From Data Warehouse to a New Trend in Data Architectures – Data Lake

Elisabeta Zagan[†] and Mirela Danubianu[††]

[†,††]Faculty of Electrical Engineering and Computer Science, Stefan cel Mare University, Suceava, Romania
[†,††]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

Summary
Data Lake is a new concept of approaching and analyzing large volumes of different types of data, emerging with the evolution of technology and the new generation that came with new requirements, multiple resources and media information. In this paper, we will present the new Data Lake concept, highlighting the latest developments in the field. We also perform a critical analysis of the advantages and disadvantages currently offered by Data Warehouses, and at the end, comparing the two concepts we argue the answer to the question if Data Lake will replace Data Warehouse in the near future. In this context, our main contribution refers to a qualitative and comparative study on Data Lake and Data Warehouse, highlighting the advantages and improvements Data Lake brings to the storage of large data volumes.

Key words:
Data Lake, Data Warehouse, ETL, ELT, Schema on Read, Schema on Write.

1. Introduction

In recent years, the Data Lake concept has been increasingly used due to the explosive growth of data needed to be stored, processed, and analyzed as well as due to the increased need for information as a basis for good quality decisions. Globally, there has been an exponential increase in data volumes generated by activities in areas of the most diverse nature, from social media, to commercial, industrial, scientific or cultural activities, that it has become a real challenge to successfully manage this huge volume of data. In every company in every field everyone wants to extract as many advantages as possible by analyzing and studying the amount of information they have at their disposal.

Lately, industry leaders have been looking for the most efficient and effective techniques to quickly and easily extract and process the necessary data for new strategies development and to get the most out of all the information they have.

Today's business leaders have understood that the data is the key to success because they allow to understand the needs of customers, competitors and markets. Only by analyzing this information they can act and take the right decisions for a guaranteed success, minimizing as much as possible the errors in the strategies and development decisions they have to take at some point to achieve success and existence on the market.

So, some time ago, the Data Warehouse concept has emerged. This has been a big step forward and has led to an increase in efficiency with which valuable information is obtained, making important contributions to data access and data transfer simplification, as well as to combine data from multiple sources, to answer all the questions that an organization or a company may have. But even so it is impossible to anticipate any question or requirement of a report that a company might need at some point in its process of development and refinement. Statistics may vary from year to year, from one month to another and sometimes even from one day to the next.

In addition to these shortcomings, a series of new data types coming from web, social media, servers, sensors, applications and various devices have been generating a real explosion in the amount of data, which organizations are struggling to store, understand and process. An example would be that 15 to 20 years ago, companies did not expect that in the future, it would be necessary to keep control of social media's “likes”.

In a traditional Data Warehouse, most organizations will probably ignore data from a variety of sources, as volumes are too high, storage costs may become prohibitive, and data may have different types and forms becoming difficult to integrate and manipulate. These limitations lead to the permanent loss of valuable information that will definitely lead to substantial financial losses within companies.

So in recent years, we have been witnessing the emergence and development of new concepts and technologies, Data Lake being one of them since 2014.

Lately, there have been more and more controversial discussions about the two Data Lake and Data Warehouse technologies, discussions that attempt to argue which of these technologies are best for storing large volumes of data. In what follows we will make a comparative study between the two technologies trying to provide an answer.

The first section of the paper contains a brief introduction, highlighting the contributions of the authors. Section 2 presents main characteristics of a Data Warehouse,

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

whereas Section 3 addresses the defining features of a Data Lake. Section 4 makes a comparative analysis of the two and argues the main features, such as data scaling, data types, data structure, storage, accessibility, security and data applicability. Section 5 focuses on discussions regarding the other similar projects published in the literature. The paper ends with the final conclusions in Section 6.

2. Data Warehouse

According to Inmon [1], Data Warehouse can be defined as "a collection of integrated, subject-oriented databases designed to supply the information required for decision-making". In a Data Warehouse there are collected data from many operational systems of the enterprise. Often these are completed by data from external sources. A Data Warehouse targets enterprise-wide subjects such as customers, sales and profits. These subjects go beyond the boundaries of a single process requesting data from multiple sources for a complete picture.

Data Warehouse is not a product, but an environment that combines several components into a system architecture which provides current and historical information, almost impossible to find in traditional business databases. The general architecture of a Data Warehouse is presented in Figure 1.

An important feature of Data Warehouse is that it stores both detailed data and data on different levels of summarization. The contents of a Data Warehouse is described by metadata, data that are used to represent other data. These metadata are the summarized data that leads us to detailed data and they are stored in the warehouse as well.

Because they come from multiple and very different sources, the data may be inconsistent and of low quality. In the first development stage of a Data Warehouse, a vast amount of time is directed to data sources analyze and data understand in order to accomplish all those operations that produce unitary and well-structured datasets for the warehouse. The Extraction, Transformation, and Loading (ETL) processes include operations for data unification, integration, cleaning and data transferring from data sources into Data Warehouse [3]. The traditional Data Warehouse architectures are based on a well-structured scheme, designed and implemented before uploading the Data Warehouse. The ETL process is the one through which, data of different types are brought to the right shape to be loaded into the warehouse. This technique is called Schema on Write.

Throughout the time, with the increasing data volumes, this method has proven to be no longer sufficient and as a consequence, experts have focused on a new approach, namely Schema on Read. According to the Schema on Read technique, a schema is created only when the data is read. Data is structured only after they are read, allowing the storage of unstructured data. Since it is not necessary to a priori define a scheme, it is easy to introduce new data source at any time. This new type of schema approach has come together with a new data-saving and data-processing technology, the Data Lake [5].

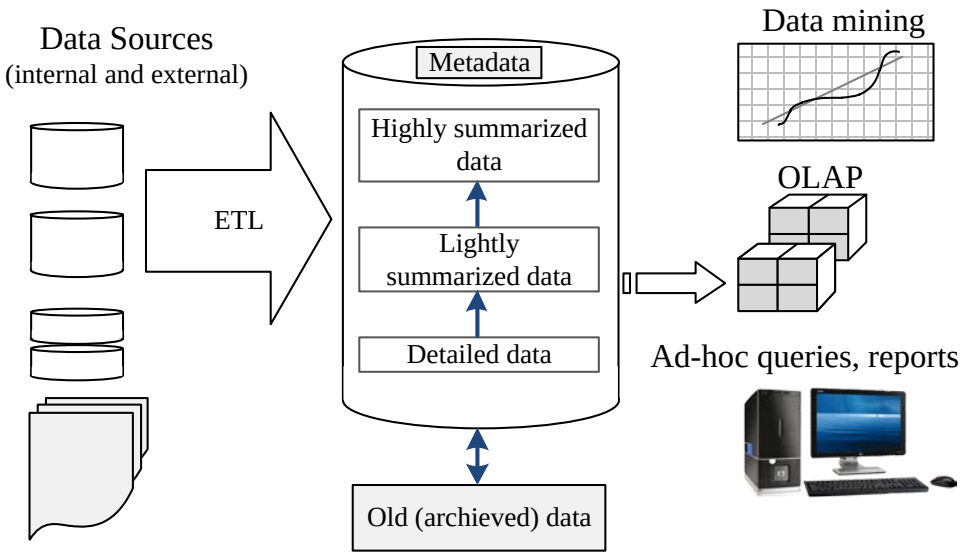

Fig. 1 The general architecture of a Data Warehouse [2]; OLAP - Online Analytical Processing

3. Data Lake

Data Lake, as the name says, can be seen as a lake with data in, a lake where will be stored all the data that reaches him. Data Lake is a modern data storage technology that saves data in their raw form for future use.

The data is ingested in its raw form so that different users can then access this data lake in order to analyze, structure and finally process them. Once it becomes operational, users can access all of data from the Data Lake, and for this, different search engines have been developed, search engines that are specifically designed to interrogate the data found in the Data Lake, so that the data structure is very flexible.

Data Lake has the ability and power to store all types of data even if they have or not a major importance in the initial data analysis process. Within a company, requirements may vary from year to year, and in the case of Data Lakes with all the data stored in their raw format from the very beginning of the development process, it will have the ability to respond to any report in the future. Also, data are stored right from the start so organizations can go back to any point in the past and make different analysis on old stored data.

This approach is also possible due to the storage hardware technologies. Data Lakes have a cheap storage technology that makes storage space scaling from terabytes to petabytes to be fairly economical.

As we have noted above, Data Lakes use Schema on Read technique which offers more flexibility for using large volumes and different types of data.

Schema on Read technique uses Extract, Load, Transform (ELT) processes. These involve data integration to be transferred in a raw state from a source server to a Data Lake located on a target server, followed by preparing them for future use.

Taking into account the market demands, the experts in field have found this method of saving data in their most natural form even at the risk of being more difficult to recognize. The information lost through the Schema on Write are proved to be precious information, so the experts sought to avoid losing them. By this new method it was shown that the data can be stored in their raw form to be formatted or structured when necessary in a certain analyzing process.

Table 1 presents a brief comparative study between the two integration processes: ETL and ELT.

Table 1: Comparative study between ETL and ELT

	ETL Extract Transform Load	***ELT Extract Load Transform***
Processing Data	i. Evaluate and extract data. ii. Define data structure applying business rules. iii. Validate and clean data. iv. Store accurate data.	Collect and store raw data
Client Data	Answering business reports	i. Evaluate data. ii. Define and apply different data schema. iii. Answering business reports.
When to use	Small and medium business organizations	i. Available to all business sizes. ii. Ideal for larger volume of data.

In Data Lakes, because the data are saved in their raw form they can be accessed at any time by any user who can explore these data in different ways, so that will always find the correct answer and in the shortest possible time without being limited by a rigid and well-established structure as with Data Warehouses.

At the same time, if the answers are useful and it is thought that they could still be requested in the future, then a scheme can be created, that leading to some kind of automatism for the future requirements, so this scheme can be reused as often as needed. On the other hand, if the obtained answers proves to be irrelevant and uninteresting, then they can be deleted without having made any changes to the structure and without having spent developing times. Figure 2 shows the architecture of a Data Lake at the conceptual level.

The companies that have used other analyzing techniques can easily move to the new, modern method because it does not replace old methods of analysis but simply it complements them.

By building up a modern data architecture, organizations can continue to capitalize on their existing investment in analysis by collecting all the data that they have ignored or that they have been constrained to delete so far, all of this allowing analysts at the same time to obtain data and business information faster.

Designing or implementing a Data Lake is in process of development. Research on Data Quality, Data Security, Data Life Cycle and Data Gravity approaches are still in progress.

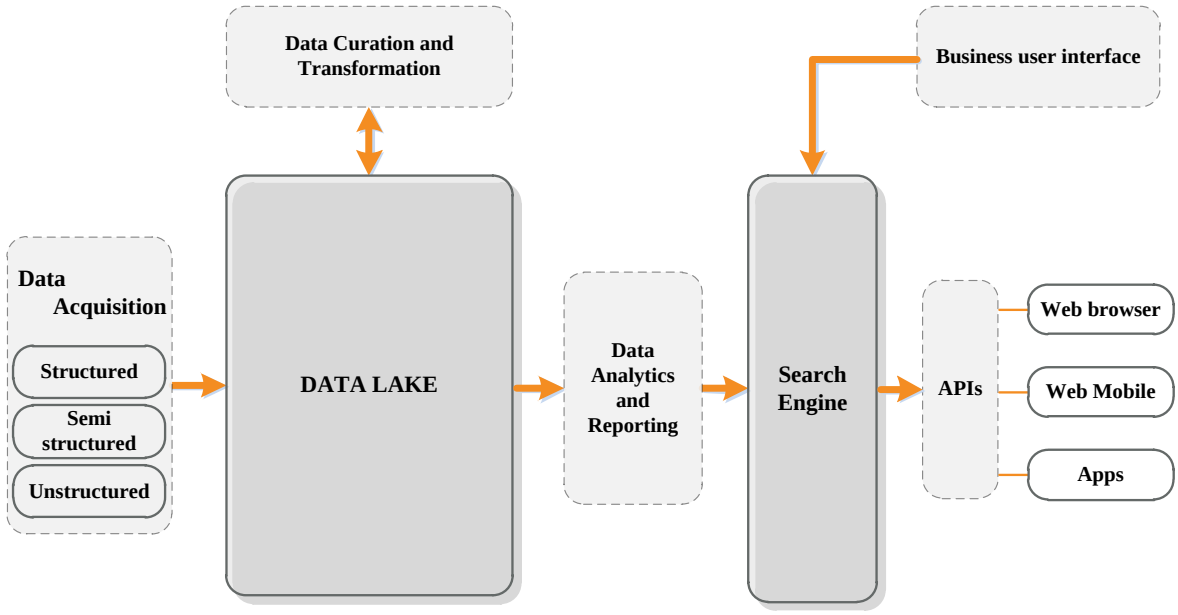

Fig. 2 Data Lake Concept

4. Data Lake vs. Data Warehouse – a comparative analyze

Data Warehouse is a technology that can be defined as a data repository integrated from one or more different sources, and these data are very carefully cleaned and checked before being stored to determine the integrity of the data. Data found in a Data Warehouse are structured and used on very precise purposes within a company.

Here are some of the most important features of a Data Warehouse:

i. Data is very well organized and structured.
ii. The data is properly cleaned of inaccuracies, corrupted data, duplicate data or discrepancies.
iii. Data are not stored unless their purpose and use is well defined.
iv. All company data are stored in one place respecting the same data format.

The process used to extract data from different data sets is the ETL (Extract - Transform - Load) process that ensures data accuracy (Table 1).

Because of its integrated nature, a Data Warehouse spares business users from the need to learn, understand, or access operational data in their native environments and data structures [2].

The main features of a Data Lake are:

i. More flexibility, a major plus in that you do not have to hold all the answers before doing an analysis.
ii. Ability to store raw data so that the algorithm of data selection and analysis can always be improved by finding new and new answers.
iii. Unlimited data query methods.
iv. The possibility of using different tools to get a perspective on the data.
v. Elimination of data silos.
vi. Using an efficient data management platform.
vii. Uses the ELT (Extract-Load-Transform process for ingesting and processing data.
viii. Possibility to obtain results from unlimited data types.
ix. Possibility to store all structured, semi-structured, unstructured data in one place.

The main challenge with a Data Lake, to make data useful, mechanisms for cataloging and securing data have to be implemented. Without these elements, the data can't be found or secured, eventually leading to a "data swamp". In order to meet the needs of the general public, data lakes need to have governance, semantic consistency and access controls.

In Table 2, the main features, such as data scaling, data types, data structure, storage, accessibility, security and data applicability of Data Warehouse and Data Lake are taken into consideration and contrasted.

Table 2: Data Warehouse versus Data Lake

	Data Warehouse	*Data Lake*
Scale	**Medium/Large data volume** at a moderate cost	**Very large data volume** at a low cost
Data	**Accurate data** (structured, semi-structured)	**Raw data** (structured, semi-structured, unstructured data, binary data, transactional systems data, sensors data, application activity data)
Schema	**Schema on Write** (ETL process), schema being defined before data is stored	**Schema on Read** (ELT process), schema being defined after the data is stored
Storage	**Costly** depending on the needs	**Low-cost** storage with very large volume of data
Agility	**Time-consuming** on making changes to the data structure, being a highly structured data bank. **Less agile**, having a fixed configuration.	**Ability to easily change** data models and queries. **Highly agile**, can be configured and reconfigured whenever necessary
Security	**High security** performance	**Still in process**, new techniques in development
Use	Mostly in the **business industry**	Mostly used in **scientific fields** and **business** with large volume of data

5. Related works

D. E. O'Leary attempted to examine the concept of Data Lake by contrasting it with the existing solutions, taking the Data Warehouse as an example [4]. As it is defined in [4], DataLake uses a classic database approach that consists in allowing users to access original databases coming from multiple sources.

As a unique data repository of an organization, Data Lake is conceptually speaking in contrast with the classic Data Warehouse, which extracts, transfers and loads (ETL) data. The author summed up some of these differences in the article [5].

In the article [6], the authors approach the transfer from the Data Warehouse to Data Lake pointing out the advantages and possibly the issues that a modern Data Lake architecture can raise, trying to outline the definition of Data Lake at the end of the article.

Data Warehouse is considered by the author to be the only solution to provide accurate and reliable data across large organizations. And yet Data Warehouse, with the emergence of the Internet Of Things (IoT) paradigm, the use of Smartphones and various applications, has been overtaken by the impossibility of storing so much information, as well as being of different data types.

Managing and preparing them to be saved has become difficult to manage. The potential of all these new data emerged lately is not yet well known and not well explored and analyzed. And here a new concept emerged, namely Data Lake. In the definition proposed in [6] for Data Lake, the authors want to present theoretical solutions about data governance principles as the key principles of the Data Lake definition.

6. Conclusions

By analyzing the two data storage methods discussed earlier, we can say with certainty that there is no competition between Data Lake and Data Warehouse, as they are two completely different things, even if they seem to serve the same requirements, Data Lake is a much more versatile technology and with larger resources than a Data Warehouse.

If we look more closely at the rapid evolution of IoT, Data Lake is becoming more and more the optimal solution for storing data from this branch of technology. A Data Lake makes it easy to store and run analytics on IoT data to discover ways to reduce operational costs and increase quality.

Each of the two technologies will serve for well-defined purposes so that both will continue to occupy an important role in various fields or processes of activity. Choosing a technology to the detriment of the other will always depend on a thorough process of analyzing the advantages and disadvantages that each of them presents in order to best serve the intended purpose.

Acknowledgments

This work was partially supported from 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] W. H Inmon. "Building the Data Warehouse," QED Technical Publishing Group, 1992, ISBN: 0-89435-404-3.

[2] 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.
[3] M. W. Humphries, M. W. Hawkins, and M. C. Dy. “Data Warehousing: Architecture and Implementation,” Prentice Hall, 1999, ISBN-10: 0130809020.
[4] D. E. O’Leary. “Embedding AI and Crowdsourcing in the Big DataLake,” in IEEE Intelligent Systems, vol. 29, no. 5, pp. 70-73, Sept.-Oct. 2014. doi: 10.1109/MIS.2014.82.
[5] C. Madera, and A. Laurent. “The Next Information Architecture Evolution: The Data Lake Wave,” Proceedings of the 8th International Conference on Management of Digital EcoSystems (MEDES), pp. 174-180, Biarritz, France, Nov. 2016. doi: 10.1145/3012071.3012077.
[6] H. Fang, “Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem,” 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, China, 2015, pp. 820-824. doi: 10.1109/CYBER.2015.7288049.

Elisabeta Zagan received the M.Sc. degree in Automation and Applied Informatics from the Stefan cel Mare University of Suceava, Suceava, Romania, in 2005. He is currently a Ph.D. student at the Department of Computers of Stefan cel Mare University of Suceava. His research interests include databases, Big Data, Data Lake and Data Warehouse.
Address: Str. Universitatii nr. 13, 720229, Suceava, Romania.
E-mail: elisabeta.b@gmail.com.

Mirela Danubianu received the M.Sc. degree from the Faculty of Electrotechnics Craiova, Department of Automation and Computers in 1985, and PhD from the Department of Computer Science, Stefan cel Mare University of Suceava in 2006. Mirela Danubianu is currently Associate Professor and head of Computers Department of Stefan cel Mare University of Suceava, Romania, with more than 10 years of teaching and research experience.
Address: Str. Universitatii nr. 13, 720229, Suceava, Romania.
E-mail: mdanub@eed.usv.ro.