Data Lake Approaches: A Survey

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

***Abstract*—The explosion of new data: social media, commercial, industrial, health, school, etc. appeared in recent years, has led to the emergence and development of new technologies and techniques of data management. The old technologies of data storage and data processing are beginning to be overwhelmed by the large volume of data and their variety. Data Lake is one of the latest technologies that seem to be in the spotlight in the last period. In this article, we analyze some of the recent approaches and architectures using Data Lake, approaches that have tried to cover several shortcomings encountered with the advent of these new technologies.**

Keywords—Data Lake, Data Warehouses, Databases, Data Analysis, Data Mining

I. INTRODUCTION

In any company, from any field of activity, the aim is to extract the maximum amount of benefits that could be obtained through the study and analysis of the volume of data at their disposal. Lately, market leaders have been searching for more advanced and efficient techniques to quickly and easily extract the information necessary for new development strategies and to obtain maximum profit from all the data in their possession. Each organization owns a different amount of data - either simple or more complex, sometimes becoming too complex to be easily managed even by the most experimented ones in the field. Given that the technology has evolved and companies have started to collect more and more data, it has become necessary or, rather, vital to creating reports and analysis in various company branches. Analytics was an answer to the needs of growing businesses.

Today's business leaders have understood that data is the key to success, by understanding the demand of the clients, competitors and the market in general. Only by analyzing this information can they take action and make the right decisions for a guaranteed success, minimizing the errors in development strategies and decisions, which they need to take at some point to achieve success and existence on the market.

Data storing technology from recent years have used the ELT (Edit-Transform-Load) process as the base storing technique, whereby data was first processed, cleaned, and then stored. Through this method, data that did not pursuit, a specific goal was eliminated. The eliminated data proved to be valuable data in time, considering that it is impossible to anticipate any question or requirement of a report which a company might need at some point in its development and improvement process. The statistics may vary from year to year, from one month to another and sometimes even from one day to another. In addition to these shortcomings, several new types of data have emerged from the web, social media, comments, servers, sensors and various devices that have generated a real explosion in the volume of data that organizations are struggling to store, to understand and process them [1]. For instance, 15 – 20 years ago, companies did not expect that in the near future it would be so necessary to keep a record of "likes" on different social media, as they could provide vital information since they represent direct feedback from users in the online environment. Thus for the past years, the explosion of these new types of data has determined the appearance and development of new technologies and data management techniques. Lately, Data Lake seems to be in the spotlight.

Data Lake is a new working method that simplifies and improves Big Data storage [2], management and analysis by using natural, raw data from different sources. A Data Lake is, essentially, a storing place for structured and unstructured data, a Big Data analyzing tool, a resource of raw data that can be accessed, distributed and analyzed. Data Lake's main features are [3]:

- Possibility to store all types of structured, semi-structured, unstructured or binary data, data from transactional systems, sensor data, data from different applications.
- Storing a very large amount of data at a low cost, the data storage amount can be increased without having to modify the data storage structure and diagram/design.
- Algorithms of selection and raw data analysis can always be improved, thus companies are able to obtain new and new answers that they can take into account in their development strategies;
- It is a highly agile structure, it can be configured and reconfigured any time it is necessary.
- The possibility of getting results from unlimited types of data will allow companies to obtain various

This work is supported by the project "Integrated Center for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control", Contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic

978-1-7281-6870-8/20/$31.00 ©2020 IEEE

information that can improve the quality of services at any angle.

- More flexibility, a major plus being the fact that not all answers are needed before conducting any analysis.
- The possibility of using different tools to get a perspective on the data.
- It uses the Extract-Load-Transform (ELT) process to store and process data, thus not filtering the data before storing it.
- Another important reason for using Data Lake is that big data analysis can be performed much faster.

This paper is structured as follows: after a brief introduction in the first section, a survey on data lake approaches for all companies is presented in section II. Conclusions are presented in section III.

II. A Survey on Data Lake Approaches

This research aims to carry out a study on the main Data Lake architectures. The originality of the work is that it summarizes, in a single material, these architectures with their broad characteristics. In scientific articles, we can find different architectures that attempt to discover optimal solutions for different issues that occurred along with the new technology, such as Data Quality, Data Security, Data Life Cycle, Data Gravity and User Interface.

A. CoreDB

CoreDB service [4] is an open-source Data Lake which offers researchers and developers a single REST API application that allows data and metadata organizing, indexing and querying.

CoreDB manages several database technologies and offers an integrated design for security and tracking data changes. Its purpose is to lay forward a simple data and metadata storage and management solution, given that organizations have been confronting a tremendous amount of different types of data, i.e. structured, semi-structured and unstructured, while analysts have had to operate a large number of digital information generated via social networks, blogs, online communities and mobile applications that create a complex Data Lake.

In Fig. 1 the full CoreDB open-source architecture and its main components are displayed. Organizing and indexing this vast amount of data is challenging and can only be managed by experts in the field, who use the latest technology. CoreDB is the solution the authors provide; this open-source employs numerous database technologies (from the relational database, NoSQL) and provides a built-in design to support the following functions:

- Access security and control: to provide database security, authentication, access control and data encryption.
- Data tracking and data source: to collect and aggregate metadata, including descriptive, administrative and temporary metadata, and drawing a provenance graph.

CoreDB enables analysts to build a Data Lake, to create sets of relational data and/or NoSQL within the Data Lake, to apply different CRUD (Create, Read, Update and Delete) instructions and query entities on these data sets. CoreDB allows Elastic Search due to the Apache Lucene (lucene.apache.org/) search engine, which has a powerful indexing and full-text search system. CoreDB has a built-in design in order to provide a top security system along with tracking and provenance support.

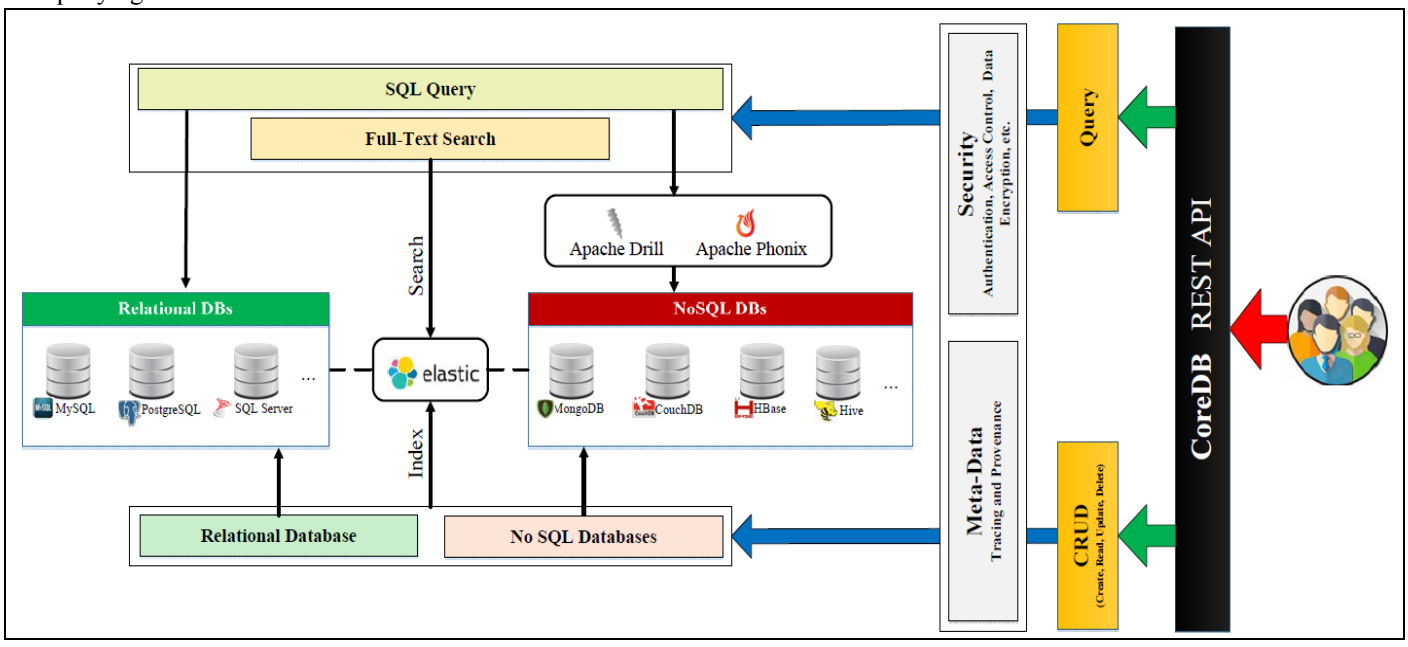

Fig. 1. CoreDB Architecture [3].

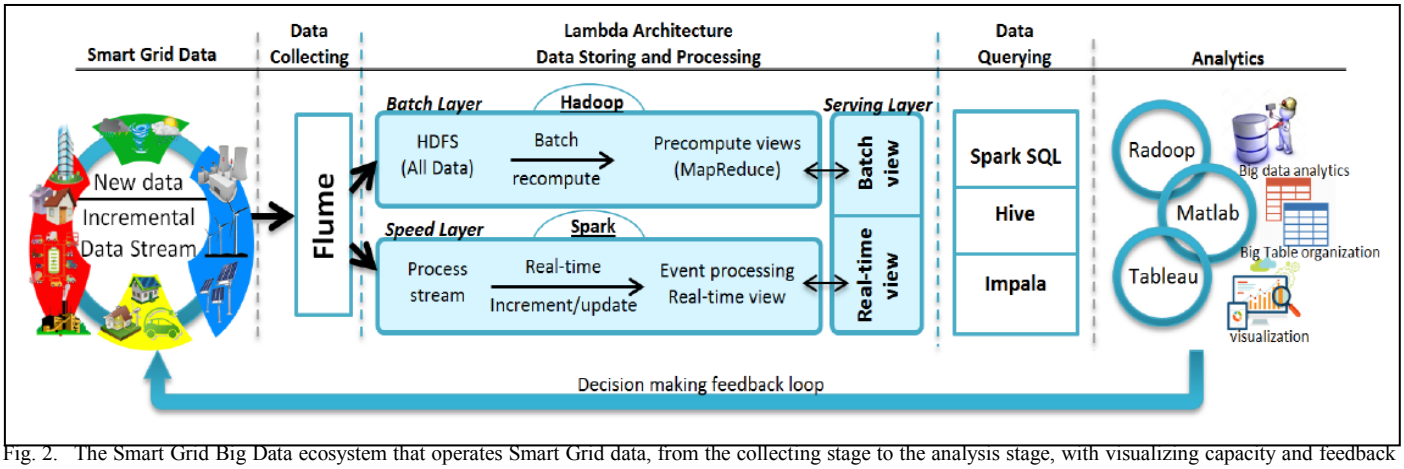

Fig. 2. The Smart Grid Big Data ecosystem that operates Smart Grid data, from the collecting stage to the analysis stage, with visualizing capacity and feedback loop [4].

B. *Smart Grid Big Data*

Smart Grid Big Data [5] ecosystem is based on the latest Lambda architecture, which is able to manage high quantities of data and carry out a batch and real-time operations. The ecosystem uses Hadoop Big Data Lake to store various types of Smart Grid Data, including Smart Meter, pictures and videos in order to be manipulated later through different processes. Cloud Data Lake storing was used because it enables different types of data from different sources to be stored in one place. To test the capacity of the ecosystem displayed in (Fig. 2), real-time data visualization and extraction applications were performed.

This ecosystem, implemented on a Cloud computing platform focuses on using the latest technology components and platforms provided by companies such as Google and Facebook, in order to withstand challenges from smart grid data. It uses a Hadoop Big Data Lake to collect various types of Smart Grid data, including pictures and video files and it is capable of efficient real-time or almost real-time extraction of massive data sets in order to improve decision-making for future benefits. Performing Smart Grid Data Mining applications on the last level this ecosystem can be used not only by scientists but also by common users.

C. *Big Data Technologies for Public Data Lakes*

In [6] a framework for public data lakes that would control and protect the confidentiality of distributed data is presented. Its main purpose is to stimulate the distribution of valuable data in Big Data analysis processes and to promote the development of new Big Data technologies.

To increase the confidentiality of data, public data lakes should be redesigned to prevent data consumers from redistributing public data to other consumers without the consent of data providers. The key technologies that allow the distribution and analysis of valuable data between so-called unreliable consumers, while data confidentiality can be controlled through the pay-per-use payment method provided by Cloud platforms are proposed. For ensuring data security, data which must become public should not be stored directly into Data Lake storage environments because they may have major problems losing control over them against unreliable consumers. It is designed an architecture (Fig. 3) where each data provider resorts to Cloud storage to store data that is due to become public, because, in the Cloud, authorized consumers can access data through remote connections (RPC – Remote Procedure Calls) or through applications (API – Application Programming Interfaces). Thus, Data Lake remains to ensure only the storage of metadata and data set management information, ensuring the necessary data. Here, metadata will hold information related to size, shape, source, usage restrictions, prices, etc. – information necessary to the consumers for the purpose of renting them.

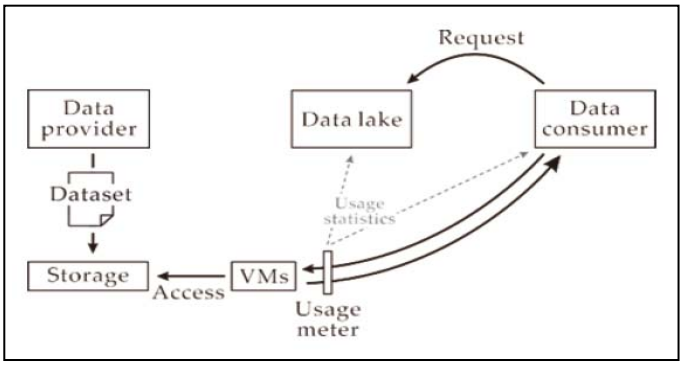

Fig. 3. Data distribution protocol in the proposed Data Lake architecture [5].

D. *Ceph*

At present, there is no doubt about the importance of mass data storing and processing technologies used in the background, in order to satisfy present-day megatrends in the field of Information, Communication and Technology focused on IoT, Big Data, Cyber Physical System (CPS) and AI. To assess the storage performance of large volumes of data, there were conducted a series of experimental tests on the Ceph open-source environment, using the Abyss storing environment, and examined the network response performances using Korea Advanced Research Network (KOREN) [7]. These tests were made for both domestic use, as well as external websites. In order to improve the storing security and performance of data distributed in the Abyss

cluster based on the Ceph open-source, several tests were conducted in this study, such as performance tests on the disk-type storage devices, tests on traffic and network connections, and also security tests on Cuckoo sandbox and Yara malware tests. Aiming at solving the one-way Data Lake issue called Garbage Dump, the authors proposed the application of mathematical topologies and Machine Learning (ML) technology. Therefore, a Data Lake framework (Fig. 4) was proposed in order to test a few methods of organizing data from a Data Lake for later use: mathematical topologies and automated learning technologies. Mathematical topology examines a collection of data subsets that match certain properties, the returning set forming a topological space. This method is used to study the form of data. Machine learning is based on algorithms that learn to recognize and make predictions on data.

E. Data Lake Introspection tool (DLI)

In [8] a tool for inspecting and managing Data Lake was built. The tool works by extracting metadata from the Hive database, from a shared Hadoop platform, which contains a supply of multi-terabyte real data. This metadata was used to draw a chart of the relationships between entities through column correlation, allowing them to apply social network analysis techniques (SNA – Social Network Analysis) in order to discover important properties of the accumulated data, such as detecting unknown previous relationships between data entities. Data Lakes usually offer organizations data that can be manipulated towards obtaining valuable information.

In order to acquire this information, experts should be able to identify these data sets, while encountering the following major problems: finding and using relevant data; risk management and data security. So, it's obvious the utility of using the Data Lake Introspection tool (DLI) towards achieving new data relationships by merging Hive tables stored in the Hadoop environment.

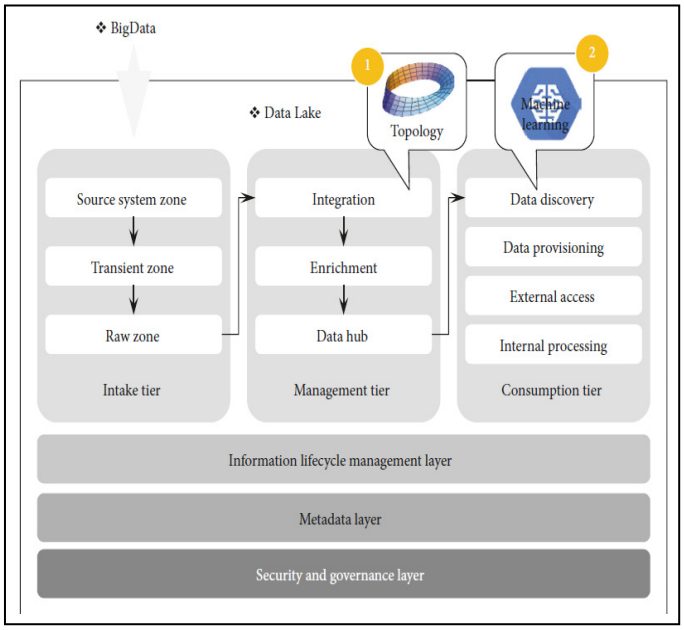

Fig. 4. Data Lake framework draft architecture [6].

F. Azure Data Lake

A method of analyzing large amounts of data by resorting to Fuzzy Search libraries is presented in [9]. This analysis was performed on an Azure Data Lake situated on a Cloud platform [10]. The developed solution grants total and complex control over data during the entire EPS (Extract, Process and Store) process in a Big Data Lake, data that is extracted, processed, converted and later stored in well-defined places and in a format that is optimal for analysis. The potential of Fuzzy libraries from the Azure Data Lake in the process of searching data Big Data is demonstrated. One of the most important advantages is the unlimited scalability that allows for fast adaptation to the increase of the data volume. The second biggest advantage is the scaling simplicity of the fuzzy query process without reconfiguring the entire runtime environment (hardware and software), as opposed to solutions based on Hadoop/Spark and NoSQL databases, which would require a reconfiguration of the entire cluster if scaled (for instance, adding new cluster nodes).

The general design for the Azure Data Lake is illustrated in Fig. 5. There are delimited two main parts of this architecture:

- Data Lake Store – DLS (which provides petabyte scaling, unlimited storage for the DL data lake).
- Data Lake Analytics – DLA (which enables efficient and scalable analysis of data stored in the Big Data Lake, parallelizing analysis on an infrastructure distributed in Cloud Azure).

G. Personal Data Lake

Some issues challenging a Data Lake can be solved through the Personal Data Lake (PDL) architecture [11]. It starts from the premise that solving issues in a restricted plan as in the case of PDL and applying strategies for storing and processing data on a restricted/personal level would later lead to a global architecture that would successfully meet the problems of privacy, security and gravity of data. In Fig. 6 two different paths of personal data are displayed. The first one is the conventional way, where personal data is shared across multiple service sites with which the users interact.

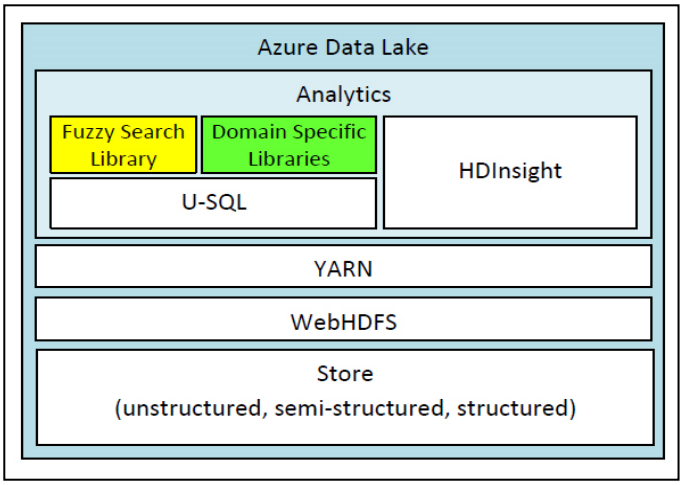

Fig. 5. The general architecture proposal for Azure Data Lake with the Fuzzy Search Libraries [8].

In this case, the data consumers who collect personal data for data analysis purposes could access the data without the user's knowledge. The second path is a personal data lake serving as a repository for personal data consolidation, which provides data management, security and query interfaces for third-party data consumers.

The connection between the data lake and the concepts of data gravity is seen as a perfect solution for storing a large variety of personal data and offering a secure entry point for third-party queries. The personal entry point of lake querying, if efficiently connected, can lead to a powerful global platform for personal data extraction and analysis and, also, supports a fair and healthy expanding data market economy [12].

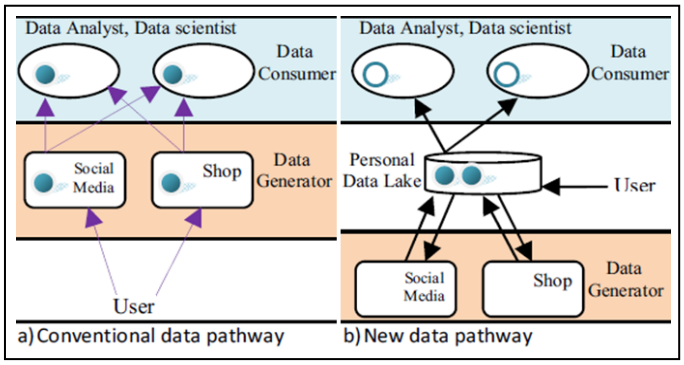

Fig. 6. Personal data paths [9].

III. Conclusion and Future Work

Data Lakes are becoming increasingly more important for marketing strategies regarding big companies. On closer examination of the rapid evolution of the Internet of Things (IoT), the Data Lake seems to have steadily become the optimal solution for storing data coming from this branch of technology. It is well-known that Data Lakes are data repositories where organizations collect and store all the data they need in order to analyze it for specific purposes. The nature, form, or source of the data are irrelevant in this new concept. These data "flows" come in multiple formats: structured data (data from a traditional relational database or even spreadsheets: rows and columns), unstructured data (social, video, e-mail, text, etc.), practically, any type of data. Once stored, this data can be accessed anytime to be analyzed. Data Lake has also the capacity to keep this data for a longer period of time, thus making long term data analysis possible.

Analyzing the data is not enough, thus it is followed by a series of actions that will lead to reaching the final goals in different organizations. Data Lakes are suitable for using large quantities of algorithm data that will contribute to real-time analysis. Therefore, the Data Lake has proven to be exactly what organizations need for Big Data Analytics in a mixed data environment.

Acknowledgment

This work was partially supported by the project "Integrated Center for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control", Contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic Competitiveness co-funded from the European Regional Development Fund.

References

[1] M. Danubianu, T. Socaciu, and A. Barila, "Some aspects of data warehousing in tourism industry," Stefan cel Mare University of Suceava, Fascicle of The Faculty of Economics and Public Administration, 2009(1 (9)), 290-296.

[2] E. Zagan, M. Danubianu, "From Data Warehouse to a New Trend in Data Architectures – Data Lake," IJCSNS International Journal of Computer Science and Network Security, vol.19, no.3, March 2019

[3] N. Miloslavskaya, A. Tolstoy, Big Data, Fast Data and Data Lake Concepts, 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

[4] 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.

[5] A. A. Munshi and Y. A. I. Mohamed, "Data Lake Lambda Architecture for Smart Grids Big Data Analytics," in IEEE Access, vol. 6, pp. 40463-40471, 2018. doi: 10.1109/ACCESS.2018.2858256.

[6] 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.

[7] Cha, Byung & Park, Sun & Kim, Jongwon & Pan, SungBum & Shin, JuHyun, "International Network Performance and Security Testing Based on Distributed Abyss Storage Cluster and Draft of Data Lake Framework. Security and Communication Networks," 2018, pp. 1-14. doi: 10.1155/2018/1746809.

[8] 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.

[9] 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.

[10] Microsoft Azure cloud platform & services, https://azure.microsoft.com/en-us/, accessed: 2019-12-10.

[11] 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.

[12] V. Kleek, Max, and K. O'Hara. "The Future of Social is Personal: The Potential of the Personal Data Store.", in,Daniele Miorandi, Vincenzo Maltese, Michael Rovatsos, Anton Nijholt & James Stewart (eds.), Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society. Berlin, DE, Heidelberg, DE,Springer-Verlag, 125-158, 2014.