MOL2NET, International Conference Series on Multidisciplinary Sciences
MOL2NET 2021, International Conference on Multidisciplinary Sciences, 7th edition

# Data Lakes Technologies and It's Significant Implementations

*Ajit Singh*[a] *, Sultan Ahmad*[b]
[a] *Patna Women's College, Patna, India*
[b] *College of Computer Engineering and Sciences, Saudi Arabia.*

.

**Abstract:** This exploratory research of data lakes in big data times is a prominent topic for both academia and industry. One of the main motivations behind is that companies need to cope with more data than ever before, and the problems of how to analyze even how to store data are becoming more and more challenging in many industries. The occurrence of the concept of a data lake to meet such big data problems is enlightening and will most likely be considered in any relevant big data strategy. This idea is still on the way to prove itself out and inevitably it gives rise to much attention as well as much criticism. Luckily, more and more positive voices towards data lakes are emerging and give highly appreciation to the concept and even propose some workable and innovative suggestions to make improvement to the practical implementation. This study introduced basic background information of data lake implementation and can give valuable suggestions and insights to practitioners. After presenting and summarizing most of the popular implementation of data lakes from data professionals, three different approaches were introduced. All of these approaches have both advantages and disadvantages, and companies need to consider their own business needs and requirements to make a wise choice.

**Keywords:** Data Lake, Implementation, Hadooo, Data Virtualization

## Introduction

As big data is changing people's life in even every aspect more disruptively than ever before, companies are also inevitably getting more and more involved with big data challenges. They are experiencing pressure of handling various incredibly increasing amounts of data, unstructured and semi-structured, integration with legacy data, and the most important thing, with what their data means for business and how they can manage to use more efficiently and effectively.

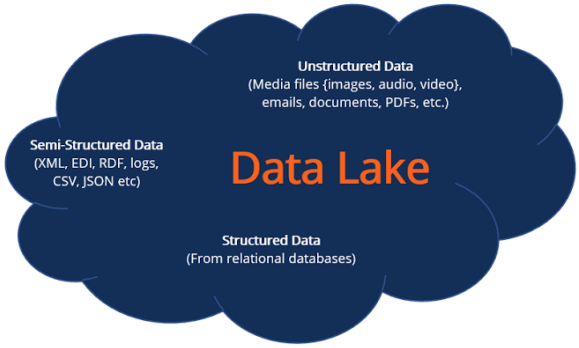


Figure 1: Data Lake Architecture

On the one hand, just as Mr. Bill Schmarzo, the CTO of EMC[2], mentioned in an interview[10] that, people are bringing all kinds of big data technologies into their companies and then just wait there for magic to happen. But the reality is almost negative for them. Companies tend to have a misunderstanding in the sense that new technologies stand for advantages, competence and value. However, as we can see from the survey results, although companies are indeed setting out to get prepared for big data challenges, the results do not necessarily go to what exactly they are expecting for.

On the other hand, numerous kinds of technologies are available for anyone to choose, with different features and advantages, whether free or commercial. Actually, some companies are suffering from too many options to choose from and are not sure if they can come up with a cost-effective plan or not. Currently, there are several vendors offering data lake – related commercial products and services, such as Pivotal. What's more, there is not yet any guideline available for data lakes practitioners to carry out a data lake on their own.

## Three Approaches

In this research, three approaches to get a data lake in an enterprise are proposed and briefly introduced, aiming at giving some hints for practitioners where to start to think if they want to bring in a data lake. The ideas of these three methods have been discussed with experts such as Dr. Florian Neukart.

Companies may implement a data lake,

- ✓ via data virtualization;
- ✓ completely depend on Hadoop;
- ✓ through combination of heterogeneous data sources either optimized for storing and processing unstructured data (document stores, key value stores) and structured data (traditional relational databases).

Data lakes are unique in a way that they can store and process both unstructured and well-structured data smoothly, unlike traditional database technologies.

### Data Virtualization

The basic idea of data virtualization is pulling together data without consolidating it in a central data warehouse physically. Instead, an abstract virtual data layer is created in order to connect distributed data from disparate sources as if it is stored in one central common place. Obviously, the original data remains where it is and there is no physical transport of data at all, while data is virtually connected together to some extent.

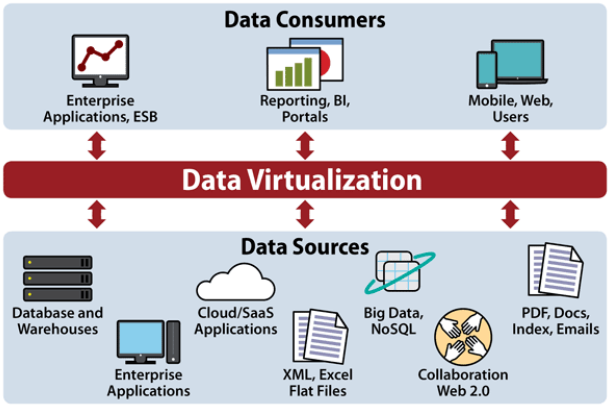


Figure 2: Data Virtualization

With data virtualization, companies now get possibilities to put all of their data, maybe all over the world, in one virtual place, acting like an enterprise data lake actually, with the help of some data virtualization tools and technologies. This method has several advantages.

As no physical data transport happens, there is high potential in saving costs, such as for servers, maintenance costs, licensing costs for additional data marts and DWHs, savings related to operations, etc. Additionally, the implementation is relatively painless and it can give organizations fast return on investment (ROI), compared with the two other approaches.

This method can avoid resistance of data owners handing over their treasure since companies have no need to ask them to “donate” their data.

It can achieve faster time to business intelligence report and information delivery, since researchers can have quick and easy access to data, via just one platform, without physically consolidating data, but abstracting data from disparate sources to get a full view.

There are concerns as well, that organizations should be aware of.
Performance problem. This is said to be the most significant pitfall. Nonetheless, it can be very much solved with throwing more technologies on it, together with proper tuning. Related technologies can be optimization techniques (refer to Section 2.6), in-memory computing. The rise of commodity servers can also

help to improve the performance of data virtualization.

Consistency in data across all the sources. Companies need to make sure that the different data that they want to access via data virtualization should be treated defined consistently. This is the first issue that should be settled down before using data virtualization techniques.

It's better to starting with piloting in small scale projects that companies can succeed on, and prove it out. If it can win success and then companies can continue to go from there and grow.

### All in Hadoop

Another option would be to build a data lake based on Hadoop, which means that the power of a Hadoop cluster in order to store data of all kinds, thus both structured and unstructured is leveraged .

This method involves moving data into one Hadoop system physically, including extracting metadata, loading, setting up new hardware, etc. With this approach, organizations can gradually have an enterprise data lake that built on the whole Hadoop ecosystem, together with its related vendors, providing many additional functionalities and capabilities. For example, Hadoop data lakes have a wide variety of data access approaches, like spanning batch, streaming, real-time and interactive, in-memory, etc.

Hadoop data lakes enable companies to "*store everything, analyze anything and build what you need*", as introduced in an online open course[32] for data lakes. It means that companies can store almost all kinds of data in its native form as well as full context of data and its usage lineage, which can definitely help companies to tap into more insights about customer behaviors and how to run business process more efficiently. Gaining more and more raw data can empower the business with the data insights required, so the business can build right applications upon data lakes, then bringing in more innovation and value, creating new and more data, pushing the data cycle to repeat itself. To some extent, data lakes accelerate the speed of this store-analyze-build cycle, via which companies can do lots of analytics, such as in-database analytics, in-memory analytics, massive parallel processing, etc.

Apart from those popular advantages showed in Section 2.4.2.4, Hadoop data lake also has some other unique features deserves attention. Firstly, it allows for different industries to have a data lake that has specific analytic applications tailored for its own data need. Different industries (e.g., healthcare, retail, telecommunications) even organizations may have different types of data (e.g., sensor, clickstream, geographic,

social, etc). Second, as Hadoop allows for distributed storage and easy accessibility, Hadoop data lakes are becoming more and more welcomed in organizations that increase their exposure to mobile and cloud-based applications, Internet of Things (IoT) .

### Combined-approach

There is another choice for companies that wish to have a Hadoop-based data lake works as a complement to their EDWs, which means that companies can store unstructured data in Hadoop system while remain well-structured data or other legacy data as where it is, whether stored in relational database or managed by other suitable storage technologies. As shown in the results of Q18 in the questionnaire, half of the respondents declared that their companies join data lakes with EDWs.

Nevertheless, this approach is not very handy during implementation and utilization, compared with previous two approaches. Companies that adopt this approach need to come up with feasible solutions to some problems, which mean, they have disadvantages to overcome. As pointed out in the questionnaire, these disadvantages are mainly related to privilege management, security and a consistent authorization concept. If data are not stored in one consistent system, business users may face different data access control issues. They cannot reach the data they want quickly.

## Business Implications

Bringing a data lake into an enterprise implicates much more than just technical issues. Unlike complicated issues, such as technological problems, which can gradually be solved by various approaches along with time, complex issues are ones that are human related and are more likely to remain the states what they are, and difficult to be solved along time passing, even may become worse and worse if no proper and effective remedial measure is taken. The same is true for developing a data lake in a data-driven company. The reasons are as following.

Firstly, the concept of data lakes actually calls for a new way to think about how should people treat company's data, whether viewing it as personal or departmental property or, instead, the treasure and value that belongs to the whole enterprise. This should entail a cultural change that requires people to show openness towards what they think they should have the right to possess, such as data or related professional competence, but may actually belong to the whole enterprise. When being asked to share information about what they are doing, how they are doing and what data they own, people are often reluctant to do so, due to being afraid of losing jobs or value of

their own in their organization. It is not surprising to see that the survey results also show this phenomenon, which revealing that companies are always encouraging their people to share data and knowledge but will never force them to do so. This organizational cultural change requires both time and efforts from the upper management and executives.

Secondly, due to the need to couple with big data challenges, it seems like that data lakes can be a nice choice to start with. However, as Bill Schmarzo points out that[3], IT people would better, firstly, convince the business side to cooperate with them, winning their support and understanding of what's going on with tackling big data, and then prove it out to the business guys with better business performance. In short, it is not that good for IT to play a lone hand in bringing in data lakes in an enterprise! Rather, IT should gain the business to back it up and achieve an alignment between them about big data counter strategy.

Thirdly, there is not yet any best practice of data lakes available for reference. Practitioners are trying out every different method to improve the whole ecosystem for data lakes. Although the concept of a data lake looks very much appealing, companies should never overlook its accompanying potential risks and pitfalls, at least till now, such as data governance, data security and legal issues, which can also be seen from the survey responses. Without good data governance, data lakes can easily end up being dirty and unusable. Satisfying data quality and data lakes performance are not that easy to achieve, unless good data governance is guaranteed. Just like the real natural lake, if there is no guarder to keep track of things like who fished in this lake, who poured what into this lake, how many fishers are there currently, what are the sources that stream into this lake, etc, then the lake will definitely end up being like dirty still waters.

For the same reason, data lakes is said be to the promising big "data warehouse" to hold all the company data, consolidated or raw data, so management work such as keeping track of who used lakes and how he used are more than crucial. Not only we need record data usage history but also control data access, achieving faster authorization time for required different datasets, while higher security level to sensitive data. Companies can try to have a separate department that takes over all the issues related to enterprise data lakes. People in this department have the authority to grant or deny data access to all requests. This process may require companies to give special training to their staffs about legal affairs and privilege management. Insuring security of the company data, both internal and external, is vital. For instance, customer data is for sure sensitive but information about staffs of a company is also

significantly private, like healthcare data. This separate data lake department should be armed with enough knowledge in such as law and regulations.
.A mature data lake takes time and “cultivation” (Brian Stein, Alan Morrison, 2014). A data lake will gradually mature as user interaction and data governance performance grows and gets better – the interaction that continually refines the data lake and the data discovery will make the lake mature. In conclusion, the idea of a data lake to be one single data repository for organizations to work more efficiently with their data can be a great solution to tackle the challenges brought by big data problems. Building a small data lake firstly and then filling it in with more and more raw data, together with what have been built already there in the lake by the users will make the data lake the most promising treasure and property of an enterprise in its near future.

## Conclusion

The topic of data lakes is not just a technical issue. Rather, it also has respect to corresponding vital business implications, from an organizational point of view. Data lakes demand openness from people towards data in a company. At present, people tend to view departmental or other sensitive data as their own possession and are often reluctant to share data with others. Besides, this new lake concept calls for a far more advanced data management, or say data governance methods. If data are well-structured or in small size of amounts, there is no problems with the conventional approaches at all. But once integrating all kinds of data in one big lake thing, different troubles are coming out. Notwithstanding creating a transparent alike atmosphere, in an enterprise, between users and data is awesome and enlightening, the security and legal issues related to data lakes still remains vital but troublesome, demanding more attention and efforts from upper management to make an organizational cultural change.

## References:

[1]. Barb Darrow (2013). “Pursuing big data utopia: What realtime interactive analytics could mean to you”. [Online] available: <https://gigaom.com/2013/03/21/pursuing-big-data-utopia-what-realtime-interactive-analytics-could-mean-to-you/>

[2]. Barry Devlin (2014). “Data lake muddies the waters on big data management”. [Online] available: <http://searchbusinessanalytics.techtarget.com/feature/Data-lake-muddies-the-waters-on-big-data-management>

[3]. Bill Inmon (1999). “Data Mart Does Not Equal Data Warehouse”. NOV 20, 1999.

[Online] available: <http://www.information-management.com/infodirect/19991120/16751.html?zkPrintable=1&nopagination=1>

[4]. Brian Stein, Alan Morrison (2014). “The enterprise data lake: Better integration and deeper analytics”. Technology Forecast: Rethinking integration Issue 1, 2014. [Online] available: <http://www.pwc.com/technologyforecast>

[5]. Dan Woods (2011). “Big Data Requires a Big, New Architecture”. [Online] available: <http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/2/>

[6]. Donald R. Cooper, Pamela S. Schindler (2008). “Business Research Methods”. In its Anniversary 10th Edition. Pp 370.

[7]. EMC$^2$ white paper 1 (2015). “Federation Business Data Lake – Enabling Comprehensive Data Services”.

[8]. Frank Lo (2015). “What is Hadoop? What is MapReduce? What is NoSQL?”. [Online] available: <https://datajobs.com/what-is-hadoop-and-nosql>

[9]. Gregory Chase (2014). “10 Amazing Things to Do With a Hadoop-Based Data Lake”. [Online] available: <http://blog.pivotal.io/big-data-pivotal/features/10-amazing-things-to-do-with-a-hadoop-based-data-lake>

[10]. Gwen Shapira (2011). “Hadoop and NoSQL Mythbusting”. [Online] available: <http://www.pythian.com/blog/hadoop-and-nosql-mythbusting/>

[11]. Singh, A . (2021). Data Science and Human Behaviour Interpretation and Transformation . Journal of Learning and Teaching in Digital Age , 6 (1) , 1-7 .

[12]. Singh, A. (2021). Communication Coroutines For Parallel Program Using DW26010 Many Core Processor, Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics , vol. 3, no. 1, pp. 15-20, Retrieved from DOI: https://doi.org/10.35882/ijeeemi.v3i1.3

[13]. Singh, AK (2020). Applications of IoT in Agricultural System. Acad. Res. J. Agri. Sci. Res. 8(5): 485-491

[14]. Singh, A. (2019). Data Publishing Techniques and Privacy Preserving. International Journal for Information Security Research (IJISR), Volume 9, Issue 3, pp 881-890.

[15]. Singh A (2019) Architecture of data lake. Int J Sci Res Comput Sci Eng Inf Technol 5(2):411-414.

[16]. Singh, Ajit, Enabling Researchers to Make Their Data Count (March 24, 2019). Available

at SSRN: http://dx.doi.org/10.2139/ssrn.3359161

[17]. Singh, Ajit, Best Approach to Read a Scientific Article (April 16, 2019). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3372784

[18]. Singh, Ajit, Implementation of the IoT and Cloud Technologies in Education System (May 3, 2019). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3382475

[19]. Singh, Ajit, Writing Research Proposal for MS/MPHIL/PhD Program (February 9, 2021). Available at SSRN: http://dx.doi.org/10.2139