INTERCATHEDRA

3(48), 2021, 151–157

http://www.up.poznan.pl/intercathedra/
pISSN 1640-3622
eISSN 2545-2045

http://dx.doi.org/10.17306/J.INTERCATHEDRA.2021.00130
Accepted for print: 11.07.2021

Ajit Singh✉

Patna Women's College, Bihar, India

METHODS FOR DATA LAKES IMPLEMENTATION

Abstract. The exploratory research of data lakes in the times of big data is a prominent topic for both academia and industry. One of the main motivations for the study is that companies need to cope with more data than ever before, and the problems associated with analyzing and even storing data are becoming more and more challenging in many industries. The emergence of the data lake concept to solve big data problems can be helpful in any significant big data strategy. Now, the use of lake data is being further tested and inevitably generates much attention as well as much criticism (Darrow, 2014). Luckily, more and more voices of appreciation towards data lakes are emerging. What is more, workable and innovative suggestions to make improvement to the practical implementation of this solution are proposed. This study introduced basic background information of data lake implementation and can give valuable suggestions and insights tousers. 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, making it imperative that companies consider their own business needs and requirements when implementing them.

Keyword: Data Lake, Hadooo, Data, Virtualization, Big Data, analytic

INTRODUCTION

As big data is changing people's life in 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 data, integration with legacy data, and the importance of data to the business and the method of its effective and efficient use.

On the one hand, just as Mr Bill Schmarzo, the CTO of EMC^2, mentioned in an interview (Hurwitz et al., 2019), 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 different. Companies tend to have a misunderstanding in the sense that new technologies stand for advantages, competence, and value. However, as one can see from the survey results, although companies are indeed setting out to get prepared for big data challenges, the results are not always in line with their expectations.

On the other hand, numerous kinds of technologies are available for anyone to choose, with different features and advantages, whether free or commercial. In fact, some companies suffer from having too many options to choose from and are not sure if they can come up with

✉Ajit Singh, Department of Computer Science, Oatna Women's College, Patna University, Bihar, India, e-mail: ajit_singh24@yahoo.com

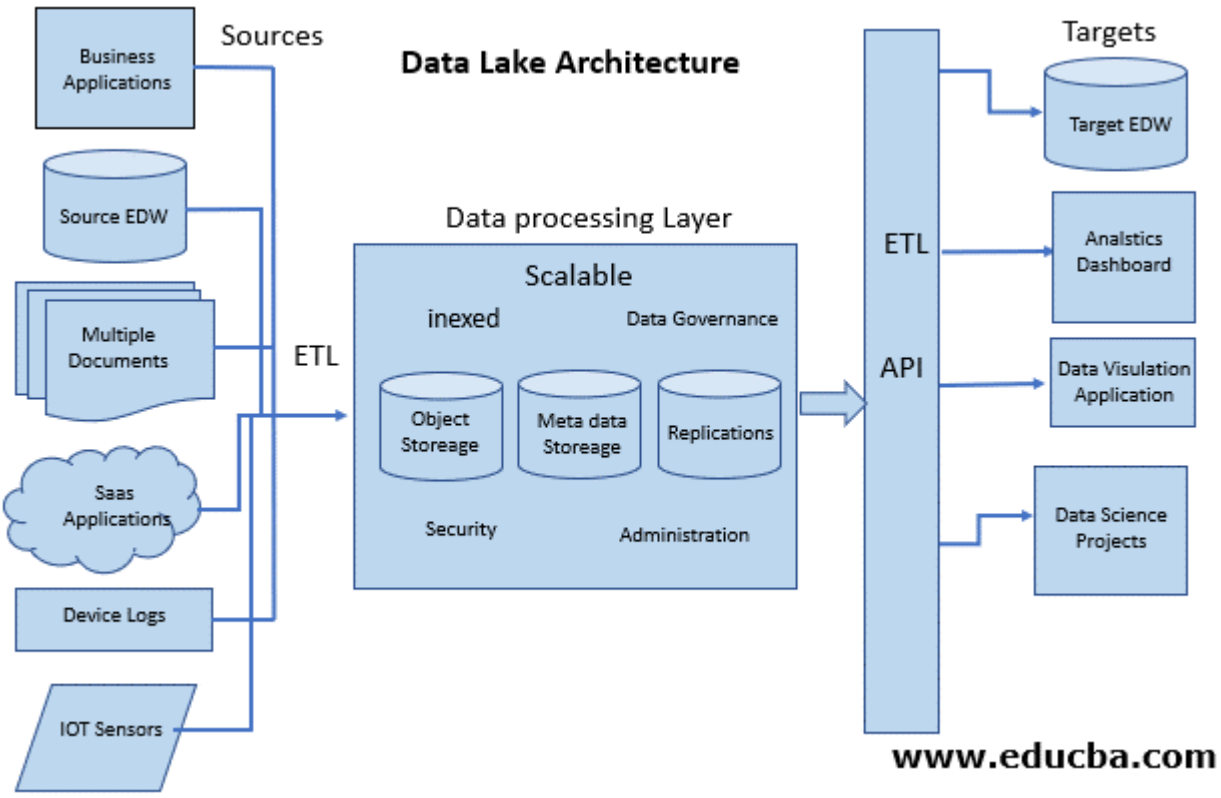

Fig. 1. Data Lake Architecture
Source: https://www.educba.com/data-lake-architecture/

a cost-effective plan or not. Currently, there are several vendors offering data lake related commercial products and services, such as Pivotal (Devlin, 2014). What is more, there is not yet any guideline available for data lakes users to carry out a data lake on their own.

The data lake concept was initially coined by the CTO of Pentaho, named James Dixon, on one of his blogs (EMC² white paper 1, 2011). Later, many information leaders and vendors raised a number of varied, but still consistent, meanings concerning the concept. On the other hand, there were also those who held negative attitudes and contrary opinions towards data lakes. Although there are gaps or even misunderstandings among the conveyed ideas and concepts, some consensuses, which are regarded as the dreams that data lakes can bring to a current business, to serve as a part of the company's counter strategy for the challenges stemmed from big data issues, can be found.

Given that the concept of a data lake may be considered new,the adaptation of a data lake for different industries may vary considerably and so far only a few companies are using it successfully, further research on business requirements, suggested architectures and possible implementation methods is needed (Stein and Morrison, 2014).

LITERATURE REVIEW

Carrying out a literature review on traditional data lake development approaches is aiming at helping readers better understand the significance and urgency for enterprises to make a change to tackle the big data challenges. Many companies are in the process of big data strategy transformation. Some of them are taking steps out to facilitate the transformation associated with plethora of big data technologies and solutions from all kinds of vendors, while some other are still making their minds to get prepared to take the initiatives to change.

This literature review aims to facilitate to make the aforementioned decisions by presenting deficiencies of

traditional data warehousing systems and what advantages that a data lake can bring to them.

MULTIVOCAL LITERATURE REVIEW

Since academic literature on data lakes is scarce, a new method of reviewing, called multivocal literature review (MLR), was selected as the review method to obtain basic research information. MLR is a method of reviewing literatures that are accessible on the Internet and are generally non-academic topics related (Devlin, 2014). This method is suitable for collecting preliminary requirements of data lakes and other contemporary topics, for instance data virtualization and unstructured data.

In this paper, the Google search engine has been used to collect data source for MLR. Keywords for querying included "data lakes", "data lakes requirements", "data lakes implementation", "unstructured data" and similar. However, other similar terms of data lakes, like enterprise data hubs and landing zone, are not included in this research, which can be considered as its limitation.

To conduct search concerning the part of concepts of data lakes and implementation requirements, keywords "data lakes" and "data lakes requirements" were used. As determined by the Google ranking algorithm, the sources were accessed from 12 April 2021 to 24 April 2021. After reading the title and introduction, sources that are identical to each other are discarded. Finally, 78 posts were gathered and analyzed as the input data for MLR process.

The results of MLR process consists of two parts. The first part is all the content concerning Scientific Fundamentals. The other part consists of a preliminary set of requirements for implementing and utilizing data lakes successfully established based on a comprehensive analysis, extraction and grouping of the sources found online. The set of requirements consists of features that might have significant impact on successful implementation and utilization of data lakes (Lo, 2016).

RESEARCH METHODOLOGY

As big data is changing every aspect of people's life 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 importance of data to the business and the method of its effective and efficient use.

On the one hand, just as Mr. Bill Schmarzo, the CTO of EMC[2], mentioned in an interview (Hurwitz et al., 2019), 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 different. Companies tend to have a misunderstanding in the sense that new technologies stand for advantages, competence and value. However, as one can see from the survey results, although companies are indeed setting out to get prepared for big data challenges, the results are not always in line with their expectations (Bleiberg and West, 2018).

On the other hand, numerous kinds of technologies are available for anyone to choose, with different features and advantages, whether free or commercial. In fact, some companies suffer from having 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 (Woods, 2011). What's more, there is not yet any guideline available for data lakes users to carry out a data lake on their own (Lawson, 2016).

This study presents and describes three approaches to implementing a data lake in the enterprise. The purpose of this study is to facilitate making the decision associated with the implementation of data lakes. Companies may implement a data lake:

- via data virtualization,
- using 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 that connects distributed data from disparate sources as if it is stored in one central common place is created. Obviously, the original data remains where it is and there

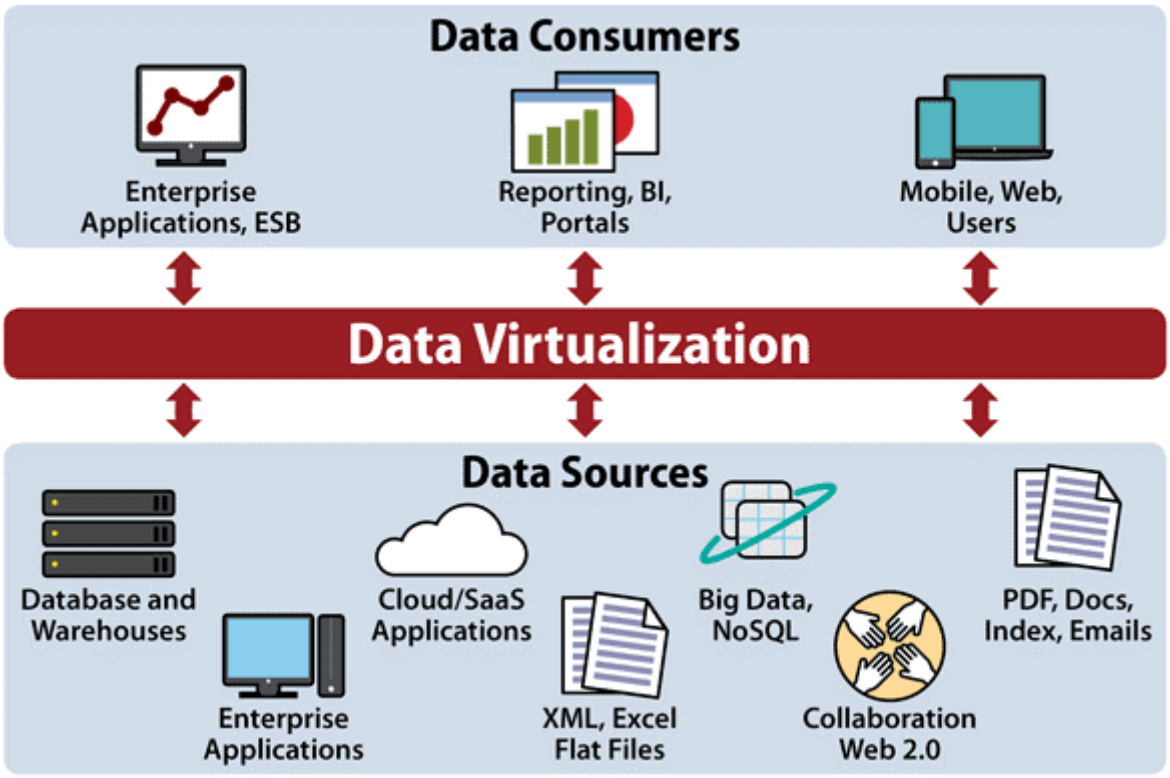

Fig. 2. Data Virtualization
Source: https://www.datamation.com/big-data/what-is-data-virtualization/

is no physical transport of data at all, while data is, to some extent, virtually connected (Cooper and Schindler, 2019).

With data virtualization, companies using some data virtualization tools and technologies have the possibility to store all of their data from all over the world in one virtual place acting like an enterprise data lake. This method has several advantages.

As no physical data transport happens, there is high potential in saving costs, e.g., for servers, maintenance costs, licensing costs for additional data marts and DWHs, savings related to operations, etc. Additionally, compared with the two other approaches, its implementation is relatively problem-free and it can provide fast return on investment (ROI).

This method can avoid resistance of data owners handing over their data since companies have no need to ask them to "release" their data.

This enables faster report and information delivery times because researchers can quickly and easily access data through a single platform, without physically consolidating data, but abstracting data from different sources to get the big picture. There are also concerns l that organizations should be aware of.

Performance problem. This is said to be the most significant problem. Nonetheless, it can be very much solved with using more technologies, together with proper tuning. Related technologies can involve optimization techniques, 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 defined consistently. This is the first issue that should be settled down before using data virtualization techniques.

It's better to start with piloting in small scale projects that companies can succeed on and test the method. If the approach proves beneficial, companies can continue its use and develop it.

USING 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 (Stein and Morrison, 2014).

This method involves moving data into one Hadoop system physically, including extracting metadata, loading, setting up new hardware, etc. Implementing this approach, organizations can gradually have an enterprise data lake that built on the whole Hadoop ecosystem, together with it's 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 (EMC² white paper 1, 2011).

As stated in an online open course for data lakes, Hadoop data lakes enable companies to "*store everything, analyze anything and build what you need*" (Stein and Morrison, 2014). 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 learn more about customer behaviors and how to run business process more efficiently (Lo, 2016). Gaining more and more raw data can provide the company with the data insights necessary for better use of data lakes, bringing in more innovation and value, creating new and more data and pushing the data cycle to repeat itself.

Hadoop data lake also has some other unique features that deserve attention. Firstly, it allows for different industries to have a data lake that has specific analytic applications tailored for their own needs. Different industries (e.g., healthcare, retail, telecommunications) and even organizations may possess different types of data (e.g., sensor, clickstream, geographic, social, etc) (Klein, 2014). 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) (Bleiberg and West, 2018).

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 where it is, whether stored in relational database or managed by other suitable storage technologies.

Nevertheless, this approach is not very convenient for implementation and use, compared with previous two approaches (Chase, 2017). 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 (Shapira, 2015).

BUSINESS IMPLICATIONS/CHALLENGES

Firstly, the concept of data lakes actually calls for a new way to think about how one should treat company's data, whether viewing it as personal or departmental property or, instead, 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 reveals that companies are always encouraging their employees 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 (Widom, 2019).

Secondly, due to the need to address the challenges connected with big data, it seems that data lakes can be a good choice to start with. However, as Bill Schmarzo points out (Inmon, 2016), 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. Users 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 the potential risks and pitfalls accompanying its implementation, e.g. data governance, data security and legal issues, which are also evidence by the survey responses. Without good data governance, data lakes can easily end up being unusable. Satisfying data quality and data lakes performance are not easy to achieve, unless good data governance is guaranteed This can be compared to a real lake, which also needs a guard who keeps an eye on who fishes in it (Lo, 2016), what gets into the lake, how many anglers are there, what are the sources of the inflow to this lake, etc. Otherwise, the lake will be dirty and unusable.

CONCLUSION

The topic of data lakes is not just a technical issue. It also has significant business implications from an organizational point of view. Data lakes require people to be open about the data possessed by the 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 data governance methods. If data are well-structured or stored in small amounts, conventional approaches are sufficient. Although, once integrating all kinds of data in one big data lake, various problems might emerge. Regardless of the benefits of creating transparent relationships between users and data in the enterprise, the security and legal issues associated with data lakes remain important but challenging and require more attention and effort from senior management due to the need for a cultural shift within the organization.

REFERENCES

Bleiberg, J., West, D.M. (2018). In the Future We Will Store Data Not in a Cloud But in a Lake. Brookings, accessed 25 May 2021, available from: http://www.brookings.edu/blogs/techtank/posts/2014/07/28-big-data-lakes

Chase, G. (2017). 10 Amazing Things to Do With a Hadoop-Based Data Lake. Pivotal, accessed 23 May 2021, available from: http://blog.pivotal.io/big-data-pivotal/features/10-amazing-things-to-do-with-a-hadoop-based-data-lake

Cooper, D.R., Schindler, P.S. (2019). Business Research Methods. 10th Edition. New Jork: McGraw-Hill Education, 370–377.

Darrow, B. (2014). Pursuing big data utopia: What realtime interactive analytics could mean to you. Gigaom, accessed 20 May 2021, available from: https://gigaom.com/2013/03/21/pursuing-big-data-utopia-what-realtime-interactive-analytics-could-mean-to-you/

Devlin, B. (2015). Data lake muddies the waters on big data management. TechTarget, accessed 20 May 2021, available from: http://searchbusinessanalytics.techtarget.com/feature/Data-lake-muddies-the-waters-on-big-data-management

EMC^2 white paper 1 (2011). "Federation Business Data Lake – Enabling Comprehensive Data Services", Hopkinton, Massachusetts.

Hurwitz, J., Nugent, A., Halper, F., Kaufman, M. (2019). Unstructured Data in a Big Data Environment. Accessed 26 May 2021, available from: http://www.dummies.com/how-to/content/unstructured-data-in-a-big-data-environment.html

Inmon, B. (2016). Data Mart Does Not Equal Data Warehouse. Accessed 24 May 2021, available from: http://www.information-management.com/infodirect/19991120/16751.html?zkPrintable=1&nopagination=1>

Klein, J. (2014). Relational Data Lake. Accessed 24 May 2021, available from: https://jorgklein.com/2014/12/02/relational-data-lake/

Lawson, L. (2016). Another Barrier to Data Lakes: The Metadata. Accessed 30 May 2021, available from: http://www.itbusinessedge.com/blogs/integration/another-barrier-to-data-lakes-the-metadata.html

Lo, F. (2016). What is Hadoop? What is MapReduce? What is NoSQL? DataJobs, accessed 23 May 2021, available from: https://datajobs.com/what-is-hadoop-and-nosql

Shapira, G. (2015). Hadoop and NoSQL Mythbusting. Pythian, accessed 24 May 2021, available from: http://www.pythian.com/blog/hadoop-and-nosql-mythbusting

Stein, B., Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. Technology Forecast: Rethinking Integration, 1(4), 355–365.

Widom, J. (2019). Research Problems in Data Warehousing. Stanford University. Proc. of 4th International Conference on Information and Knowledge Management (CIKM), Nov. 2005.

Woods, D. (2011). Big Data Requires a Big, New Architecture. Forbes, accessed 23 May 2021, available from: http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/2

Singh, A. (2021). Methods for data lakes implementation. Intercathedra 3(48), 151–157.
http://dx.doi.org/10.17306/J.INTERCATHEDRA.2021.00130

METODY WDRAŻANIA JEZIOR (REPOZYTORIÓW) DANYCH

Abstrakt. Badania nad jeziorami danych (ang. *data lakes*) w czasach stosowania big data są znaczące zarówno dla środowisk akademickich, jak i przemysłu. Jedną z głównych motywacji jest większa niż kiedykolwiek wcześniej ilość danych, z którymi muszą radzić sobie firmy, a problemy związane z analizą danych, a nawet ich przechowywaniem, stają się coraz większym wyzwaniem w wielu branżach. Koncepcja jezior danych w celu sprostania problemom związanym z big data jest pomocna i najprawdopodobniej będzie stosowana w każdej istotnej strategii big data. Pomysł ten jest wciąż w trakcie sprawdzania i poddawany krytyce. Na szczęście pojawia się na jego temat coraz więcej pozytywnych głosów, wysoko oceniających tę koncepcję. Proponuje się nawet innowacyjne sugestie dotyczące tego rozwiązania. W tym opracowaniu przedstawiono podstawowe informacje na temat wdrożenia data lakes oraz spostrzeżenia, które mogą być przydatne dla praktyków tej koncepcji. W artykule przedstawiono i podsumowano większość popularnych metod na wdrożenie jezior danych i scharakteryzowano trzy główne z nich. Każda ma zarówno zalety, jak i wady, a firmy muszą rozważyć własne potrzeby biznesowe i wymagania, aby wybrać najkorzystniejsze z rozwiązań.

Słowa kluczowe: Data Lake, Hadooo, dane, wirtualizacja, big data, analiza