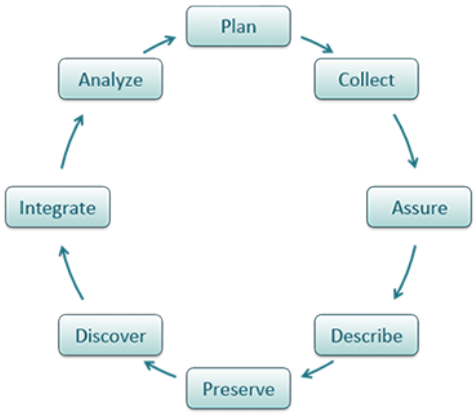

Fig. 1. Data Life Cycle.

Internet of things supports such approaches since sensors produce huge volumes of data that are not all consumed but that are usually considered as *potentially interesting*. It should be noted that data lakes store raw data, which excludes preprocessing. However, new models are emerging with fog computing that consider that some treatments can be embedded within the sensors.

3 Method

In this paper, we claim that some data must be deleted from (or not even put in) data lakes by the organization in charge of the lake. The questions are then:

- can some data be retained for a (given) period, never put into the data lake or even deleted from the lake at some period? For example, we may think about an analogy with the withdrawal of a medicinal product that threatens or could involve risk patients.
- when data is put into the data lake, to which spatial and temporal scale must the data be considered?

We report and discuss below some first research positions:

- providing solutions to extend naive data aggregation and sampling (e.g., choosing data distributions);
- providing decision support systems to help end users deciding on where to point the cursor between storing all data as raw data or preparing and sampling data;
- relying on and copying supply chain principles to decide what data and when putting into the data lake.

3.1 Forgetting Functions

When dealing with large data with the idea of saving storage memory consumption, it may be interesting to compress the data. This is even the case before storage when data are collected as sensors are designed for measuring at some time periods.

[1] has proposed so-called *forgetting functions* in data warehouses with two main principles, either sampling the data or aggregating.

We propose to extend this approach so as to be used in data lakes. The main challenge is then that it is difficult to evaluate the *utility* of the data regarding their usage as data lakes are not meant to serve *immediate* needs but rather store data in case it may be useful *in the future*.

Moreover, if the data are aggregated in order to be compressed, we claim that distributions must be used rather than simple functions as *mean, median, min, max*. This will allow to describe and model the uncertainty, imprecision and variability of the data. Formalisms as fuzzy logic or possibility distributions may be used for this purpose.

3.2 Providing Decision Support Systems

The goal here is to design and build decision support systems that will drive the management of available data. As previously mentioned, storing and managing all the data fed to the lake is antinomic with a sustainable science. Thus some data must either be removed from the lake or idled, compressed and left aside. The criteria that will guide the decision making need to be defined. A common approach can be to define criteria through usage, frequency, etc. We state that other properties of the data can be relevant to maintain a sustainable data lake, or data lakes, as a dynamic resource. Considering the features of interest is one venue that might hold over a certain period but might be reconsidered further on. Part of the research is to seek functions that will determine criteria for managing the life span of the data.

3.3 Imitating Natural Data Lakes

Natural lakes are ecosystems with many organisms which are subjugated (governed) by chance and necessity. Chance is gene mixing, mutations, etc. Life reproduces itself with a prolixity far superior to the level of acceptance of the system. There are therefore regulations that are carried out through the mechanism of natural selection (only the fittest ones survive and reproduce). Natural selection is the constraint imposed on living things. Chance does not produce information; it only produces complexity in the sense of Kolmogoroff [8]. Necessity is the one that produces information. In the nature, information is what has managed to evolve [4].

Complexity in a data lake can be illustrated by the addition of new data, the heterogeneity of the data but also by the product of data crossover (relationship search for example). The need is to retain the data that will provide useful information and eliminate data that will never be used. So we have to imagine a system that retains the potentially useful data.

To complicate the comparison, it must also be noted that nature sometimes keeps useless things (for example, there are DNA sequences that are never coded into proteins and therefore seem useless). So nature does not really delete all the data, it probably puts it to sleep. DNA that is not translated into protein can be altered (mutation) without, on the face of it, any consequence for the system (organism, ecosystem). The homeostasis of the system therefore depends on both a sorting by natural selection and a preservation of elements apparently not used.

This can be compared to data lakes (DNA being comparable to raw data) for which we thus claim for the necessity to delete data without eliminating everything.

3.4 Imitating Supply Chain Management

The supply chain is a set of integrated corporations and processes which is included sourcing the raw materials, manufacturing the products and transferring the finished products to the customers [11]. Supply chain management is a

manner to coordinate and organize all processes and activities according to the supply chain goals [12]. Since data lakes collect all types of data and transform them for the final user, we can consider data as a product and data lakes and data governance as supply chain management.

There are some supply chain strategies that reduce or eliminate the useless, non-valuable or destructive products or activities, as for example lean management and green supply chain. Lean management is beneficial strategy that improves quality and profit by emphasizing on wast reduction. The main purpose of lean strategy is to wipe out all processes and products which do not create value for the chain [10].

In recent years, successful supply chains do not only try to gain more profit, improve product quality, increase service level and reduce final price regardless of ecological affairs, but also they attempt to pay more attention to the environmental consideration and issues which influence on economic and social systems [13], the firms and companies are pressured from ecologists and government to reduce or eliminate the pollutant processes and materials throughout the supply chain which have destructive effects on the environment.

Green supply chain management has become popular and strategic to improve environmental performance by green procurement and production throughout the entire Stages of a products life cycle. The green vision encompasses all decision-making process of the supply chain, as for example Eco-design,considering environmental criteria for supplier evaluation,green production, green transportation,green purchasing and revers logistics (disposal) [14].

There are a lot of standards and criteria to build a green supply chain. According to these standards, all of the products and activities that have ecological and environmental standards remain in chain otherwise they do not have permission to be present in supply chain.

We can define the proper criteria and standards for the products that are allowed to enter in supply chain or remain in it up to the last stage of the chain and if the products or activities do not have appropriate standards and requirements they will not be allowed to enter into the chain and if they were entered, they would be removed form it.

Data in data lakes act like products in supply chain. Therefore, if some data are poor and do not meet the criterion will not get into the data lakes, from our point of view, data in data lake could be never put into or deleted from data lake when they threat the veracity of data lake just like the harmful or non-valuable products in green supply chain.

Acknowledgments

Supported by PHC CEDRE 42415YJ, French Ministry of European and Foreign Affairs (MEAE), French Ministry of Higher Education, Research and Innovation (MESRI) and Lebanese Ministry of Education and Higher Education (MEHE).

4 Conclusion

Efficient data governance could be a practical solution to preserve data lakes from huge amount of useless data. Therefore, we can easily answer to the questions of this issue, some data can be retrained for a period or be removed form the lake if they increase the risk of data lake's authenticity.

In this work, we propose preliminary directions based on comparisons with supply chain management and natural lakes on the one hand, and on forgetting functions and decision support systems on the other hand. Criteria and solutions will be developed in our future work and assessed on real data lakes from both scientific and business data.

References

1. Boly, A., Hébrail, G.: Forgetting data intelligently in data warehouses. In: 2007 IEEE International Conference on Research, Innovation and Vision for the Future in Computing & Communication Technologies, RIVF 2007, Hanoi, Vietnam, 5-9 March 2007. pp. 220–227. IEEE (2007). https://doi.org/10.1109/RIVF.2007.369160, https://doi.org/10.1109/RIVF.2007.369160
2. Fang, H.: Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In: International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). pp. 820–824. IEEE (2015)
3. Gartner: Gartner Says Beware of the Data Lake Fallacy. http://www.gartner.com/newsroom/id/2809117 (2014)
4. Gaucherel, C., Gouyon, P., Dessalles, J.L.: Information, The Hidden Side of Life. ISTE (2019)
5. IBM: Governing and Managing Big Data for Analytics and Decision Makers. http://www.redbooks.ibm.com/abstracts/redp5120.html?Open (2014)
6. Jones, N.: How to stop data centres from gobbling up the worlds electricity. Nature **561**, 163–166 (09 2018). https://doi.org/10.1038/d41586-018-06610-y
7. Khatri, V., Brown, C.V.: Designing data governance. Commun. ACM **53**(1), 148–152 (2010). https://doi.org/10.1145/1629175.1629210, https://doi.org/10.1145/1629175.1629210
8. Li, M., Vitnyi, P.M.: An Introduction to Kolmogorov Complexity and Its Applications. Springer Publishing Company, Incorporated, 3 edn. (2008)
9. Madera, C., Laurent, A.: The next information architecture evolution: the data lake wave. In: Chbeir, R., Agrawal, R., Biskri, I. (eds.) Proceedings of the 8th International Conference on Management of Digital EcoSystems, MEDES 2016, Biarritz, France, November 1-4, 2016. pp. 174–180. ACM (2016). https://doi.org/10.1145/3012071, http://dl.acm.org/citation.cfm?id=3012077
10. Martnez-Jurado, P.J., M.F.: Lean management, supply chain management and sustainability: a literature review. Journal of Cleaner Production **85**, 134–150 (2014), https://doi.org/10.1016/j.jclepro.2013.09.042
11. Mentzer J, Witt W.D, e.a.: Defining supply chain (sc)management. Journal of Business Logistics **22**(2) (2001), http://dx.doi.org/10.1002/j.2158-1592.2001.tb00001.x
12. Simchi-Levi, D., K.P.S.L.E.: Designing and Managing the supply chain Concepts, Strategies and Case studies. New York: McGraw-Hill Publishing (2003)

13. Zhu, Q., Sarkis, J.: Relationships between operational practices and performance among early adopters of green supply chain management practices in chinese manufacturing enterprises. Journal of Operations Management **22**, 256–289 (2004)
14. Zhu Q, Sarkis J, L.K.: Green supply chain management implications for closing the loop. J Transp Res Part E **44**, 1–18 (2008)