Cloud DATA LAKE: The new trend of data storage

Elisabeta ZAGAN[1,2] and Mirela DANUBIANU[1,2]
[1] Faculty of Electrical Engineering and Computer Science, Stefan cel Mare University, Suceava, Romania
[2] Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD), Stefan cel Mare University, Suceava, Romania
elisabeta.b@gmail.com, mdanub@eed.usv.ro

2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) | 978-1-6654-4058-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/HORA52670.2021.9461293

***Abstract*—In databases field, the term Data Lake is increasingly common, which is a new raw data storage technology to undergo further advanced processing and analysis. Today there are for different ways to implement Data Lake architecture, namely: Data Lake On-Premises, Cloud Data Lake, Hybrid Data Lake and Multi-Cloud Data Lake. Each of these architectures has their own advantages and disadvantages, and yet the new trend is to go in Cloud. In this work, we will briefly explain what a Date Lake is, what the for different architectures are, and we will broadly present the major benefits of the Cloud architecture and its shortcomings in order to provide a preliminary guide on the implementation of Data Lake architecture.**

Keywords—Data Lake, Data Lake architecture, Data Lake On-Premises, Cloud Data Lake, Hybrid Data Lake, Multi-Cloud Data Lake

I. INTRODUCTION

More than 15 years ago, there were few third-party data set processing solutions, but as the capacity to generate data increased, the need for data storage and processing has also increased. The large amount of data has brought with it the need for significant changes in the architecture of storage and processing systems. Along the way with the help of new technologies, new adaptive hardware and software systems have been deployed to meet the new requirements for big data storage and management [1].

In databases field, in recent years, the concept of Data Lake is increasingly common, this can be seen as a lake of data, a lake where absolutely all the data that reaches it will be stored in order to be subsequently accessed and explored by different users [2]. Data Lake is a modern raw data storage technology that brings many advantages, due to the fact that, within a company, it allows the absolute storage of all data in their raw format, so that it can be subsequently subjected to advanced analysis processes to obtain crucial information unforeseen at the beginning of the storage process. Fig. 1 shows the structure of a Data Lake at the conceptual level [3].

This article is structured as follows: Chapter I a brief introduction, Chapter II presents the 4 types of Data Lake architectures, Chapter III describes the main advantages and challenges of building a Data Lake in Cloud, Chapter IV presents the related work, and in Chapter V the main conclusions are drawn.

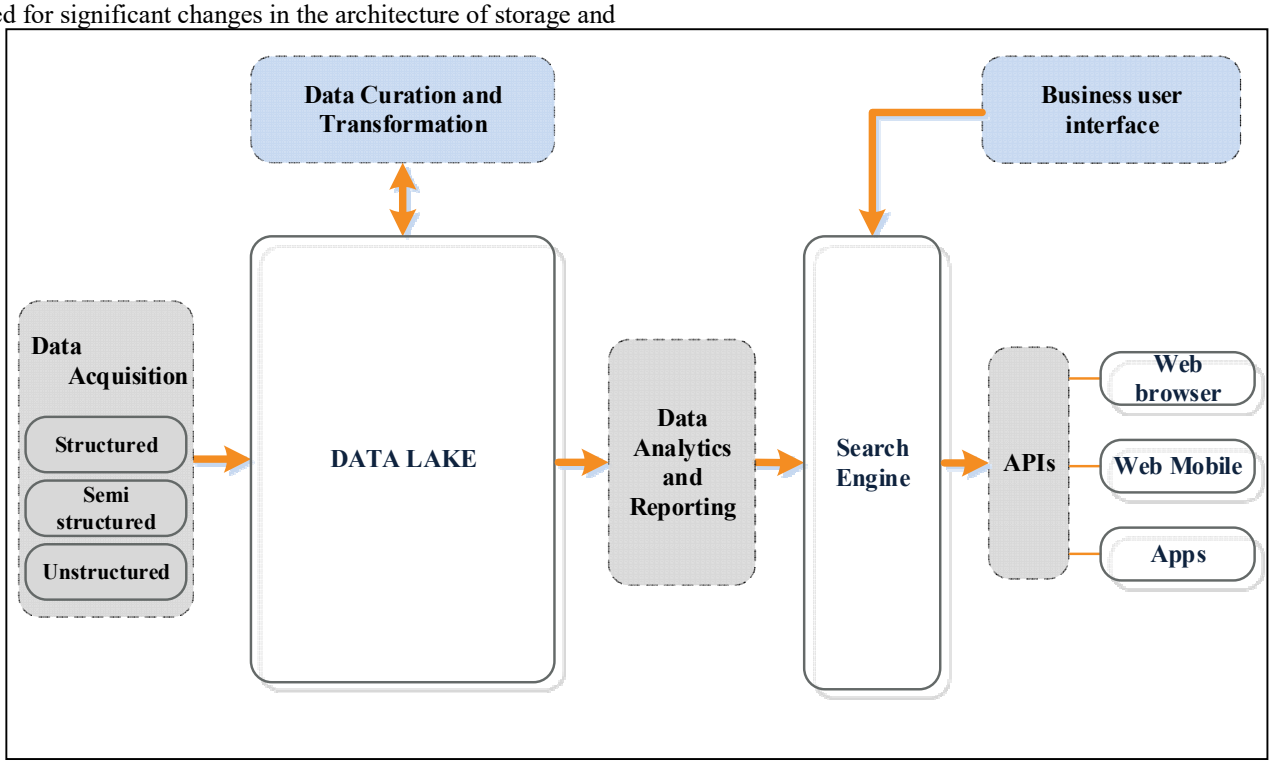

Fig. 1. Data Lake concept.

This work is supported by the project ANTREPRENORDOC, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847.
Equations

978-1-6654-4058-5/21/$31.00 ©2021 IEEE

II. Types of Data Lake Architectures

Data Lake architecture can be implemented in several ways (Fig. 2). It all depends on the choice of architecture that best fits the needs of a business based on costs and hardware and software resources:

- **Data Lake On-Premises:** Data Lakes physically deployed locally require the management of both hardware and software technologies. The major disadvantage is that this double task requires major engineering resources and experts in the field who need to improve their knowledge with new and rapidly evolving technologies. It is difficult and costly to constantly invest in a team of experienced engineers in the field to guarantee quality service and advanced data security. Another problem is that hardware resources must be monitored so that the ingested data does not reach the available storage limit that could lead to significant data loss issues. The most common technology used to implement a Data Lake On-Premises is the open source Hadoop framework from Apache..
- **Cloud Data Lake:** The major advantages are: the immediate availability of creating a Data Lake, the hardware scalability that is done automatically, the response speed and the low costs for the engineering team who should have knowledge of managing Cloud services that are user-friendly. Each Cloud services provider offers complete tutorials about their own services, and on the other hand social networks are also a good source of training materials. The downside may be that Cloud services are paid, and over time, it remains to be seen whether it can be more expensive than having a local architecture and thus a team of engineers. Among the largest Cloud services providers who have managed to deploy and deliver high-performance Data Lake solutions, we mention Microsoft Azure, Amazon S3 and Google Cloud..
- **Hybrid Data Lake:** Allows Data Lake to be maintained locally as well as in Cloud, which can present a number of benefits but also a number of challenges. The major advantage is that less relevant data can be stored locally to reduce storage costs, taking advantage of the Cloud speed services on important data instead. However, this architecture involves even higher costs for the team of IT experts and engineers who should have very good technical skills for both storage environments in order to achieve the communication between them.
- **Multi-Cloud Data Lake:** This last type of Data Lake architecture is the one that proposes using combined Cloud services from different providers, for example, there are large companies that use both AWS and Microsoft Azure to manage and maintain Data Lakes. Having multiple Data Lakes on different Cloud platforms means benefiting from the advantages of each platform, but requires greater technical skills to achieve the communication between them..

Mostly, the Date Lake term is associated with Apache Hadoop, the data being stored in HDFS clusters that are physically owned locally within companies. Unfortunately, many of these local Data Lake projects have failed to successfully complete a reliable architecture due to the complexity of implementing such architecture and the great efforts of managing the entire system. Data Lake, as we well know, must provide services for collecting, storing, analyzing and securing all types of data in one place. Cloud technology has evolved so much recently that Data Lake in Cloud is, in some ways, beyond the Data Lake On-Premises.

In recent years, many companies with large volumes of self-managed data on platforms such as Hadoop tend to migrate to the Cloud [4] due to the many advantages we will highlight below.

III. Advantages and Challenges of Building a Data Lake In Cloud

Moving data to the Cloud, today has become an affordable technique for all businesses of all sizes. Among the major advantages we can mention:

- **Storage capacity:** In the Cloud, storage is unlimited and eliminates the problems that can occur when expanding and maintaining a local data storage network.
- **Cost effective:** Cloud storage providers offer different services at different costs. This helps companies pay for exactly the services they need at the different stages of development, different from local deployment that requires an approximate estimation of costs considering their future development prospects.
- **Central repository:** The Cloud has the advantage of a centralized location to store all types of data that can be accessed from any location. This greatly simplifies the complexity of an IT team's assignments, as it is freed from the responsibilities that would involve local deployment.

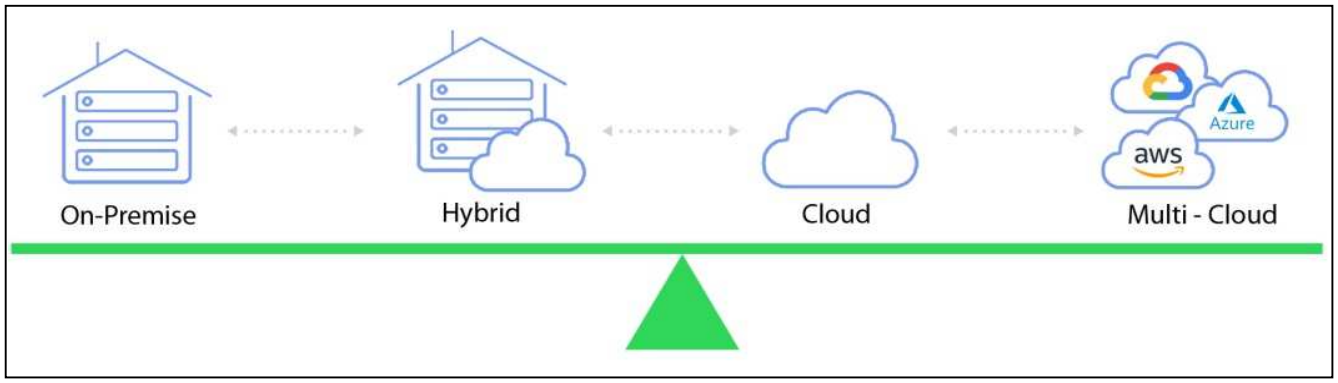

Fig. 2. Data Lake architectures types.

- **Data security:** All companies have the major responsibility to protect their data. With Data Lake designed to store all kinds of data, including information such as financial data or sensitive customer data, security becomes a priority and fundamental to data storage. Cloud service providers guarantee data security according to the shared security responsibility model.
- **Automatic scaling:** Modern Cloud services are designed to provide immediate scalability of data, so businesses don't have to worry about expanding hardware capacity when needed.
- **User-friendly interface:** Another advantage of Cloud services is the user-friendly interface at any level. This advantage is not to be neglected due to the fact that it facilitates access to the fewest experts in the field.

Migrating data and infrastructure to the Cloud as we have seen has multiple benefits that reduce operating costs within companies (Fig. 3) however there are a number of challenges related to longer-term costs, data analysis and migration data [5].

In terms of costs over time, Cloud providers charge companies for storage on time rather than size. Over long periods of time, a Cloud may be more expensive than a local storage. This remains to be assessed by each company according to the possibilities for local management of a storage environment, as IT experts in this field in some regions may be difficult to find, so implicitly there will be higher costs of creating and owning a full team of reliable experts to provide high-quality services [6]. The main advantage of using a Data Lake for storing raw data consists in the benefits obtained from analyzing these data. The ability to transform, organize and combine different data sources is a huge benefit of Data Lake, but requires an equally robust analysis solution. Most Cloud providers offer advanced analysis solutions but less customizable as in the case of locally owned Data Lakes. Here we can mention, however, that Cloud service providers are constantly improving their services, and the companies that want to implement a Data Lake locally to manage independently must always rely on a professional IT team that is not always easy to form.

When it comes to data migration, moving data to the Cloud can face some difficulties being a complex process that requires a preliminary evaluation to choose a Cloud service provider that suits personal needs. Regardless the type of the architecture you choose to build a Data Lake, we can say that for companies that want to discover the benefits of a Data Lake, they can start with the services offered by Cloud providers. Following the experience gained on the Cloud, it will be possible to correctly assess the need of implementing a Data Lake locally, taking into account the costs, the implementation time and the complexity of such an architecture.

IV. Related Works

In recent years, the concepts of 'Big Data' [7], [8], Hadoop and Data Lake has been used more and more often due to the sudden increase in information and data that needs to be stored, processed, and then interpreted into various statistics.

The above mentioned authors offer code [9] lines to illustrate the ease with which a data lake can be created and queried and the way in which data security and control can be achieved. An example of an application is given towards the end of the paper, presenting a scenario where an analyst is interested in the government health budget by involving the audience's thoughts and opinions on social media. With CoreDB they created a Data Lake where more than 15 million social media pieces of information related to Australia's government budget were entered using a simple JAVA code. Subsequently, they created a relational database where they saved data concerning the budgetary healthcare program, such as registered doctors and nurses in Australia, hospitals and chemist's shops, healthcare funds, medical tools and devices, medicine, medical conditions and keywords related to health. Various full-text searches were conducted on all this data, highlighting the capacity of data indexing in CoreDB. The solution presented in [10] provides the capacity to process not only structured data, but also unstructured data stored in its natural form in the data lake, using a unified interface of extractors that display data as a set of rows. The solution also simplifies the data fuzzy transformation process using U-SQL SELECT expressions, unlike Hadoop/Spark-based processing which requires adaptation to a certain processing pattern, such as MapReduce and the implementation of dedicated processing functions.

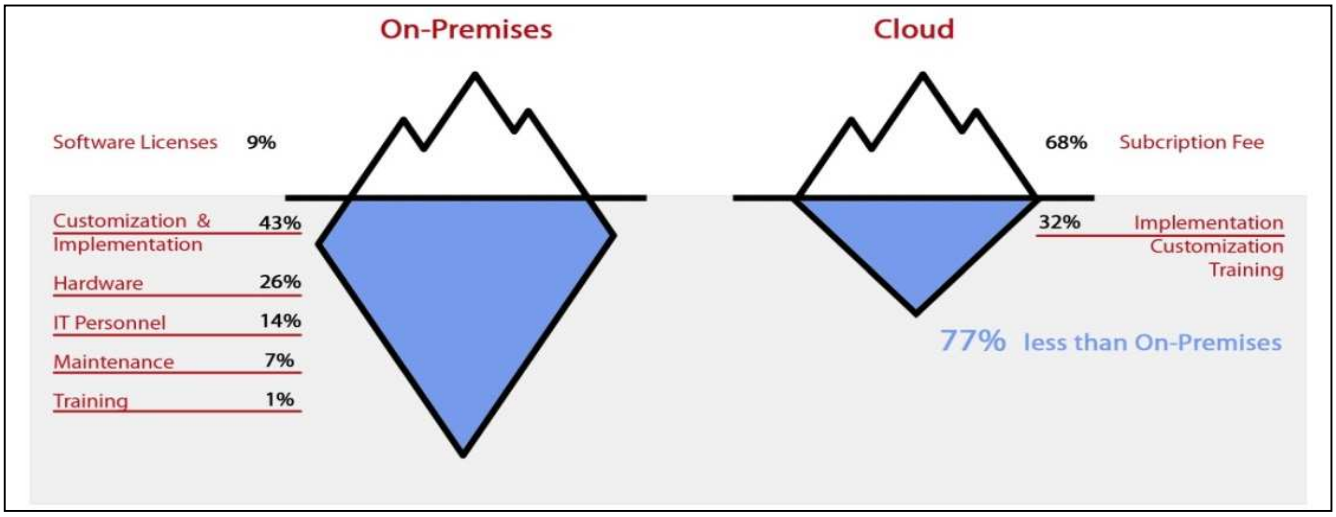

Fig. 3. Cloud vs On-Premise total costs.

Fuzzy techniques supplied as functions (UDFs) directly accessed by U-SQL queries not only allow data to be processed declaratively, but also optimise U-SQL performance plans in order to dispense fewer resources and reduce the data processing cost. The authors' intent was to develop the Fuzzy Search Library for Big Data Lake as a universal tool that provides methods of fuzzy processing and scaling of data on the Azure Cloud for various analysed data fields. By massively overlapping computing/calculations related to data extraction, processing and storage on the Cloud, the authors claim that a significant decrease in the data processing time can be achieved.

The concept of Data Gravity was coined by Dave McCrory [11] to describe the phenomenon in which the number or quantity and the speed with which services, applications and even customers are drawn to data, increases as the amount of data increases. Based on this assumption, the authors in [12] outline a series of features a PDL must have in order to speed this data gravity, such as data and application detachment, data application mobility and play-and-plug assistance for portable code. In [13] the authors point out the utility of using the Data Lake Introspection tool (DLI) towards achieving new data relationships by merging Hive tables stored in the Hadoop environment. In order to acquire this information, experts should be able to identify these data sets, while encountering the problems related to finding and using relevant data, risk management and data security. Unlike current Data Lakes, which send users unencrypted data, the proposal presented in [14] requires the data provider to be in possession of a platform in the Cloud, which the data consumer could use to create a virtual machine (VM) to carry out the public data analysis tasks. VM would access the public data free of charge, while in the Cloud a hardware or software firewall would be configured to monitor the data output traffic used for charging data consumers.

V. Conclusions

To be truly useful, a Data Lake must be able to easily store data in its raw form, to facilitate the exploration of these data, to automate data management activities and to enable the use of a wide range of data analysis technologies.

However, most local Data Lakes cannot efficiently manage all of a company's data. Moreover, today's Data Lakes must allow the ingestion of data from different sources, each of which provides data at a different frequency. Without proper data quality and proper governance, even well-built locally Data Lakes can quickly become data swamps - unorganized data that is difficult to use, understand and share with multiple users. The greater the amount and variety of data is, the more significant this problem becomes. Other common problems in On-Premises Data Lakes include poor performance, management difficulties and scaling.

In recent years, Cloud services providers have evolved and successfully managed to meet the requirements of either structured or unstructured data, therefore it is possible to successfully deploy a Data Lake in the Cloud. The new trend is to move and implement a Data Lake in Cloud, which provides complete services for the entire process of data ingestion, storage, processing and analysis with a high level of security, which are constantly improved by Cloud service providers.

Acknowledgment

This work is supported by the project ANTREPRENORDOC, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847.

References

[1] E. Zagan, M. Danubianu, "HADOOP: A Comparative Study between Single-Node and Multi-Node Cluster", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 12 Issue 2, 2021. doi:10.14569/IJACSA.2021.0120207

[2] J. Dixon, Pentaho, "Hadoop and Data Lakes," Retrieved 10 Aug 2017. [Online]. Available: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

[3] E. Zagan, M. Danubianu, "From Data Warehouse to a New Trend in Data Architectures - Data Lake", International Journal of Computer Science and Network Securiity (IJCSNS), Vol. 19, No. 3, March 2019, pp. 30-35.

[4] R. Ghazi, D. Gangodkar, "Hadoop, MapReduce and HDFS: A Developers Perspective", Procedia Computer Science, Volume 48, 2015, Pages 45-50, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.108.

[5] https://www.slideshare.net/sugarcon/cloud-session-7-cloud-computing-software-as-a-service-and-sales-forecasting, accessed: 2021-03-02

[6] https://www.informationweek.com/big-data/top-trends-in-data-lakes/a/d-id/1338651?, accessed: 2021-02-10

[7] S. S. Al-Rifai, A. M. Shaban, M. S. Muayad Shihab, A. S. Mustafa, H. A. ALHALBOOSI and A. M. Shantaf, "Paper Review On Data Mining, components, And Big Data," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020, pp. 1-4, doi: 10.1109/HORA49412.2020.9152919.

[8] M. R. Naqvi, M. Arfan Jaffar, M. Aslam, S. K. Shahzad, M. Waseem Iqbal and A. Farooq, "Importance of Big Data in Precision and Personalized Medicine," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2020, pp. 1-6, doi: 10.1109/HORA49412.2020.9152842.

[9] A. Beheshti, B. Benatallah, R. Nouri, V. M. Chhieng, H. T. Xiong, and X. Zhao, "CoreDB: a Data Lake Service," In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17). ACM, New York, NY, USA, 2451-2454. doi: https://doi.org/10.1145/3132847.3133171.

[10] B. Małysiak-Mrozek, M. Stabla and D. Mrozek, "Soft and Declarative Fishing of Information in Big Data Lake," in IEEE Transactions on Fuzzy Systems, vol. 26, no. 5, pp. 2732-2747, Oct. 2018. doi: 10.1109/TFUZZ.2018.2812157

[11] D. McCrory, "Data Gravity – in the Clouds," [Online]. Available: http://blog.mccrory.me/2010/12/07/data-gravity-in-the-Clouds/ , 2014.

[12] C. Walker and H. Alrehamy, "Personal Data Lake with Data Gravity Pull," 2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, 2015, pp. 160-167. doi: 10.1109/BDCloud.2015.62

[13] A. Farrugia, R. Claxton and S. Thompson, "Towards social network analytics for understanding and managing enterprise data lakes," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 1213-1220. doi: 10.1109/ASONAM.2016.7752393

[14] Y. Chen, H. Chen and P. Huang, "Enhancing the data privacy for public data lakes," 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, 2018, pp. 1065-1068. doi: 10.1109/ICASI.2018.8394461.