Sylwia Gierej

The main barriers to effective use of Big Data

Zarządzanie. Teoria i Praktyka nr 2 (20), 39-45

2017

Sylwia Gierej
Międzynarodowy Chiński i Środkowo-Wschodnioeuropejski Instytut Logistyki i Nauki o Usługach, Wydział Inżynierii Zarządzania, Politechnika Białostocka

Zarządzanie. Teoria i Praktyka 20 (2) 2017
ISSN 2081-1586
e-ISSN 2449-9730
str. 39-45

THE MAIN BARRIERS TO EFFECTIVE USE OF BIG DATA / Główne bariery w efektywnym wykorzystaniu Big Data

Adres do korespondencji:
e-mail: s.gierej@pb.edu.pl

Streszczenie
Na obecnym etapie rozwoju gospodarki informacja ma szczególne znaczenie. Dzięki postępowi technologicznemu bardzo szybko generowane są duże ilości danych. Zjawisko Big Data skłania coraz więcej przedsiębiorstw do zwrócenia szczególnej uwagi na posiadane zasoby informacji, możliwość zakupu zewnętrznych danych i ich analizy w celu osiągnięcia wartości biznesowej. W artykule przedstawiono główne problemy i bariery w prawidłowym wykorzystaniu potencjału dużych zbiorów danych. Za owe bariery uważa się ograniczenia technologiczne, kwestie prawne dotyczące bezpieczeństwa danych, problemy z właściwym zarządzaniem i jakością danych oraz przeszkody w monetyzacji danych. Zwrócono uwagę na kwestie, które powinny zostać uwzględnione przez organizacje decydujące się na wdrożenie rozwiązań Big Data.

Słowa kluczowe: Big Data, Bug Data, jakość danych, monetyzacja.

Abstract
At the present stage of the development of the economy information is of particular importance. Thanks to technological progress, large amounts of data are generated very quickly. The Big Data phenomenon is driving more and more businesses to pay special attention to their data, the possible purchase of external data, and their analysis to achieve business value. The article presents the main problems and barriers to the proper utilization of the potential of large data sets. These barriers are technological constraints, legal issues related to data security, problems with the proper management and data quality, and obstacles to the data monetization. Attention was paid to the issues that should be included by organizations deciding to implement Big Data solutions.

Key words: Big Data, Bug Data, data quality, monetization.

JEL Classification: L15

1. INTRODUCTION

The current level of economic development is largely dependent on advanced information technology. Due to the development of IT systems and mobile platforms, the conditions under which organizations operate are evidently changing. The progress in the Internet area. Things has made it possible to collect data that was difficult to obtain a few years ago or even impossible. Collections of currently collected data are huge and are constantly growing at a very fast pace. This contributes to the growing interest in the Big Data concept, which is to briefly manage and analyze mass data volumes with high flow rates and diversity. It has been recognized as one of the key technologies in smart management strategy and future global development. In dealing with such large data sets, it is essential to focus on the appropriate analysis and assessment of which resources are actually useful. Maintaining data on your own or external servers requires a lot of financial investment. For this reason, it is important for companies to focus their attention not only on data collection but primarily on their effective analysis and elimination of overdue and worthless data. Due to the flow rate and the variety of data produced, distributed and used over

a few seconds to a few hours, as well as the structural and unstructured form of data, it appears that existing data analysis methods do not cope with proper management of data. For this reason, the term Big Data should be understood not only as a mass data collection, but also as human resources, organizations, and the right technology to provide valuable information. In Big Data analysis it is crucial to extend the perspectives and thought horizons, allowing you to change your field of vision, take a closer look at your information resources, and then skilfully select those that are worth attention and continue your analysis. (Lee, Sohn, 2016, p.25-38).

The statistics show that the global demand for enterprise experts specializing in the analysis of large data sets is growing rapidly. According to the information presented by McKinsey & Company in 2018 in the United States alone will be 14-19 thousand Big Data research analysts and 1.5 million managers who can practice this knowledge effectively in order to make effective decisions. This demonstrates that companies recognize the potential of acquiring, processing and analyzing mass data collections, and utilizing information obtained to optimize processes and strategy planning. (McKinsey, 2011). At the same time, organizations are catching on on the trend of large-scale use of data, focusing on their storage without analyzing whether they actually carry a particular value for the enterprise. This results in a situation where the costs of maintaining the data that does not bring any measurable benefits are incurred. As a result, instead of gaining and benefiting from the Big Data potential, the company loses its inexpensive management of its data volumes. The article further discusses the Big Data concept and highlights the major barriers to effective use of large data sets.

2. BIG DATA CHARACTERISTICS

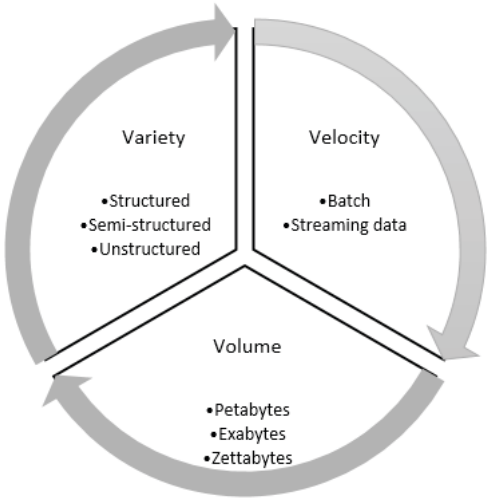
By analyzing the available publications, you may encounter several definitions that discuss the concept of Big Data. So far, none of them has been officially accepted and fully accepted. The first publications on the subject of large volumes of data come from the late 1990s. The authors of one of the first definitions are M. Cox and D. Ellsworth, according to which Big Data is a large dataset that needs to be expanded to extract information values. (Cox, Ellsworth, 1997).

With the passage of time and technological advances, the definitions associated with Big Data evolved, gaining more attributes. Technologies that allow the processing of large volumes of data of a completely different nature (quantitatively and qualitatively), such as Hadoop and HBase, make it much easier to implement exploration techniques and support other data science efforts. Consequently, Big Data should be understood not only as data sets but also as a set of techniques for processing and analyzing them.

A very popular and often cited model is the 3V model (Fig. 1), which takes into account the three main attributes defining Big Data (Doug, 2001):

- Volume – This feature refers to the quantity and means the high dynamics of data growth requiring advanced technologies for efficient processing;
- Velocity – data is streaming delivered to the organization in real time, which requires sufficient computing power to provide fast response time and isolate only relevant information.;
- Variety – refers to the heterogeneity of the data, the lack of structuring, occurrence in different formats (numerical data, text, image, sound) generated by various sources (internal data of the organization, external data from online sources: social networking sites, purchased databases, etc.).

Figure 1. Model 3V Big Data



Source: Korfiatis (2013)

It can be stated that the presented model is the basis for the development of the definition and enriching it with further attributes. There are 4V, 5V and 6V models in the literature (Fig. 2), which most accurately defines the nature of Big Data. By analyzing the available sources, it turns out that the authors are enriching the basic 3V model with the following features:

- Variability – represents a change in the intensity of data over time and indicates the subordination of flows to certain periodic cycles and trends, eg. the increase in traditional and electronic sales during Christmas, increased interest in hotel services during the holiday season, increased traffic in social media during the election parliamentary (Tabakow, Korczak, Franczyk 2014);
- Veracity – is directly related to reliability and quality of data, it is a data cleaning from unnecessary information that disrupts the final analysis results (Fouad, Oweis, Gaber, Snasel 2015);
- Value – refers to the ability to capture unique information from large data sets that significantly influence organizational efficiency by supporting decision-making and can be an additional revenue stream through the monetization of your information assets (Tabakow, Korczak, Franczyk, 2014).

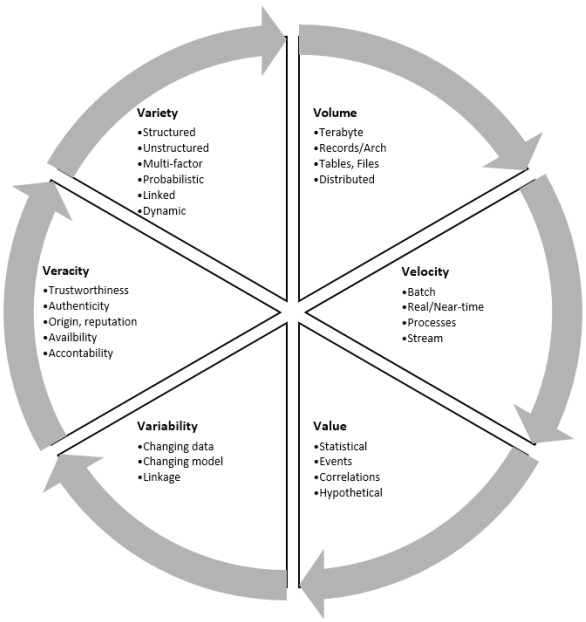
Presented models include the most popular Big Data definitions. However, the 9V model is already in the literature, which extends the 6V model by three further features (Owais, Hussein 2016):

- Validity – is directly related to the truthfulness of the data, means that the data is correct and consistent with purpose, such data is one of the key factors in the decision-making proces;
- Volatility – it refers to variability and draws attention to the instability of some Big Data (structural data retention policies) that can be destroyed after some time;
- Visualization – in order to understand information from mass data volumes, it is necessary to analyze them and then to present the results in the most readable and comprehensible manner to the recipient; Data is presented in the form of infographics, which facilitate the use of Big Data in decision-making.

In addition, the authors of the 9V model share these Big Data features into five categories: Collecting Data, Processing Data, Integrity Data, Visualization Data and Worth of Data. This division is the result of the subsequent stages of the formation and operation of Big Data (Owais, Hussein 2016). A diagram of the distribution of features according to categories is presented in Figure 3.

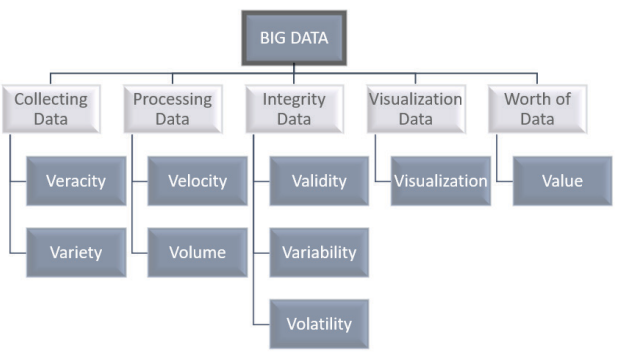
In the literature you can already meet the concepts of Big Data 1.0 and Big Data 2.0. This is reflected in the analogy to the acquisition of Internet technology by the business sphere. At the early stages of Web 1.0, businesses used Web technologies primarily to network, enable e-commerce, and improve the efficiency of their operations. Referring to the Big Data 1.0 concept, it can be stated that the current purpose of the business is to build the right technology base to process large volumes of data to support current business. The economics of the Web 2.0 era have begun to make great use of the possibilities of interaction on the Web. Companies have recognized this potential and are currently using it extensively to improve their offerings and more accurate adjust to consumers needs. Similarly, after mastering the technology of flexible data mass processing, organizations in the Big Data 2.0 era will take full advantage of the information potential they will be able to achieve (Provost, Fawcett, 2013).

Figure 2. Model 6V Big Data



Source: Own elaboration based on (Fouad, Oweis, Gaber, Snasel 2015).

Figure 3. Five Categories CPIVW of the Big Data with their 9 V's Characteristic

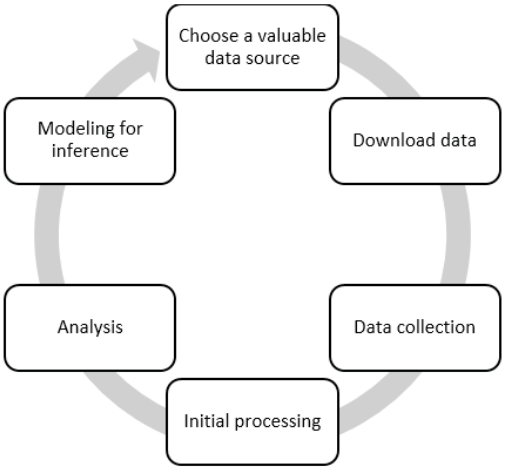


Source: Owais, Hussein (2016).

2. TECHNOLOGICAL BARRIERS TO BIG DATA IMPLEMENTATION

Technological barriers to the development of the Big Data concept stem from the need to build the right architecture, understood as the foundation of the entire data processing cycle. The data mining process is presented in Fig. 3. It has an iterative nature because often after the analysis, there is a need to obtain more data to form more valuable and certain conclusions, so the whole process starts afresh.

Figure 3. Big Data processing cycle



Source: Own elaboration

Technologies used for Big Data analysis should provide as much support for the stages presented in the schematic. One of the most important points is to define relevant data sources from a business point of view. This allows the GIGO (Garbage In, Garbage Out) rule to be avoided, which says that even the best analysis will result in poor results if the entry data is insignificant or erroneous (Zastrożna, 2013, p.69). This requires defining data filters to select only relevant data while not taking unnecessary data, which generates unnecessary costs. Also important is the automatic generation of metadata describing data. Large amounts of data generated through the development of the Internet Things, because of their volume, far exceed the capabilities of today's databases. Difficulties have also begun to make traditional data processing as well as presentation of results. This resulted in a longer waiting time for the results of the analysis. On the other hand, real-time information is expected to be processed on a real-time basis. Streamflow of data generates the need to support current data and update them. This requires modern systems to handle them immediately. Data flow speed is primarily a challenge for the data stack management platform. This is related both to the storage layer and the query processing. Both must be very fast and scalable. Despite continuous development, these are still limited technologies. Variety of data is another problem in this regard. Data appears in different formats and models. There is a need for new technologies and new tools to integrate these data to find the relationships and relationships between the apparently different areas where the data is located. Diversity can also be said about the data structure, rather than the disordered data, which is a separate challenge. There are technologies to deal with different types of data, yet their seamless integration remains a challenge (Tabakow, Korczak, Franczyk, 2014).

In addition to finding the right data processing technology, businesses are struggling to provide the right hardware for these technologies. Enabling the collection and processing of large data sets forces the provision of adequate infrastructure. Big Data collections require dedicated servers. Investing in such solutions is extremely costly. Therefore companies are not too willing to bear such high costs without a return on investment guarantee. On the other hand, the development of cloud computing makes it possible to use external resources, which, compared to investing in your own infrastructure, is much more profitable for economic reasons. Despite the growing popularity of data storage in the cloud, there are still concerns on the part of entrepreneurs about data security and limited access to resources. Many are still convinced that data stored in the cloud is easier to capture by unauthorized individuals than collections stored on internal servers. This is a misleading stereotype because it often appears that data stored within an organization's internal infrastructure is more likely to be vulnerable to cybercriminals. One of the key tasks of cloud computing companies is to provide security. Their activities are largely focused on building cyber-attack barriers, which outperforms the solutions used within the internal resources of many organizations. Another concern with the use of cloud-computing services is the limited access to data resulting from any network access problems, for example, as a result of a failure (Dziembek, 2010; Wielki, 2016).

3. LEGAL AND SECURITY ISSUES

The legal and security concerns associated with the development of the Big Data concept stem mainly from the way data is stored, the already mentioned issue of cloud computing. It is a very convenient form of storage and editing of data, both by individuals and organizations of

all kinds. It allows access to documents from virtually anywhere where, we have access to the Internet.

To understand how strong the relationship is between cloud computing and legal standards and how it affects data security, it is important to briefly explain how cloud computing works. Data stored on the network is not stored in a near-undefined virtual space. They are collected in physical, large server rooms, and the user has access to them through the network. In other words, the files are located on the Internet, but their actual location is physical servers. The location of these servers can be quite important from a data security perspective. It is related to the laws of the country in which the servers are located. Server location information alone provides too much information. In order to explore the subject of data security, you should familiarize yourself with the laws applicable in the areas where the servers of your provider are located. However, this is time consuming and requires some knowledge and skills in the interpretation of legal norms. It is therefore important to consult the service provider's certificates. The ISO 27018 standard is important here. This standard specifies the confidentiality of data and states that the service provider will not use it for marketing purposes and will inform the customer if access is required by the government. In addition, ISO 27018 also imposes an obligation on the provider's staff to maintain absolute confidentiality of the data (Jankowski, 2016). Other noteworthy certifications are the EU Directive 95/46 / EC, which sets out the minimum level of protection of personal data and the standards applicable in the European Union and the United States - ISO / IEC 27001 and ANSi / TIA-942. These are the norms and regulations that should be known both to individuals and to businesses. Consumers should be aware of how their data can be used and in which situations abuse occurs. On the other hand, in the case of entrepreneurs, this knowledge is necessary in order to avoid unintentionally breaking the law by misuse of the customer data, which results in consequences (Wieczorkowski, 2015).

It should be noted that even international certificates and statutory statutory provisions are not replaced by good practices. That is why many of the security issues depend on the users themselves. First and foremost, it is important to remember that user inferiority and poor password setting are a much more common cause of data loss than infrastructure errors. To best protect data, it is best to store them in the cloud, hard disks, and manually back up. This will not eliminate the risk of data loss a hundred percent, but will significantly reduce it (Jankowski, 2016).

4. PROBLEM OF DATA MANAGEMENT AND DATA QUALITY

According to analysts' estimates, more than 60% of the data currently stored in businesses is backed up by overdue information, known as „Bug Data." This is over 3 exabytes of needless data. By the end of 2016, their cost of living has exceeded the threshold of $ 50 billion. Current research shows that 41% of global companies have not been modified for 3 years and 12% have never been open in the last 7 years. Leading IT companies say only 20% of their data is used to streamline business processes, and 33% of them are classified as ROT, Redundant, Obsolete or Trivial. This all points to the importance of keeping a proper classification of data and assigning access to them to individual employees who need to understand enterprise data policies by regularly acquiring knowledge and increasing competence in data analysis (Lachowski, 2016).

Big Data's growing trend has led businesses worldwide to pay more attention to business analytics, data collection and use. Study „Going beyond the data. Turning data from insights into value by KPMG International shows that 82% of companies surveyed admit that using up-to-date data and implementing advanced Big Data analytics in the cloud helps to better understand the needs and preferences of their customers. It also has a significant impact on the business prospects of the company (KPMG International, 2015).

By failing to update the database, which is a source of information about customers, the company can get lost in the huge collection of „Bug Data". The use of data analytics has become one of the main causes of revenue in more than 60% of IT companies. According to 83% of entrepreneurs, their products and services became more profitable and started to generate more profits. 59% of companies consider Big Data to be a key element of their operation, and 29% consider it to be very important for the company's development. Access to relevant, up-to-date data is often crucial in achieving market advantage. Big Data is the digital capital of the enterprise, which simultaneously becomes strategic and development capital. This should focus the company on ensuring the quality of the data, and above all, their actuality. The role of the company is to ensure a proper relationship between the data used and the actual state of knowledge about the market and the consumer (Lachowski, 2016).

According to V. Cerf, the lack of ongoing analysis and updating of data may be one of the reasons for the arrival of „Digital Dark Age". It meant the stage of digitization

development dominated by unstructured, unstructured, unprocessed, raw data, and also archived data, or copies of information that businesses store on their servers. Cerf claims that this kind of data invasion will become the biggest challenge facing data analysts in the coming years (Maffeo, 2015).

5. OBSTACLES IN DATA MONETIZATION

Monetization is defined as „the action of exchanging information products or services for legal means of payment of equivalent value." In other words, it is a digital conversion of capital, which is a large set of data collected by an enterprise - in analogue capital, which can be operated on the market or through which it will be possible to optimize its business activities. Data monetization is therefore an opportunity to directly and indirectly gain knowledge about the digital behavior and interests of the clients of the enterprise or the Internet users. (Ross, Wixom, 2015).

Among the indirect benefits of data monetization are: optimizing business process efficiency, reducing business risk, supporting new product development and exploring new markets, and building and strengthening partner relationships. The direct benefits of data monetization can be seen as market data turnover (data warehousing), expansion of existing products or services, information acquired during data analysis, sales of raw data through brokers and offering access to analyzed data to other companies in the subscription model (Cloud Technologies, 2016).

Availability and freshness of data is the basis for monetization. However, European companies have a serious problem with it. Among the continents best suited to data are Asian companies: as many as 63% of Asian companies claim to be able to generate revenue from their data. This means that it is the Asian market that copes with the Bug Data problem. Second place is the United States, with 58% of the efficiency in the monetization of large data sets. The last place was in Europe, where 56% of the companies surveyed benefited from the monetization of digital information. The recent position of European data monetization companies proves that businesses have a serious problem with separating Big Data from „Bug Data". Their main problem is currently evaluating the usefulness and credibility of Big Data resources, as well as assessing the suitability of data from external sources that companies have not considered so far. (Lachowski, 2016).

6. SUMMARY

With the advancement of digital technology, new data acquisition opportunities have emerged. The volume of data volumes and the speed of their generation so far far outstrip their processing and analysis capabilities. The potential inherent in the development of Big Data is undeniable. It should be noted, however, that the efficient use of large data sets and their transformation into enterprise capital depends on several important factors.

First and foremost, appropriate infrastructure must be provided to enable the processing of Big Data datasets. This is a challenge not only for technology providers but also a big barrier for entrepreneurs. Implementing IT solutions often involves a lot of financial effort, and defining and ensuring return on investment is often difficult to define. In order to reduce the costs of investing in the development of IT infrastructure, some businesses choose to use cloud computing services. However, this solution is associated with certain dilemmas and limitations. The key here is to ensure the security of stored data. Much of the responsibility rests with this situation on entrepreneurs themselves, who should ensure good practice to reduce the risk of data loss. Companies that decide to move their information resources to an external server should also be familiar with the legal regulations that protect their data and verify to what extent they are respected by the service provider. Knowledge of law in this field will not only protect your company data from getting into unauthorized hands, but will also affect the proper use of your data by the company. The vast majority of these are customer data and preferences, so it is important that they are properly protected. Such data is one of the key values of an enterprise, but on one condition: when they are updated and reflect the actual market situation. As numerous studies and reports show, providing the right quality of data is a huge problem. At the moment, most companies are primarily focused on collecting data rather than paying attention to their ordering and updating. This contributes not only to the potential benefits of Big Data, but can also generate unnecessary costs generated by storing large volumes of useless data. Only high-quality current data can provide the basis for monetization, either directly or indirectly, of the company's digital information capital.

ACKNOWLEDGEMENTS

The research was conducted within S/WZ/1/2014 project and was financed from Ministry of Science and Higher Education funds.

REFERENCES

1. Cloud Technologies (2016). *4 kroki do monetyzacji danych w firmie [4 steps to monetize your business]*, <http://biznestuba.pl/biznes-na-zywo/4-kroki-do-monetyzacji-danych-w-firmie/>, 12.05.2017.
2. Doug, L. (2001). Data Management: Controlling Data Volume, Velocity, and Variety, *Application Delivery Strategies*, META Group (currently with Gartner).
3. Dziembek, D. (2010). The cloud computing services in the support of the activity of virtual organization. *Zeszyty Naukowe Uniwersytetu Szczecińskiego*, Vol. 598 No. 58, 289-297.
4. Fouad, M., Oweis, N., Gaber, T., Snasel, V. (2015). Data Mining and Fusion Techniques for WSNs as a Source of the Big Data. *Procedia Computer Science*, Vol. 65, 778-786.
5. Jankowski, P. (2016). *Wszystko o chmurach. Bezpieczeństwo danych w chmurze [All about clouds. Data security in the cloud]*, <http://www.komputerswiat.pl/centrum-wiedzy-konsumenta/uslugi-online/wszystko-o-chmurach/bezpieczenstwo-danych-w-chmurze.aspx>, 05.05.2017.
6. KPMG International (2015). *Going beyond the data. Turning data from insights into value*. <https://assets.kpmg.com/content/dam/kpmg/pdf/2015/08/going-beyond-the-data-turning.pdf>, 07.05.2017.
7. Lee, H., Sohn, I. (2016). *Fundamentals of Big Data Network Analysis for Research in Industry*. John Wiley & Sons Limited, 25-38.
8. Korfiatis, N. (2013). Big Data for Enhancing Learning Analytics: A Case for Large-Scale Comparative Assessments. *Communications in Computer and Information Science*, Vol. 390, 225-233.
9. Maffeo, L. (2015). *Google's Vint Cerf on how to prevent a digital dark age*. <https://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age>, 10.05.2017.
10. McKinsey Global Institute (2011). *Big data: The next frontier for innovation, competition, and productivity*. <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>, 03.05.2017.
11. Owais, S.S., Hussein N.S. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Application*, Vol. 7(3), 254-258.
12. Provost, F., Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media, Inc.
13. Ross, J., Wixom, B. (2015). *Data Monetization: Cashing In on Your Data*, <http://cisr.mit.edu/blog/documents/2015/03/11/1-5-2015_0311_datamonetization_ross-pdf/>, 10.05.2017.
14. Tabakow, M., Korczak, J., Franczyk, B. (2014). Big data –definitions, challenges and information technologies. *Business Informatics*, Vol. 1(31), 138-153.
15. Wielki, J. (2016). An Analysis of the Chances, Opportunities and Challenges Connected with Utilization of Cloud Computing and Big Data as Convergent Technologies. *Annales Universitatis Mariae Curie-Skłodowska, sectio H – Oeconomia*, Vol. 50 No. 2, 163-173.
16. Wieczorkowski, J. (2015). Big Data – Social and Legal Issues, *Nierówności Społeczne a Wzrost Gospodarczy*, Vol. 44, 341-353.
17. Zastrożna, M. (2013). *Google Analytics dla marketingowców [Google Analytics for Marketers]*. Gliwice: Helion.