Cooperation of Business Intelligence and Big Data in one Ecosystem

Matej Černý[1,*]

[1]University of Economics in Bratislava, 852 35 Dolnozemská cesta 1, Bratislava, Slovakia

Abstract. This paper is focused on the issue, how the business can analyze all data types (structured and unstructured) in one cooperative environment. With structured data handle Business Intelligence and with unstructured data on the other side Big Data. As a solution to this issue, we have suggested our Business Intelligence and Big Data ecosystem. This model - the ecosystem is based on already proven data processing processes running in Business Intelligence and in Big Data areas. Both processes are integrated into one unit. We have also described their common functioning.

1 Introduction

Business Intelligence (BI) is a modern term used to refer to a set of technologies and processes that use data to understand and analyze business results [1, 2] designates it as a process that involves gathering sufficient information at the right time and in a usable form and analyzing it so that it can be used for a positive impact on business strategy and tactics.

According to Rainer et al. [2], BI supports comprehensive and non-recurring decision-making with a focus on tactical management and knowledge workers, which as is defined above, can be extended to include strategic and operational management levels. Consulting company Gartner [3] and Mallach [4] also agree with the above definitions. Most authors associate this term mainly with the creation of a unified enterprise data warehouse, respectively data marketplace and also advises it in the area of knowledge management, which also concerns.

According to the authors Laudon & Laudon [5], is the term Business Intelligence used to describe: „the infrastructure for warehousing, integrating, reporting, and analyzing data that come from the business environment, including big data. The foundation infrastructure collects, stores, cleans, and makes relevant information available to managers."

Most of the data collected by businesses today are in the form of transactional data that can be easily written in the form of two-dimensional tables, in rows (records) and columns, and are stored in a relational database.

Such databases containing structured data can be used in the enterprise as a data source for further analysis. However, in order to use them, it is necessary to extract, cleanse and store this data first in a central data warehouse, where it will be possible to perform further complex analyses to meet the needs of the decision-making process. A data warehouse can be defined as a database or a set of databases that contain business data from various sources in the enterprise as well as from external systems that cover all aspects of business processes in all functional areas, all aspects of products and services and customers and partners [2]. The data warehouse contains both historical and current data [5]. This integrated data is used

* Corresponding author: matej.cerny@euba.sk

to support data analytics and decision-making [6]. The Creation and operation of a data warehouse is costly [7].

A data warehouse that contains only a subset of all data stored in a data warehouse is called a data mart [6]. Data mart is a cheaper and smaller version of the data warehouse and is intended to support only a specific group of users in the enterprise, e.g. sales department, or strategic business unit [7].

To load data from different sources and insert them in a data warehouse is used data extraction, transformation, and load (ETL) process. This process extracts information from internal and external databases, transforms it using a set of business definitions, and uploads it to the data warehouse.

Another term is Data Analytics (DA). This concept is more focused on tools and techniques for analyzing and understanding data. It includes techniques like OLAP, statistics, models and data mining (Data mining is a process for extracting hidden, unknown, but potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, and random data [14]).

BI and DA are essentially about integrating all the information flows produced by a company into one coherent corporate data set (data warehouse) and by using modeling, statistical analysis tools (such as normal distribution, correlation, and regression analysis, Chi-square analysis, forecasts, and cluster analysis) and data mining tools give a sense to all this data so that managers can make better decisions and plans.

Summarizing the above knowledge, we will get a comprehensive overview of the functions and capabilities of BI. The entire BI data processing process could be described in these steps. Raw transaction data are obtained from internal systems such as CRM, ERP or other OLTP systems. This data can be enriched with data from external sources. Using the ELT process, data are transferred to the data warehouse, where they are stored. From data warehouse are distributed to the data mart according to the needs of functional areas of the company. In the data marts are performed the required analyzes with the help of decision-support systems and then are provided to the end-user in a visual form through the presentation layer.

Recent trends in organizations' data retention need point to the need to retain unstructured or semi-structured data, such as images, videos, sound recordings, or text files from various sources such as visitor web records, email messages, social media content, as well as machine-generated data from sensors or automated systems (e.g. stock systems, etc.) [13], and such data is not suitable for processing and storage using traditional relational databases. This data typically consumes a huge capacity of data storage at the level of petabytes or exabytes, so we are talking about billions or even trillions of records from various sources.

Data in such volumes and forms are called Big Data. Authors such as Laudon & Laudon [5], Stair et al. [2] and Baltzan [6] agree on the Big Data definitions. According to them, Big Data can be defined as the collection of large data sets, including both structured and non-structured data, which cannot be analyzed using traditional methods and database tools. In terms of processing such diverse and mainly unstructured data, it could not be longer use standard database systems and relational databases, therefore the use of NoSQL (and Not Only SQL) respectively non-relational database management systems [6]. These use a more flexible data model and are designed to manage large volumes of data on many distributed devices and allow easy scaling. They are useful for accelerating simple queries over large volumes of structured and unstructured data, including the Web, social media, graphics, and other forms of data that are difficult to analyze with traditional SQL-based tools. NoSQL also supports SQL queries.

A comprehensive software ecosystem for data processing and storage was brought by Hadoop. It is an open-source software platform that enables distributed parallel processing and storage of huge amounts of data on many commercial computers [5]. It also allows the

interconnection of these unstructured data sets with relational databases [2]. It divides the big data problem into smaller sub-problems, distributes them among thousands of cheap computer processing nodes, and combines the result into a smaller data set that is easier to analyze. Hadoop can process large amounts of data, including structured transactional data, semi-structured data such as Facebook and Twitter feeds or complex data such as web server log files, or unstructured audio and video data.

For these types of data had to be developed new types of analyzes - Advanced Analytics. Advanced analytics can be characterized as an autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, usually beyond traditional business data analytics (BI) procedures to discover more detailed, deeper information, make predictions or generate recommendations. The main goal of advanced analytics is to extract useful information from large data sets and transform them into an understandable form. The major processes of Big Data include capture, curation, storage, search, sharing, transfer, analysis, and visualization [11, 12].

2 Methods

This article focuses on identifying data collection processes, which consist of storing, processing, analyzing, and compiling output for business decision making in the BI and Big Data areas. Based on the analysis of the identified processes, we try to find the intersection of these two processes. The goal is to create one common ecosystem of processing both structured and unstructured data in the enterprise. Easy to say provide the right information in the right way at the right time and for the right people at acceptable costs.

By achieving this goal, we focused on finding available scientific and professional publications that include the data processing processes in the areas of BI and Big Data and which defining these and other terms. Using the synthesis of acquired knowledge, we try to find their commonly used definitions. This is the content of the first part of this paper.

In the area of research results and discussion, we publish our proposed BI and Big Data ecosystem. This model – the ecosystem is based on already proven processes that are integrated into one unit and describe their common functioning. It requires validation in terms of its content and applicability in practice, which will require further work.

3 Results and Discussions

BI and Big Data can and should be used as one integral unit that will help an enterprise to economically accept, collect, and store any type of data in relatively any amount. Using this data, the organization can then support the decision-making process using various types of analyzes depending on the goal that it wants to achieve. It can then process, distribute and present the data for specific groups of managers, specialized employees or use it for automatized decision-making. The BI and Big Data ecosystem is shown in Fig. 1.

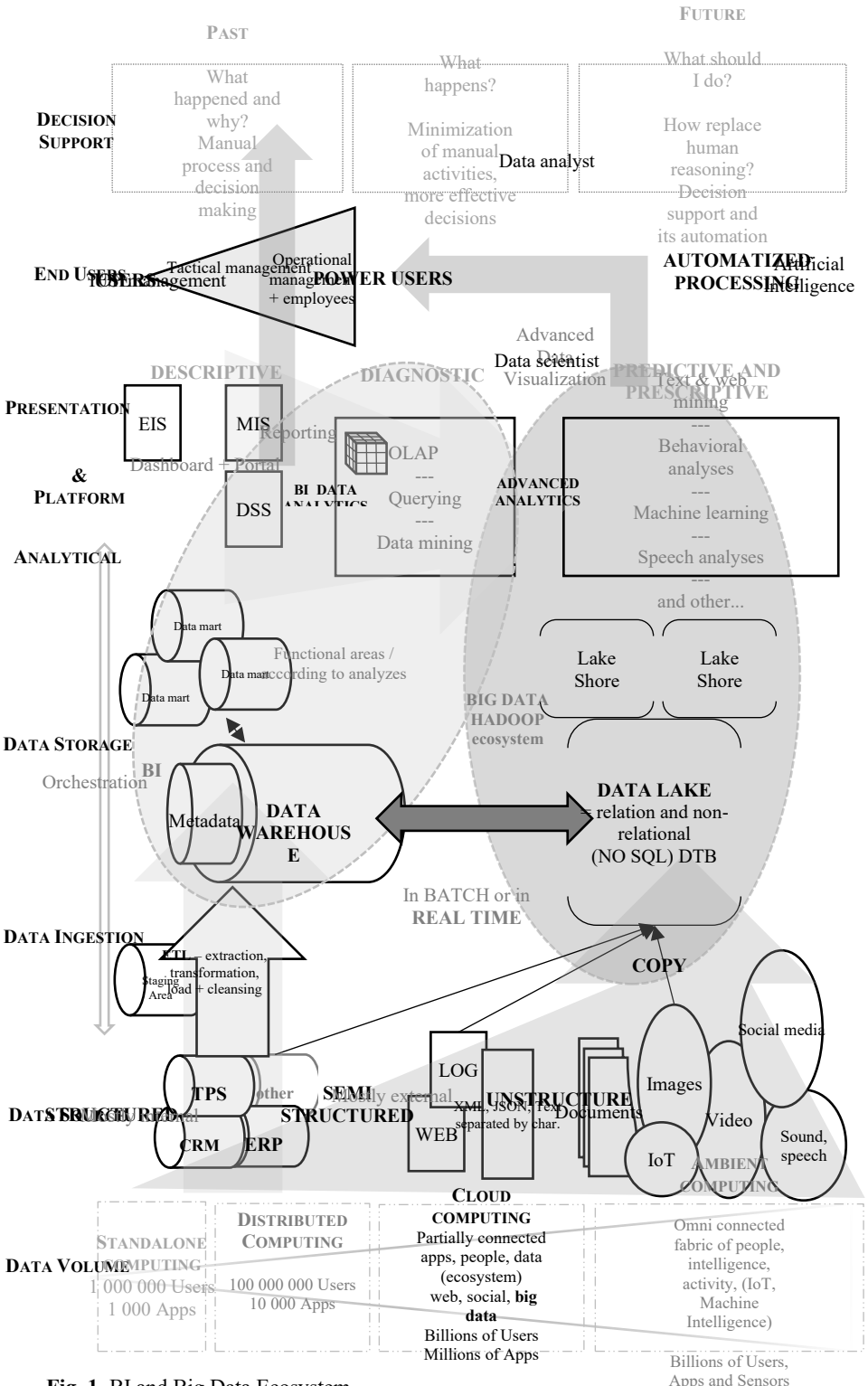

Fig. 1 BI and Big Data Ecosystem

The BI and Big Data ecosystem inherently provides all information system functions such as data collection, transmission, storage, processing, distribution and presentation. The ecosystem itself is based primarily on the BI ecosystem, which is enhanced by the specifics of Big Data.

The first column in Fig. 1 characterizes the evolution of data creation, from the beginning of computer use in enterprises to the latest ubiquitous technologies producing large volumes of diverse unstructured data.

In the second column, we are more specific, we define areas of origin of various types of data in organizations, where we go from structured more or less internal data to unstructured external data sources.

The next section deals with how the data is transferred to the central database of the company. Structured data uses a process called extraction, transformation, and load (ETL). In addition to these activities, the data in this process is also cleaned (schema-on-write). For unstructured data in large volumes, this process cannot be used and therefore the data is simply copied to the repository and before processing, they shall be edited only to the extent required for their analysis (scheme-on-read).

Structured data is stored within a central enterprise database called a data warehouse. From data warehouse it can be distributed to data marts, where it is analyzed for a specific purpose of a functional area of the enterprise. For unstructured data and not just for it, because it is also possible to store structured data at a low-cost, are used distributed storages based on relational and non-relational NO SQL (not only SQL) databases and so-called data lakes. The data lake is: „a methodology enabled by a massive data repository based on low-cost technologies that improve the capture, refinement, archival, and exploration of raw data within an enterprise“ [15] Data lake from the viewpoint of business domain load all data from source systems, no data is turned away and data are stored at the leaf level in an untransformed or nearly untransformed state. Both data warehouse and data lake are data repositories. However, they are different in many aspects from concepts, structures, and implementation [16]. For example, a traditional data warehouse stores data in files or folders, the data lake uses a flat architecture to store data. Each data element in the data lake is assigned a unique identifier and is marked with a set of extended metadata tags. When a business question arises, the data lake will be queried for all relevant data, thus creating and providing a smaller dataset that can then be analyzed to help answer the question. A data lake is something like a data warehouse for any type of data, while lakeshore marts are an image of data marts. They can be characterized, according to El Kaim [8], as data marts, each has a specific model for one limited context and processes and organizes data for analytical use.

Within the BI and Big Data ecosystem, we have identified a part of the process as an orchestration. This term could be defined according to Watts [8] as the automation (understood as running a single task on its own) of multiple activities at once. Orchestration includes automated organization, coordination, and management of computer systems, middleware and services. Orchestration uses more automated tasks to automatically perform a larger workflow or process. It could consist of multiple automated tasks and include multiple systems.

The analytical and presentation platform provides an overview of analytical methods by specific data volume levels. Here are mentioned IS-s, which use data from the data warehouse, respectively data lake, process them and shift to support the decision-making process at individual management levels. We also present analytics used in BI and Big Data.

Individual outputs from data analytics and individual types of analyses are designed for different end-users. The BI is used primarily by managers at all management levels and from power users they are data specialists. The Big Data is characterized by a high degree of sophistication in the analysis, and therefore their main consumers are data scientists who extract useful knowledge from the data for the enterprise, which can be then applied in different areas of business. The last area is automated systems that are used to automate

decisions and thus relieve managers of certain types of decisions. The last column shows the time orientation of the analyses and also the way how these analyses help the company in the decision-making process. Big Data are at the last stage of a long-term transition from systematic reporting (IT approach) to the means of automated analysis (business approach) [17].

In the literature as well as in the professional community it is not always clear what the relationship between BI, Big Data and data analytics is. The relationships between these terms were well summarized by Dedic and Stanier [9] with the visualization shown in fig. 2. Based on an analysis of available literature and validation in practice, they consider knowledge discovery (KD) to be a top-level concept that uses data analysis techniques to discover and create new knowledge in organizations. Within KD, data analytics is its object and contains various disciplines and approaches, including BI and Big Data (advanced analytics). BI and Big Data are considered by the authors as disciplines at the same. We would like to add that Big Data is a logical extension of BI, especially in the field of unstructured data processing and the necessary data analysis, and within the ecosystem, these two approaches to data work complementary.

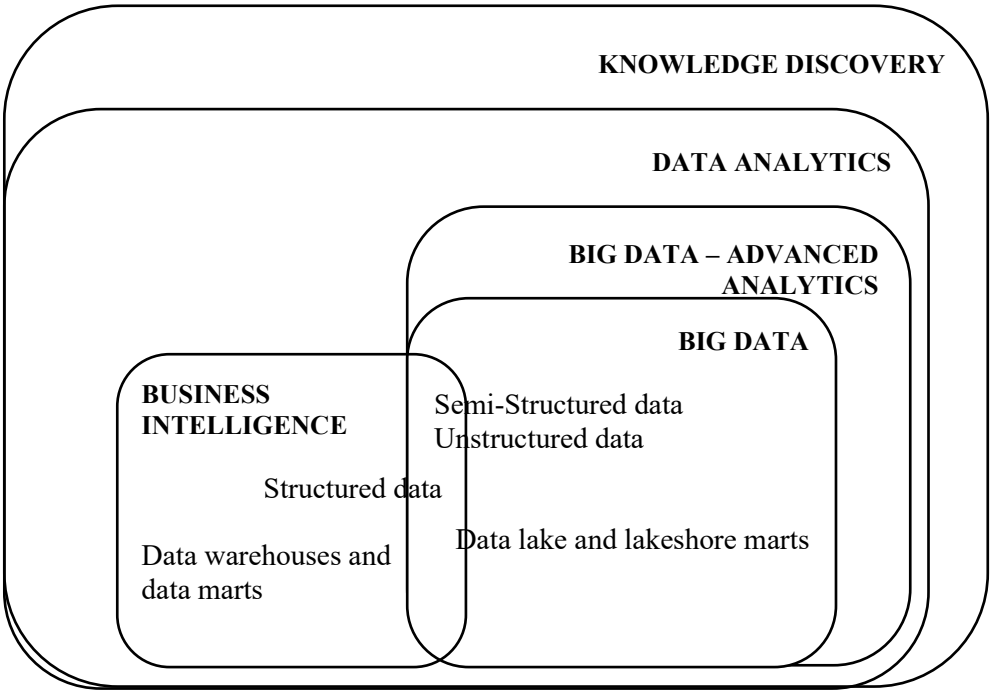

Fig. 2 Visual representation of the relationship between BI, Big Data, data analytics and knowledge discovery

4 Conclusion

Our proposed ecosystem (fig. 1) requires validation in terms of its content and applicability in practice, which will require further work.

Its compilation is based primarily on the conclusions of the authors Dedic and Stanier [9], as well as others. Our ecosystem is based on two independent frameworks, BI and Big Data, and their data processing processes, which we have integrated into a single unit. Between these two frameworks, we defined the possibilities for cooperation based on existing processes, technologies, solutions, research and practice.

Acknowledgement

The paper was elaborated within VEGA No. 1/0388/20 IT Management in Enterprises in Slovakia: International Standards and Norms Versus Individual Business Processes – proportion 100.

References

1. T. H. Davenport, J. G. Harris, *Competing on Analytics* (Harvard Business Review Press, 2007)
2. R. Stair, G. Reynolds, T. Chesney, *Principles of business information systems* (Cengage, United Kingdom, 2018)
3. Gartner. IT Glossary, Analytics and Business Intelligence (ABI). https://www.gartner.com/it-glossary/business-intelligence-bi/ [accessed 12.09.2019]
4. E. G. Mallach, *Information Systems*. (CRC Press, USA, 2016)
5. K. C. Laudon, J. P. Laudon, *Management Information Systems*. (Prentice Hall, New Jersey 2016)
6. P. Baltzan, *Information Systems*. (McGraw-Hill Education, New York, 2018)
7. R. K. Rainer, B. Prince, C. G. Cegielski, *Introduction to Information System*. (Wiley, USA, 2014)
8. W. El Kaim, Big data architecture: Hadoop and Data Lake (Part 1). https://www.slideshare.net/welkaim/big-data-architecture-hadoop-and-data-lake-part-1 [accessed 27.09.2019]
9. S. Watts, Difference between IT Orchestration and IT Automation Explained. https://www.bmc.com/blogs/it-orchestration-vs-automation-whats-the-difference/ [accessed 27.09.2019]
10. N. Dedić, C. Stanier, Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. *ERP Future 2016: Innovations in Enterprise Information Systems Management and Engineering*. **285**, 114 (2016)
11. Neaga, Y. Hao, A Holistic Analysis of Cloud Based Big Data Mining. *International Journal of Knowledge, Innovation and Entrepreneurship*. **2, 56 (**2014)
12. D.P. Shukla, S.B, Patel, A.K. Sen, A Literature Review in Health Informatics Using Data Mining Techniques, *International Journal of Software and Hardware Research in Engineering*, **2**, 123 (2014)
13. B. M. Balachandran, S. Prasad, Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. *Knowledge-based and Intelligent Information & Engineering Systems*. **112**, 1112 (2017)
14. X. Wu, V. Kumar, JR. Quinlan, J. Ghosh, Q. Yang, H. Motoda, GJ. McLachlan, A. Ng, B. Liu, SY. Philip *Top 10 algorithms in data mining. Knowledge Information Systems*. **14**, (1):1–37 (2014)
15. M. Chen, SW. Mao, YH Liu, Big Data: A Survey. *Mobile networks & Applications*. **19**, 171 (2014)
16. F. Huang, 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), 820 (2015)

17. PP. Khine, Wang ZS. Data lake: a new ideology in big data era. *4.TH Annual International Conference on Wireles Communication and Sensor Network*, **17** (2017)
18. S. Mitrovic. Specifics of the integration of Business Intelligence and Big Data technologies in the processes of economic analysis. *Biznes Informatika*. **42**, 40 (2017)