Serdica J. Computing **13** (2019), No 3–4, 171–182

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

AN APPROACH TO STORING DATA BASED ON THE DATA LAKE CONCEPT TO FACILITATE INTELLIGENCE DATA ANALYSIS

Snezhana Sulova

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 article presents the nature of the data lake concept and an approach for organizing and storing all the data, both those generated by the software systems in the organizations and those extracted from Internet sources. The approach is based on the combined usage of the concepts data lake and data warehouse and allows the storage of large amounts of data, regardless of its type, structure, or format, and allows for the integrated use of structured and unstructured data and the application of a variety of techniques for intelligent business analysis.

1. 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

ACM Computing Classification System (1998): H.2.1, I.2, K.4.4.
Key words: data lake, big data, data analysis, BI, data warehouse.

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 [1]. Another 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 [2, 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 [3]. In regard to that, this article aims to clarify the essence of the data lake (DL) data storage concept and to demonstrate its capabilities for combined usage with data warehouse for supporting smart business analyses.

2. The essence and concept of a data lake. James Dixon, chief technology officer (CTO) of Pentaho introduces a new big data storage concept called data lake [4]. 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” [5, 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 [6].

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. A lot of research shows that more and more organizations are starting to grasp the significance of DL for generating value from data [7, 8].

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).

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

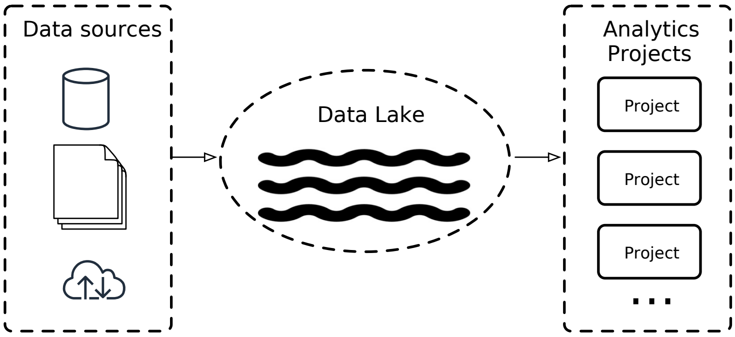

Fig. 1. A common view of data lake

- 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 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:

- low quality of the 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.

Due to the increasing amount and variety of data when creating and maintaining data lake, the biggest concerns are that it may at some point become a large set of hard-to-use or unusable data or become the so-called "data swamp". For the data lake, the quality of the incoming data is essential, and it is recommended that it be generated in a form that facilitates data comprehension and use in applications. Data that is stored without any thought as to how or what it would be used for is often redundant and unusable. Several authors have considered the importance of creating an appropriate semantic data structure, as well as retaining metadata for easy access to lake data retrieval [5, 9].

The main actions which need to be undertaken to avoid turning the data lake into a “data swamp” are:

- clearly defining the business objectives and limiting the data at the entrance to the data lake only to the necessary for analysis goals;
- data management, use of metadata in order not to start the analysis from scratch every time and to avoid an overflow with requests for data extraction.

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” [10, 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 [11, 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 [12]. 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 [13].

3. 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 [11, 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 [14], they are evolving from single applications to large-scale business Intelligent Ecosystems [15]. Their main components are: DW; ETL tools, online analytical processing (OLAP) techniques and data mining tools.

BI methods are usually applied when analysing structured datasets from a specific type of business, e. g., banking or credit institutions [16, 17]. 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, which is extracted from internet sources, different sensor devices and contributes to sentiment analysis and the study of customer opinions and 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.

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 [18]. 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.

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 [19]. 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 Business intelligence systems that are focused on providing machine learning, natural language processing and artificial intelligence to their customers.

Using DL together with DW offers a modern and optimal basis for data analysis, as shown in Fig. 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, analyses 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

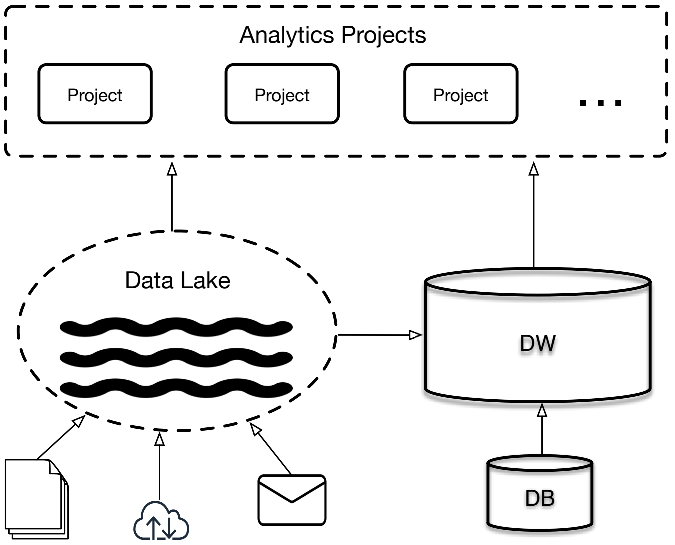

Fig. 2. Use of DL and DW

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.

The main advantages of the BI based on the combined use of DW and DL can be summarized in the following way:

- 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, 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 analyses 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.

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.

4. An approach for the combined usage of the data lake and data warehouse. Although as mentioned before the data lake concept poses a number of challenges, we believe that due to the need for accurate analysis and reliable forecasts based on more information sources, the integrated use of heterogeneous data is needed. In many cases, information extracted from the database is known to be incomplete and data extracted from internet sources is required to perform a more detailed analysis. For example, to analyze the interests of a particular customer, one can examine their feedback, comments, which can be extracted from the content of the web pages, and at the same time the data about their purchases can be taken for analysis. In this way, the information obtained will be more accurate and complete, which will also allow for more in-depth analyses that will contribute to better personal service and generate more accurate recommendations for the client.

The approach for integrated usage of DW and DL can be summarized and shown in the following main steps (Fig. 3):

1. Defining the business goals of the analytical process. At this stage, opportunities are identified to gain new knowledge that can contribute to competitive advantage and business stability.
2. Creating a framework for integrated data storage and utilization. The ways and rules for organizing the necessary data are determined. The proposed

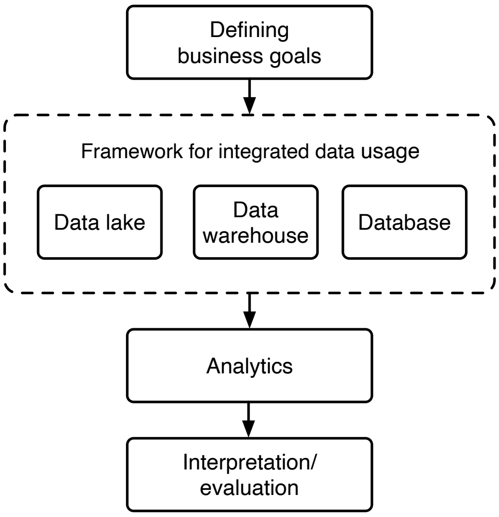

Fig. 3. The approach for integrated usage of DW and DL

framework should allow the application of modern forms of business analyses based on both structured and unstructured data.

3. Analysis of the data. Suggesting different methods to find the connections and dependencies between the data.
4. Interpreting the results and evaluation of the proposed analysis methodology.

The proposed approach makes it possible to create flexible models for data organization and management that are tailored to the information provision of each specific business. Business intelligence analyses, based on heterogeneous data sources, provide a significant advantage for companies. For example, e-commerce companies have large and diverse datasets that can be extracted for the purposes of machine learning and data analysis. DL would allow them to use data from wider datasets and queries in new ways for deeper analysis and more artificial intelligence applications.

With larger online retailers, online stores are usually integrated with other business applications. A common company database and the technology of data warehouses are used. This enables the storage and tracking of historical, archival information, the consolidation of large volumes of data from different subsystems and provides analysis and forecasting tools. At the same time, many data are generated in online commerce as a result of user interaction with e-commerce platforms, such as traffic to specific pages, recorded product reviews, or the purchase process. The specifics of analyses of semi-structured and unstructured data, such as data extracted from web pages and social networks that is obtained as a result of the use of Internet platforms, requires the storage of this data in raw format and the application of different methods of analysis. As is known for the processing of non-structured data, there are no established rules, but different algorithms are applied, and comparisons are made of the obtained results.

The foregoing leads us to conclude that smart business intelligence analyses based on the combined use of DL and DW concepts can be successfully used in e-commerce. Fig. 4 shows a model for implementing the approach of integrated use of DL and DW in e-commerce.

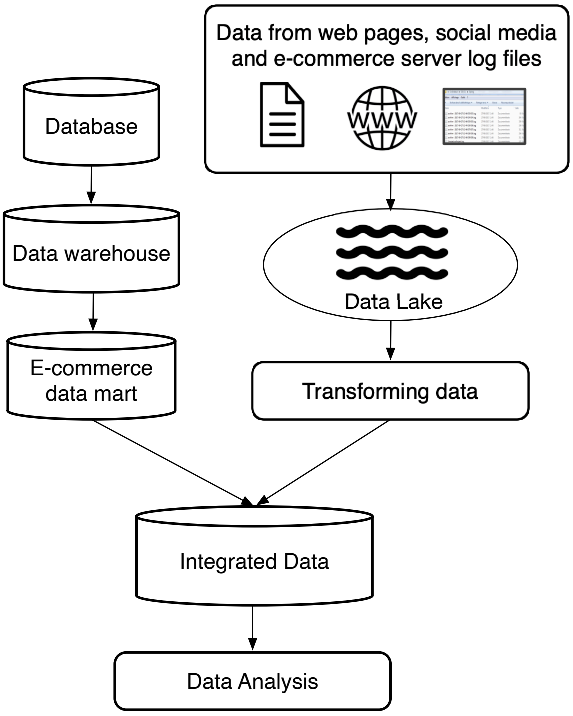

Fig. 4. A model for the integrated use of DL and DW in e-commerce

It can be seen from Fig. 4 that the data entering the company's data warehouse are extracted from the e-commerce system and other company software systems. A special data mart is created for the e-commerce data needed. The data lake contains mostly data from Internet sources, social networks, server log files, documents. If necessary, these data are processed and, after being assigned a structure, they are summarized with the data from the data warehouse. The integrated use of data enables the generation of many new business models and dependencies, which in most cases are a source for useful business knowledge [20].

The proposed model can also be used in other areas of business, where intelligent business analysis is based on multiple, heterogeneous data. In conclusion it can be said that successful implementation of the approach requires a data management strategy so that the following can be done properly:

- identify the necessary data collections and create a framework for organizing data in DL;
- choose appropriate methods for data transformation, since the data taken from the DL storage needs processing in order to obtain the necessary structure and to be suitable for automatic analysis;
- design and model an optimal data warehouse that contains the necessary data for the analyses.

In terms of software implementation of the approach, research shows that Hadoop's distributed file system (High-availability distributed object-oriented platform) is currently considered the most popular DL build technology [21, 22, 23]. 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 includes a collection of components for administering, accessing and analyzing structured and unstructured data. It is used to build large-scale projects on Yahoo!, Facebook, Oracle, in IBM's Watson supercomputer, and in Azure Cloud.

5. 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 pre-processing 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.

Developing a framework consisting of well-managed, protected and flexible data access mechanisms for data lake data will be a major part of the author's future research.

REFERENCES

[1] Grumes S. Unstructured data and the 80 percent rule. August 1, 2008. http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/, 22 July 2019.

[2] Minelli M., M. Chambers, A. Dhiraj. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. Wiley, 2013.

[3] Richards G., W. Yeoh, A. Chong, A. Popovic. Business Intelligence Effectiveness and Corporate Performance Management: An Empirical Analysis. *Journal of Computer Information Systems*, **59** (2019), No 2, 188–196.

[4] Dan W. Big Data Requires a Big, New Architecture. Forbes, July 21, 2011. https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/, 22 July 2019.

[5] Laplante A., B. Sharma. Architecting Data Lakes. Data Management Architectures for Advanced Business Use Cases. O'Reilly Media Inc., 2016.

[6] IBM Corporation. IBM Industry Model Support for a Data Lake Architecture. https://www.ibm.com/downloads/cas/DNKPJ80Q, 21 July 2019.

[7] Lock M. Angling for Insight in Today's Data Lake. https://www.ibm.com/downloads/cas/G69ZB92M, 21 July 2019.

[8] New Survey Reveals Businesses Are Bullish on Data Lakes. https://insidebigdata.com/2018/07/24/new-survey-reveals-businesses-bullish-data-lakes/, 21 July 2019.

[9] Hai R., S. Geisler, C. Quix. Constance: An Intelligent Data Lake System. In: Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16), San Francisco, California, USA, 2016, 2097–2100.

[10] Inmon B. Building the Data Warehouse. Toronto, Wiley, 2002.

[11] Curtis G., D. Cobham. Business Information Systems: Analysis, Design and Practice. Pearson Education, Canada, 2008.

[12] Marinova O. Business Intelligence and Data Warehouse Programs in Higher Education Institutions: Current Status and Recommendations for Improvement. *Economics and Computer Science*, **2** (2016), No 5, 17–25.

[13] Kuyumdzhiev I. Comparing Backup and Restore Efficiency in MySQL, MS SQL Server and MongoDB. In: Proceedings of the 19 International Multidisciplinary Scientific Geoconference (SGEM), **19** (2019), No 2.1, 167–174.

[14] Todoranova L. Trends in the Development of Business Intelligence Systems. *Izvestia. Journal of University of Economics—Varna*, **1** (2014), 98–106.

[15] Kisimov V., K. Stefanova. Design Principles and Methods for Distributed Business Intelligence Ecosystem. *Yearbook of UNWE*, **1** (2010), 215-250.

[16] Vasilev J., M. Stoyanova, E. Stancheva. Business Intelligence Data Analysis of a Loan Dataset. In: Proceedings of the 2nd Conference on Innovative Teaching Methods (ITM 2017), Varna, 17–23.

[17] Vasilev J., M. Stoyanova, E. Stancheva. Application of Business Intelligence Methods for Analyzing a Loan Dataset. *Informatyka Ekonomiczna*, **47** (2018), 97–106.

[18] CIO Media. 11 tendencii v razvitieto na platformite za biznes analizi. https://cio.bg/softuer/2008/05/12/3450655_11_tendencii_v_razvitieto_na_platformite_za_biznes, 31 July 2019. (In Bulgarian.)

[19] Bankov B. An Approach for Clustering Social Media Text Messages, Retrieved from Continuous Data Streams. *Science. Business. Society*, **3** (2018), No 1, 6–9.

[20] Sulova S. Integration of Structured and Unstructured Data in the Analysis of E-commerce Customers. In: Proceedings of the 18 International Multidisciplinary Scientific Geoconference (SGEM), **18** (2018), No 2.1, 499–506.

[21] Khine P., Z. Wang. Data Lake: a New Ideology in Big Data Era. In: Proceedings of the 4th Annual International Conference on Wireless Communication and Sensor Network (WCSN), **17** (2018), Article 03025.

[22] Yordanova S., K. Stefanova. Big Data Challenges – Definition, Characteristics and Technologies. *The Scientific Papers of UNWE*, **1** (2019), 13–31.

[23] Grosser T., J. Bloemen, M. Mack, J. Vitsenko. Hadoop and Data Lakes. BARC Research Study. https://bi-survey.com/data-lakes-usage, 7 December 2019.

Snezhana Sulova
Department of Informatics
University of Economics—Varna
77, Kniaz Boris I Blvd.
9002 Varna, Bulgaria
e-mail: ssulova@ue-varna.bg

Received October 22, 2019
Final Accepted January 28, 2020