sciendo CzOTO 2020, volume 2, issue 1, pp.166-173

THE ROLE OF BIG DATA IN INDUSTRY 4.0 IN MINING INDUSTRY IN SERBIA

doi:10.2478/czoto-2020-0020
Date of submission of the article to the Editor: 28/11/2019
Date of acceptance of the article by the Editor: 18/02/2020

Ing. Eva Tylečková[1]–*orcid id: 0000-0003-3583-2678*
Prof. Ing. Darja Noskievičová, CSc.[1]– *orcid id: 0000-0002-1154-712X*
[1]VŠB – Technical University of Ostrava–**Czech Republic**

Abstract: The current age characterized by unstoppable progress and rapid development of new technologies and methods such as the Internet of Things, machine learning and artificial intelligence, brings new requirements for enterprise information systems. Information systems ought to be a consistent set of elements that provide a basis for information that could be used in context to obtain knowledge. To generate valid knowledge, information must be based on objective and actual data. Furthermore, due to Industry 4.0 trends such as digitalization and online process monitoring, the amount of data produced is constantly increasing – in this context the term Big Data is used. The aim of this article is to point out the role of Big Data within Industry 4.0. Nevertheless, Big Data could be used in a much wider range of business areas, not just in industry. The term Big Data encompasses issues related to the exponentially growing volume of produced data, their variety and velocity of their origin. These characteristics of Big Data are also associated with possible processing problems. The article also focuses on the issue of ensuring and monitoring the quality of data. Reliable information cannot be inferred from poor quality data and the knowledge gained from such information is inaccurate. The expected results do not appear in such a case and the ultimate consequence may be a loss of confidence in the information system used. On the contrary, it could be assumed that the acquisition, storage and use of Big Data in the future will become a key factor to maintaining competitiveness, business growth and further innovations. Thus, the organizations that will systematically use Big Data in their decision-making process and planning strategies will have a competitive advantage.
Keywords: big data, industry 4.0, data quality, data security

1. INTRODUCTION

Industry 4.0, called also as 4thindustrial revolution, brings new challenges such as digitalization of production processes, interconnecting devices and machines, and integration of production and business models. Along with it the effort to connect all devices (not only machines and devices, but also cars, phones, household applianc-es, security systems, and others)into the internet of things comes. In industry, new

technologies and the possibility of incorporating sensors, cameras or other devices into any component or machine, bring the opportunity to collect data in very short intervals and very large volumes. As a result of this fact, together with digitalization and possibility of online process monitoring, the amount of data produced is constantly increasing – in this context the term big data is used. The term big data encompasses issues related to the exponentially growing volume of produced data, their variety and velocity of their origin. These characteristics of big data are also associated with possible processing problems. However, it can be assumed that the acquisition, storage and use of big data in the future will become a key factor to maintaining competitiveness, business growth and further innovations. Thus, the organizations that will systematically use big data in their decision-making process and planning strategies will have a competitive advantage. A key prerequisite of profiting from analysis of big data is data quality. On the contrary, if data quality is poor, it becomes a serious risk to the organization. From this reason it is necessary to regularly assess data quality to ensure that all the relevant data quality dimensions are met.
Big data acquisition and analysis provide essential information for objective decision-making processes. The ability to get big data, analysis and immediate responses to process changes brings the opportunity to transform any industry. With efficient information exchange between system units, real-time information transfer and an integrated management system, industry 4.0 production can be expected to be more efficient and cost-effective.

2. BIG DATA

Big data refer to large amounts of multi-source, heterogeneous data generated throughout the product lifecycle (Tao et al., 2018). Usually big data are characterized by “5 V’s” that are shown in Figure 1.

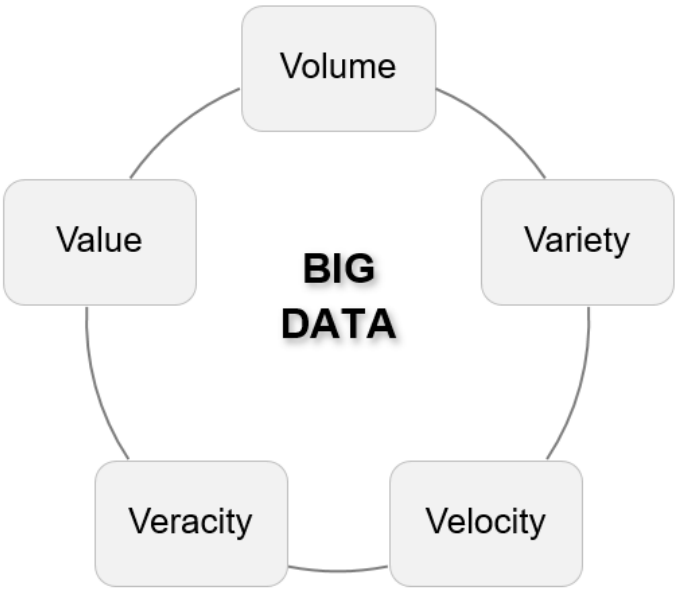

Fig. 1. The 5 V’s of big data

Source: (Own source)

Among 5 V’s are volume, that refers to size and scale of data, variety, that refers to number of different types and format of data, velocity, that refers to the high speed

with which data are generated and need to be processed (Hwang and Chen, 2017). These three V's were introduced as early as 2001 (Laney, 2001). However these main characteristics reflect big date until today – big data are data produced in large volume and many types and formats, since big data come from many different sources and at the same time require high velocity to be processed. Lately, two more V's were added to the characteristics – veracity, that refers to the quality, reliability or uncertainty of the data, and value, that points out the data are valuable only with the proper context.High veracity data have many contributions to the overall results. On the contrary, low veracity data contain a lot of meaningless records. Value refers to a problem of identifying useful and valuable data among all the big data.

The amount of data produced is constantly increasing. Currently, the data of size 1 terabyte (TB) or greater is considered to be big data (Hwang and Chen, 2017). For example, an autonomous car gathers nearly 1 gigabyte (GB) per second, that means 1 terabyte is gathered less than every 20 minutes (Amara, 2013). Furthermore, according to International Data Corporation's (IDC) forecast there will be 41.6 billion devices connected in Internet of Things in 2025 and they will be generating 79.4 zettabytes (ZB, equals 1 000 000 000 TB) of data (International Data Corporation, 2019).

Although the amount of data will grow in the future, the data themselves are only bare facts, such as measured values of quality features or records of transactions made. Without putting the data in the proper context, the data are practically meaningless (Lee et al., 2013).The value of data lies in their possible use for support of business processes and decision-making processes. From this reason, it is necessary to transform data into valuable information. From data, the information might be inferred or transformed. Information can be understood as relevant, usable, meaningful or processed data (Lake and Drake, 2014).The information in the proper context can be used to acquire knowledge. Knowledge can be understood as cognition or recognition, capability to act and understanding the reasons. To put it simply, knowledge includes know-what, know-how and know-why (Liew, 2007).

The obtaining and using information hidden in big data and getting an appropriate knowledge have become a key basis to drive competitiveness, business growth and innovations, as also emerged from studies (Manyika et al., 2011).

In addition to the positive consequences of big data, there are also some serious threats associated with big data processing. The main threats are possible cyber security risks as big data are stored, analysed and used online, and insufficient data quality, that directly and negatively affects the quality of information obtained.

2.1. Big data quality

For efficient analysis and managing the big data, a key prerequisite is data quality. Data quality can be understood as the ability of data to describe real situations objectively and correctly. Data quality needs to be monitored and managed, especially due to potential high costs of consequences of data of poor quality or errors. The business costs of data of poor quality, including irrecoverable costs, rework of products and services, workarounds, and lost and missed revenue may be as high as 10 to 25 percent of revenue or total budget of an organization (English, 1999). Data quality needs to be assessed according to some of the dimensions. Each of the data quality dimensions expresses a specific characteristic of the dataas it can be seen in Table 1.

Table 1
Dimensions of data quality

Dimension	The purpose of dimension
Accuracy	To assess whether are data in the right context and reflect reality accurately.
Availability	To evaluate the ways the user can access information if he needs it. Availability can be further distinguished according to time, place, structure and required format.
Completeness	To assess whether all data for a required context are available.
Consistency	To ensure the possibility to use data across the whole company.
Integrity	To verify whether there is a structure and relationship between data.
Timeliness	To assess if data represent reality from required point of time.
Uniqueness	To ensure a single view of the data and to assess whether all data are recorded just once.
Validity	To verify if data conform to the syntax of their definition. Data should be valid for the required context.

Source: (Novotný et al., 2005; DAMA UK Working Group, 2013).

Except for dimensions mentioned above, it ought to be also considered: usability, understand ability, relevance, simplicity and availability of data. The most frequent problems are missing data and problems with integrity. These problems may impair the predictive ability and quality of the resulting analyses.
To avoid possible losses caused by data of poor quality, the production data should be continuously analysed. The results of analysis are usually in the form of structured document and supplementary tables (Novotný et al., 2005). Based on the results of the analyses, detailed requirements for solving specific problems and specification of necessary repairs or update operations are formulated.
The analysis of data needs to be performed at regular basis. The tools for assurance the data quality are based on four interconnected processes, which are summarized in Table 2.
The main goal of data quality assessment is to ensure data quality with respect to all the characteristics mentioned above. The main activities within the data quality assessment are to identify and treat (Novotný et al., 2005):

- data, that are missing or are unusable, so data completeness is ensured,
- data, that are not stored in standard or required format, so data availability is ensured,
- data whose values represent conflicting information, so data consistency is ensured,
- data that are inaccurate or outdated, so data accuracy is ensured,
- records that are duplicates, so data uniqueness is ensured,
- data, that lack important relationship with other data, so data integrity is ensured.

Table 2
Processes of data quality assurance

Process	The description of process
Investigation	The evaluation of the characteristics of the input data, their type and format, while identifying other information previously hidden due to data inconsistency errors or non-standard formats. Investigation process may also include identifying the type of data due to predefined rules, identifying general items (such as date), and reposting the results of the analysis.
Standardization	The ensuring a uniform representation of data for the further processing, where input data come from different information systems and usually indifferent formats. The process of standardization should include transformation of data according to predefined rules, formatting data items to a consistent state, normalization of data elements with respect to language and national specifications.
Enrichment of information	The completion of the data, correcting and extending information from other sources. The process of searching links between individual records creates new information that increases the efficiency and improves further use.
Integration	The implementation of a unified control process and improving data quality within the whole information system of the company. Integration represents standard of data quality and guarantees all data conform to that standard.

Source: (Novotný et al., 2005).

If the data have no sufficient quality, the information is no longer reliable, the knowledge gained is distorted, and business decisions based on such knowledge, do not have expected impacts, so companies cease to trust their data or their information systems.

2.2. Big data security

The fact that the amount of data collected, processed and analysed is still increasing, means a challenge not only in data quality assurance, but also for security and privacy of big data. Data must be protected from external threats. Unfortunately some traditional security methods are no longer suitable for use in context of big data. For instance, methods such as firewalls and demilitarized zones have been developed to secure private computing infrastructure, however today many organizations use public clouds for storing huge amount of data collected. One possibility to ensure the security of data stored in the public cloud is encryption. If the data in database are encrypted, the extraction of information from such encrypted database will be associated with a number of questions needed to be answered – like if it is necessary to encrypted also the queries to extraction data, or who should have an access to decrypted databases (Moura and Serrao, 2015). The changing conditions also requiredata security controls ideally at the location of origin of the data, or at least as close as possible to this place to secure prompt reaction in a case of attack.

The ability to connect devices in internet of things contributes to a decentralized system architecture in which an attack can come from any connected device and negatively affect the whole information system. The use of modern technology allows shar-

ing information or data with business partners via internet – however this kind of transmission must also be sufficiently secured. A potential source of problems can be also an insufficiently secure mobile device, from which authentication processes are perform, or through which it is possible to access big data analysis and information.Similar threat presents a policy BYOD (bring your own devices). If employees use their private devices (laptops, tablets, phones) for work tasks and deal with corporate data, it is necessary to protect these devices from potential threats and information leaks.

To ensure big data security and privacy, four main areas should be secured as shown in Table 3.

Table 3
Big data security areas

Big data security area	The key activities to securing the area
Data privacy	Securing personal and sensitive data by cryptography and ensuring access to data only by authorized users by granular access control.
Infrastructure security	Securing distributed computations, for example by using Map Reduce.
Data management	Securing also data storage and transaction logs and verifying the level of data protection against unauthorized use, for example by granular audits.
Integrity and reactive security	Checking data gained from different sources to ensure their integrity, establishing the end-point validation and filtering for effective system control, using the real time analytics for detection of security incidents andto look for abnormalities or attacks in the system.

Source: (Cloud Security Alliance, 2013).

These areas ought to be secured in whole big data lifecycle, from getting data from various sources, data processing, transforming, storing, transportation and data usage.

The goal of big data security is to predict cyber-attacks, prevent or stop them and avoid unauthorized access to data in order to modify, delete or steal sensitive data. If organizations are able to identifying the patterns in which a threat or possible attack is detected, they might react immediately, improve their approach to security and prevent data misuse. A possible approach to identifying suspicious patterns is to identify statistical baseline by using historical data that present a common behaviour of system or network and after that discover unusual behaviour by comparing determined baseline against current real data from the system or network (Gulf Business Machine, 2019).

3. CONCLUSION

Industry 4.0 brings huge amount of new opportunities for improving processes, increasing the performance and competitiveness of organizations. The methods and tools used in era of industry 4.0, such as digitalization and online process monitoring, deal with large amount of data, so called big data. Big data offer the opportunity to analyse any part of a process and find the root causes of nonconformities. Hereby, big data empower companies to adopt more suitable strategies to achieve more com-

petitiveness. The information gained from big data can be used to optimize and improve performance of processes, identify deviations and abnormalities in processes, eliminate sources of variability and other problems.

Along with the advancement of technologies that can lead to online process monitoring, there is also an increasing amount of unstructured big data produced in high speed from various resources need to be stored and analysed. The way big data are collected and processed has a direct impact on the overall quality of information gained from data. Establishing a robust system for collecting, processing and analysing big data will make organizations more competitive as big data can be transform into accurate information about processes and products. Accurate information about processes and products leads to improving services, quality of products, decreasing failure rate, reducing costs, better guarantees for customers. There is no doubt that effective data management has become a key factor for future success.

It is necessary to point out that a key prerequisite for the effective use and evaluation of big data is the data quality. Only data that reflect a real state of process might be use for effective decision-making. An analysing and maintaining big data quality should be one of the key tasks in the future.

The increasing amount of data collected and stored on clouds provides more opportunities for attacks and other security risks. Although big data represent interesting and valuable opportunity to develop businesses, they are also facing vast challenges in terms of their security and privacy. Some existing data protection tools and methods cannot be effectively applied in today's conditions, so it is necessary to look for new methods for ensuring big data security and privacy.

The future challenges are not only in extracting, transforming, loading and analysing big data, but also in keeping big data secured. The security management at the age of big data ought to be more flexible and being prepared for online attacks like the data will be collecting online and in the huge amounts. At the era of big data, security and privacy will become a serious issue and important challenge for future development.

ACKNOWLEDGEMENTS

The article was supported by the specific university research of the Ministry of Education, Youth and Sports of the Czech Republic No. SP2019/62 and by the European Regional Development Fund in project "Research Platform focused on Industry 4.0 and Robotics in Ostrava" No. CZ.02.1.01/0.0/0.0/17_049/0008425 within the Operational Programme Research, Development and Education.

REFERENCES

Amara, A. D., 2013. *Google's self-driving car gathers nearly 1 GB/sec*. Kurzweil. Available from: https://www.kurzweilai.net/googles-self-driving-car-gathers-nearly-1-gbsec

Cloud Security Alliance, 2013. *Expanded Top Ten Big Data Security and Privacy Challenges*. DOI: 10.13140/RG.2.1.1744.1127

DAMA UK Working Group, 2013. *The six primary dimensions for data quality assessment*. Available from: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

English, L., 1999. *Improving Data Warehouse and Business Information Quality*, Wiley, New York.

Gulf Business Machine, 2019. *The Unspoken truth: The role of cybersecurity in breaking the digital transformation deadlock*. GBM 8th Annual Security Survey 2019.Available from: https://gbmme.com/wp-content/uploads/2019/10/GBM-Security-Whitepaper-2019.pdf

Hwang, K., Chen, M., 2017. *Big-data analytics for cloud, IoT and cognitive computing*, Wiley, Hoboken.

International Data Corporation, 2019. *The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast*. Available from: https://www.idc.com/getdoc.jsp?containerId=prUS45213219

Lake, P., Drake, R., 2014. *Information Systems Management in the Big Data Era,* Springer, Cham.

Laney, D., 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety*. META Group. Available from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lee, J., Lapira E., Bagheri, B., Kao, H., 2013. *Recent advances and trends in predictive manufacturing systems in big data environment*. Manufacturing Letters 1(1), 38-41, DOI: 10.1016/j.mfglet.2013.09.005

Liew, A., 2007. *Understandingdata, information, knowledge, and their interrelationships*. Journal of Knowledge Management Practice, 7(2), 1-10.

Manyika, J. et al., 2011. *Big data: The next frontier for innovation, competition. and productivity.* McKinsey Global Institute. Available from:https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Moura, J. A., Serrao, C., 2015. *Security and Privacy Issues of Big Data.*Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence, 20-52, DOI: 10.4018/978-1-4666-8505-5.ch002

Novotný, O., Pour, J., Slánský D., 2005.*Business intelligence: jak využít bohatství ve vašich datech*. Grada Publishing, Praha.

Tao, F., Qi, Q., Liu, A., Kusiak, A., 2018. *Data-driven smart manufacturing: An Overview and Perspective*.Journal of Manufacturing Systems,48 (1), 157-169, DOI: 10.1016/j.jmsy.2018.01.006