INTERNATIONAL CONFERENCE
INFORMATION AND COMMUNICATION TECHNOLOGIES IN BUSINESS AND EDUCATION

CONFERENCE PROCEEDINGS

Publishing house "Science and economics"
University of Economics - Varna

International Conference

INFORMATION AND COMMUNICATION TECHNOLOGIES IN BUSINESS AND EDUCATION

Conference proceedings

INFORMATION AND COMMUNICATION TECHNOLOGIES IN BUSINESS AND EDUCATION

Proceedings of the International Conference dedicated to the 50th anniversary of the Department of Informatics

2019

University publishing house "Science and economics"
University of Economics – Varna

Published reports have not been edited or adjusted. The authors are solely responsible for the content, originality and errors of their own fault.

With the support of ADASTRA, 411 Marketing Business Services.

ISBN 978-954-21-1004-0

PROGRAMME COMMITTEE

Abdel-Badeeh M. Salem, Egypt
Aijaz A. Shaikh, Finland
Aleksandar Dimov, Bulgaria
Alexander Feldman, USA
Alexander Sergeev, Russia
Antoanela Naaji, Romania
Avram Eskenazi, Bulgaria
Dariusz Mrozek, Poland
Dimitrina Polimirova, Bulgaria
Dmitriy Berg, Russia
Elena Nechita, Romania
Emil Denchev, Bulgaria
Georgi Dimitrov, Bulgaria
Ivan Popchev, Bulgaria
Jeanne Schreurs, Belgium
Julian Vasilev, Bulgaria
Kameliya Stefanova, Bulgaria
Krasimir Shishmanov, Bulgaria
Maciej Czaplewski, Poland
Maciej Pondel, Poland
Maksim Medvedev, Russia
Maria Hristova, Bulgaria
Marian Cristescu, Romania
Marina Medvedeva, Russia
Michael Voskoglou, Greece
Nadezhda Filipova, Bulgaria
Nikola Nikolov, Ireland
Niv Ahituv, Israel
Pavel Petrov, Bulgaria
Rami Malkawi, Jordan
Sara Paiva, Portugal
Silvia Parusheva, Bulgaria
Snezhana Sulova, Bulgaria
Todorka Atanasova, Bulgaria

Valentin Kisimov, Bulgaria
Valery Trofimov, Russia
Veselin Popov, Bulgaria
Violeta Kraeva, Bulgaria
Vladimir Sulov, Bulgaria
Vladimir Zanev, USA
Zdzislaw Polkowski, Poland

ORGANIZING COMMITTEE

Chairman: Prof. Vladimir Sulov, PhD

Members:
Prof. Avram Eskenazi, PhD
Prof. Julian Vasilev, PhD
Assoc. Prof. Todorka Atanasova, PhD
Assoc. Prof. Silvia Parusheva, PhD
Assoc. Prof. Snezhana Sulova, PhD
Assoc. Prof. Pavel Petrov, PhD

Secretary: Tijen Talib

FOREWORD

These proceedings contain the papers of the International Conference "Information and Communication Technologies in Business and Education" which took place at the University of Economics – Varna, Bulgaria, 18 October 2019.

The international scientific conference is dedicated to the **50th anniversary of the Department of Informatics at the University of Economics – Varna.** The conference is also dedicated to the 100th anniversary of the University. The included papers describe recent scientific and practical developments in the field of information and communication technologies, information systems, and their applications in business and education.

The papers in the Proceedings are peer reviewed and are checked for plagiarism.

C O N T E N T S

THE USAGE OF DATA LAKE FOR BUSINESS INTELLIGENCE DATA ANALYSIS

***Snezhana Sulova*[1]**
[1] University of Economics – Varna/ Department of Informatics, Varna, Bulgaria, ssulova@ue-varna.bg

Abstract

Data analysis is now becoming increasingly more important for business. The accumulation of large amounts of different types of data in organizations is a prerequisite for seeking new ways of storing, processing and analyzing them. The following paper presents the nature of the data lake concept and examines its capabilities to organize all the data, both those generated by the organization and those extracted from Internet sources. Storing large amounts of data, regardless of its type, structure, or format, allows for the integrated use of structured and unstructured data and the application of a variety of techniques for intelligent business analysis.

Keywords: *data lake, big data, data analysis, BI, data warehouse.*

INTRODUCTION

Today's dynamic society is increasingly more tied to internet technologies and intelligent systems. The advancement in information technologies and the development of artificial intelligence are leading to the creation of new ways of working and remotely accessing devices. More and more physical devices have built-in electronic elements, software that allow them to be connected to the Internet and receive, collect and exchange data. Although organizations currently work mainly with information systems in which the data is organized and managed by database management systems (DBMS), according to research, unstructured data makes up about 80% of all the information resources in them (Grimes, 2008). A research on the topic of big data indicates that unstructured data tends to grow exponentially in number and that they represent 95% of new data, a large part of which is not processed or used (Minelli, et al., 2013, p. 11).

It is well known that in order for data to be converted into information and business knowledge, it must be transformed and organized

with a specific purpose. The analysis process is becoming increasingly more complex and involves not only the generation of reports using SQL queries and the calculation of statistical dependencies, but also data mining technologies. Another tendency has been noted which involves looking for and adding new additional sources of data for processing, as well as improving the approach and technologies used to store and retrieve this data. Business intelligence (BI) is a top priority for the organizations in most industries (Richards, et al., 2019). In regard of that, this report aims to clarify the essence of the data lake (DL) data storage concept and to demonstrate its capabilities for supporting smart business analyses.

1. THE ESSENCE AND CONCEPT OF DATA LAKE

James Dixon, CTO of Pentaho introduces a new big data storage concept called data lake (Dan, 2011). The main idea is to store different types of data in a relatively cheap way and then to apply ETL functions (extraction, transformation, loading) to them. "A data lake is a central location in which to store all your data, regardless of its source or format" (Laplante and Sharma, 2016, p.2). IBM researchers say the concept is designed to provide storage of virtually inexhaustible materials in the form of raw data that analysts have easy access to (IBM Corporation, 2016).

DL can be defined as a data storage strategy that provides flexibility for organizations when working with data. It allows the same data to be structured and processed differently, and this is essential for processing unstructured data where there are no well-defined algorithms for data extraction, processing and analysis, and different approaches are usually used.

The DL concept also allows data to be stored from external Internet sources, such as social networks, devices using the Internet of things (IoT) concept, and other unstructured corporate data (Fig. 1).

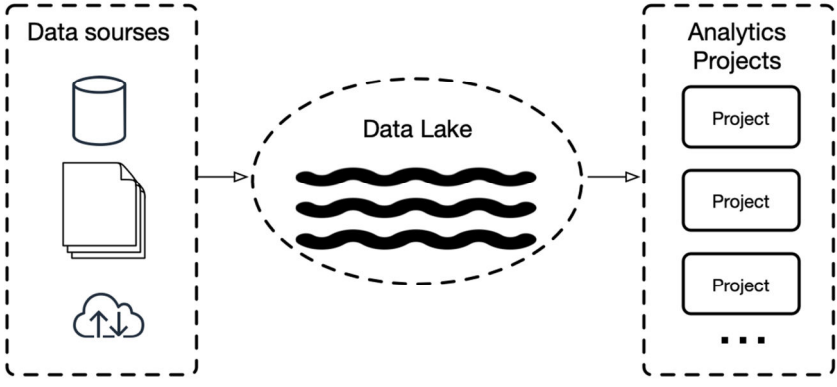

Figure 1. A common view of data lake

The following main advantages of the data lake concept can be pointed out:

- data from different types – structured, semi-structured and unstructured, can be stored;
- the types of data than can be extracted are an infinite amount;
- data can be stored in its raw form, which allows its conversion when it is needed;
- various tools can be used for extracting and processing the data;
- all data in the organization can be stored in one place, which allows a unified view of the data.

There are also several challenges associated with data lake, which is why there is still debate in literature whether this concept is necessary and useful for business. Some of the concerns connected to it include the following:

- a bad quality of data because they are received without supervision and control;
- the data process needs to be started from scratch with every data analysis;
- the performance of data operations is usually not guaranteed;
- weaknesses regarding the security and control over the access to the data;

- a danger of the data being turned into a "swamp" because they are stored without being sorted and organized into topics, categories and without maintaining metadata for them.

It should be noted that the use of DL cannot replace the traditional data warehouse (DW), which aims to integrate large-scale enterprise data into a united storage. According to Bill Inmon, DW "is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process." (Inmon, 2002, p. 31). Usually DWs are subject oriented. It is necessary to select and extract the data from the company databases and then transform, reformat and store it (Curtis and Cobham, 2008, p. 247). Therefore, creating a data warehouse takes time and resources in order to model and prepare the data. The knowledge and skills of the specialists are also important for its successful creation (Marinova, 2016). The benefits of data warehouses are numerous: they save time for users, improve decision-making processes, and help achieve strategic business goals.

Organizations looking to upgrade their analytics platforms could use both the DL and the DW storage concepts. This working method will allow them to explore how traditional analytical architectures work with new storage methods that include both relational databases and NoSQL databases. According to research, for every type of DBMS, there is an effective strategy for archiving and restoring databases, which is of crucial importance (Kuyumdzhiev, 2019).

2. DATA LAKE AND BUSINESS INTELLIGENCE DATA ANALYSIS

Business intelligence can be viewed as a generic term for a set of approaches that serve to analyze the activities and functioning of an organization and support the decision-making process (Curtis and Cobham, 2008, p. 228). The goal of BI is to enable the processing of this large volume of data easily, to support the search of new opportunities for development and to build effective knowledge-based business strategies.

Business intelligence systems are constantly evolving, new functionalities are being added to them (Todoranova, 2013), they are evolving from single applications to large-scale business Intelligent Ecosystems (Kisimov and Stefanova, 2010). Their main components are: DW; ETL tools, OLAP techniques and data mining tools.

Trends in the development of BI systems indicate that the application of the data mining technology will expand and will be applied more and more to unstructured data (cio.bg, 2018). The World Wide Web has become one of the richest sources of data. Companies have started using data mining technologies more to extract knowledge from Internet sources or the so-called Web Mining to increase the precision of their business analytics. There are many documents, data, audio and video files on the web that can be used to extract new and useful business knowledge through appropriate processing. The knowledge is generated not only by the content of the web pages themselves, but also by their unique features, the structure of the web sites and the information connected to accessing them.

BI methods are usually applied when analysing structured datasets from a specific type of business, e.g. banking or credit institutions (Vasilev, Stoyanova, Stancheva, 2017, 2018). Time series data are quite popular for storing data in different types of businesses. In these cases, BI methods are applied to data stored in relational databases. But the real view of business needs more information on opinions, moods. In this case data from data lakes are useful. This fact shows the need to create, test and apply BI methods for analyzing data stored in data lakes.

The data extracted from the Internet are in most cases unstructured, and performing an automatic analysis, generating summaries, classifications, trending and anomalies requires that they be pre-processed and given a certain structure. Processing unstructured data is not an easy task (Bankov, 2018). It is very often necessary to use different approaches that require different sections of the data. Business analytics also require rebuilding of business rules and the need to use unconverted or so-called raw data. Therefore, we believe that creating and maintaining a data lake data storage in the case of large heterogeneous data is a good base for modern BIs that are focused on providing machine learning, NLP and AI to their customers.

Using DL together with DW offers a modern and optimal basis for data analysis, as shown in Figure 2. Data from enterprise databases that are collected as a result of multiple applications in enterprise information systems are the basis for the creation of a DW, which aims to store and track historical, archival information, consolidate large volumes of data

from various subsystems, analyzes and forecasts. Unstructured and semi-structured data from other external Internet sources, such as social networks and devices using the Internet of things (IoT) concept, server log files, and more are entered and stored in a DL. In this way, the necessary data will be provided for each analytical process. Integrated data will also be able to be used, which is a good basis for obtaining more detailed and in-depth analyzes and will help to make informed decisions.

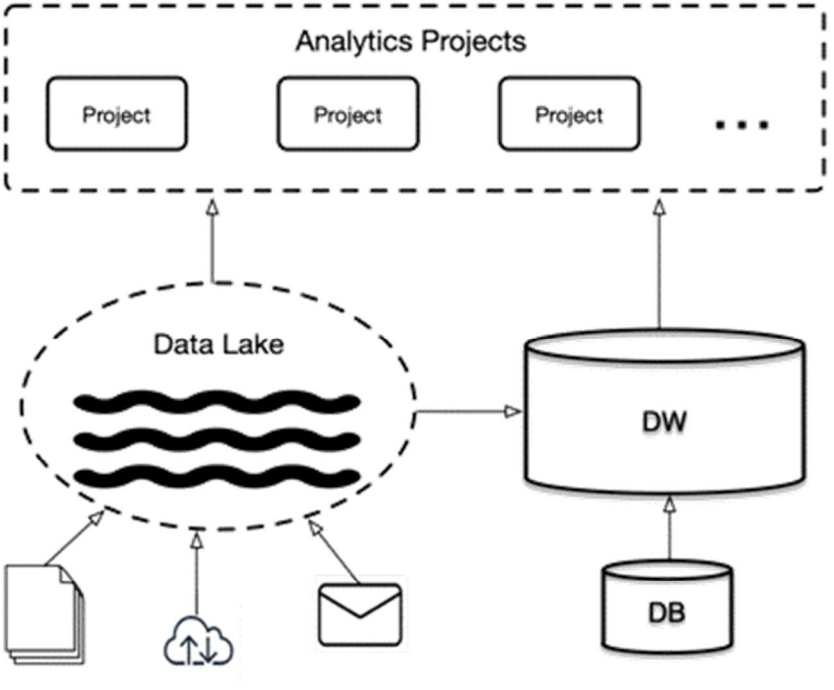

Figure 2. Use of DL and DW

The main advantages of the BI approach based on the combined use of DW and DL are the following:

- in addition to storing traditional structured data, DL allows the cheap storage of all types of data (including audio and video formats) coming from Internet applications, social networks and from various devices;

- various and optimal data analysis approaches can be used, including those that work with unstructured data – processing text, audio, video;
- enables real-time data extraction, rapid data analysis and the implementation of agile analytics schemes;
- allows working with large volumes of data, for example, which is a good basis for predicting possible future states and processes.

Although as mentioned above, there are a number of problems with the use of DL, it is considered that creating a single integrated data management framework, where they are managed with metadata that helps to find and connect information, can build a successful model for integrated data storage and management. Such a model is a good basis for conducting numerous analyzes and improving the BI strategies of companies. It would also allow the implementation of Agile BI, which is built on the idea of flexible analysis and adaptation to specific needs and is responsive to rapidly changing business conditions. A proper implementation of the DL concept would be in favor of adhering to one of the basic principles of Agile BI, which is to provide the right data at the right time for the correct analytical process.

It is proved that good organization of projects in the first phases of product development reduces errors and costs of subsequent stages (Nacheva, 2015).

We believe that the main conditions for the successful implementation of DL as a data source are the following

- clearly defining the need and purpose of using the DL;
- creating and following a data management strategy in DL;
- creating procedures for security and control over the access and use of the data;
- building DL as a new, additional source of data for analysis, rather than as a sole component of BI infrastructure.

Hadoop's distributed file system (High-availability distributed object-oriented platform) is currently considered the most popular DL build technology (Khine and Wang, 2018; Yordanova and Stefanova, 2019). Hadoop is a software framework managed by the Apache Software Foundation and designed to organize a distributed processing of big data

when using the MapReduce programming model. It is used to build large-scale projects on Yahoo!, Facebook, Oracle, in IBM's Watson supercomputer, and in Azure Cloud.

CONCLUSION

Growing volumes of heterogeneous data are a prerequisite for finding ways to extract, process and store them. Combining data across systems represents a challenge for many organizations when making management decisions. Using the DL concept can bring benefits for organizations that want to avoid the costly and cumbersome process of preprocessing storage data in a data warehouse. DL is a good approach for storing big data, but it should be noted that an incorrect design and use of it carries risks related to data quality, security and control of their access and usage. Well-trained professionals are needed to properly anticipate and plan the data lake so that it can help organizations successfully manage structured and unstructured data.

REFERENCES

1. BANKOV, B. (2018) An Approach for Clustering Social Media Text Messages, Retrieved from Continuous Data Streams. *International Scientific Journal "Science. Business. Society"*, 3(1), pp. 6-9.
2. CIO.BG (2019) *11 тенденции в развитието на платформите за бизнес анализи*. [Online] Available from: https://cio.bg/ softuer/2008/05/12/3450655_11_tendencii_v_razvitieto_na_platformite_ za_biznes/ [Accessed 31/07/2019].
3. CURTIS, G. and COBHAM, D. (2008) *Business Information Systems: analysis, design and practice*, sixth edition, Canada: Pearson Education.
4. DAN, W. (2011) *Big Data Requires a Big, New Architecture* [Online] Available from: https://www.forbes.com/sites/ciocentral/ 2011/07/21/ big-data-requires-a-big-new-architecture/#37aa44d31157 [Accessed 22/07/2019].
5. GRIMES, S. (2008) *Unstructured data and the 80 percent rule*. [Online] Available from: https://breakthroughanalysis.com/2008/08/01/ unstructured-data-and-the-80-percent-rule/ [Accessed 22/07/ 2019].

6. IBM Corporation (2016), *IBM Industry Model support for a data lake architecture*. [Online] Available from: https://www.ibm.com/downloads/cas/DNKPJ80Q, [Accessed 22/07/2019].
7. INMON, B. (2002) *Building the Data Warehouse*. 3rd Ed. Toronto: John Wiley & Sons, Inc.
8. KHINE, P. and WANG, Z. (2018) Data lake: a new ideology in big data era, *ITM Web of Conferences 17*, WCSN 2017, [Online] Available from: https://www.itm-conferences.org/articles/itmconf/abs/2018/02/itmconf_wcsn2018_03025/itmconf_wcsn2018_03025.html [Accessed 20/07/2019].
9. KISIMOV, V. and STEFANOVA, K. (2010) Design Principles and Methods for Distributed Business Intelligence Ecosystem, *Yearbook of UNWE, 1,* pp. 215-250.
10. KUYUMDZHIEV, I. (2019) Comparing Backup and Restore Eficiency in Mysql, MS SQL Server and MongoDB. *19 International Multidisciplinary Scientific Geoconference SGEM 2019*, Albena, Bulgaria, 19(2.1), pp. 167-174.
11. LAPLANTE, A. and SHARMA, B. (2016) *Architecting Data Lakes. Data Management Architectures for Advanced Business Use Cases*. O'Reilly Media, Inc.
12. MARINOVA, O. (2016) Business intelligence and data warehouse programs in higher education institutions: current status and recommendations for improvement. *Econimics and computer science: Electronic journal*, 5, pp. 17-25.
13. MINELLI, M., CHAMBERS, M. and DHIRAJ, A. (2013) *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Jonh Wiley & Sons, Inc.
14. NACHEVA, R. (2015) Principles of User Interface Design: Important Rules that Every Designer Should Follow. *Izvestia, Journal of the Union of Scientists – Varna. Economic Sciences*, pp. 140–149.
15. RICHARDS, G., YEOH, W., CHONG, A. Y. L. & POPOVIC, A. (2019) Business Intelligence Effectiveness and Corporate Performance Management: An Empirical Analysis. *Journal of Computer Information Systems*. 59(2). pp. 188-196.

16. TODORANOVA, L. Trends in the Development of Business Intelligence Systems. *Izvestia – Journal of University of Economics – Varna*, 1, pp 98-106.
17. YORDANOVA, S. and STEFANOVA, K. (2019) Big Data Challenges – Definition, Characteristics and Technologies, *The Scientific Papers of UNWE,* 1, pp. 13-31.
18. VASILEV, J., STOYANOVA, M., STANCHEVA, E. (2017) Business intelligence data analysis of a loan dataset. In *2nd conference on innovative teaching methods (ITM 2017)*, Varna: Science and economics, pp. 17-23.
19. VASILEV, J., STOYANOVA, M., STANCHEVA, E. (2018) Application of business intelligence methods for analyzing a loan dataset. *Informatyka Ekonomiczna*, 1, 47, pp. 97-106.