RESEARCH JOURNAL OF THE UNIVERSITY OF GDAŃSK

**Transport Economics and Logistics** Vol. 71 (2017)
DOI 10.5604/01.3001.0010.5734

**Dariusz Weiland**

*Department of Logistics, Faculty of Economics, University of Gdańsk*

# BIG DATA AS AN INFORMATION SOURCE IN THE DECISION MAKING-PROCESSES OF THE E-COMMERCE COMPANIES

## Abstract

This article discusses the opportunities offered by the use of Big Data in e-commerce and presents this tool as a source of information affecting the decision-making process. Some sections are devoted to introducing and presenting the perspective on information as a resource, while others attempt to define Big Data and outline the way in which Big Data may be utilised as a source of information supply in e-commerce; further parts elaborate on the challenges that information logistics has to face in order to make Big Data more adaptable in e-commerce.

**Keywords:** Big Data, information logistics, e-commerce

## Introduction

The rapid development of information technology is contributing to the ever-increasing growth of e-commerce. Since it is comparatively easy to enter this market and set up one's own business in the virtual space, there is great competition. It is difficult for any corporation to function in such an environment since what has so far been its primary assets now cease to create a competitive advantage. While the e-commerce market was still in its inception, effective marketing strategies were enough to give a company an upper-hand over its competitors, but over time, this stopped producing the desired results. The next stage in the development of e-commerce was attaining the competitive advantage not just by virtue of marketing, but also well-operating logistics; however, as we can now clearly observe, these factors are only enough to keep a company afloat and do not guarantee a permanent competitive advantage. At present, what imparts

a competitive advantage in the e-commerce market is subject to change, hence, being competitive requires the ability to make decisions instantaneously, to deviate from one's habitual strategies and what often amounts to a foundational shift in the basic assumptions lying at the core of a given business (McGrath, 2013). What also makes building a competitive advantage so difficult – except for the rapid growth in the IT sector – are the changes that are happening within the social structure, leading to what is now called an information society and which has its roots in the 90's. This type of society is equipped with modern, highly-developed means of communicating and processing information. Both within and beyond the sphere of e-commerce, there has been a steady trend for the last few years wherein it is the customers who dictate the way companies are evolving, putting forward expectations that said companies find it hard to fulfil. Being able to adjust to these changes depends to a great extent on the ability to make correct decisions, and all the decision-making processes – be it those made by the business owners or the ones made by the customers – are strictly connected to information (Chaberek, 2010). It should thus be noted how crucial information logistics is when it comes to building a competitive advantage. The decision-making process cannot be effective unless it is adequately supplied with certain resources, specifically, information. Hence, information logistics allows one to make adjustments to change more rapidly and expedites the decision-making processes. The primary purpose of this article is to discuss Big Data as a source of information in the decision-making processes of corporations dealing in e-trade. Also in the article presents a collection of tools, which can support the decision-making process for e-commerce businesses.

The article consists of six sections. The first section is an introduction. The second section presents the necessity of considering the information as a resource. The third section contains definitions of "Big Data". This section is based on a review of the "Big Data" literature. Section Four presents "Big Data" as a source of information in the e-commerce sector. In this section, the author presents possible sources of information acquisition and use in e-commerce. This section is based on the author's experience and analysis carried out by the author. Section five presents an analysis of the challenges faced by information logistics. The last section is a summary of the article.

## 1. Treating information as a resource

Information has to be viewed as a fully-fledged resource as it is actively produced, stored and sold by companies. However, we should bear in mind that it is a very specific type of resource, because information itself has certain unique properties, the most crucial of which are:

- inexhaustible renewability of information (including the possibility to process it without causing its depletion),
- substitutability,
- complementarity,
- objectivity of information,

– its virtual nature,
– synergy of information,
– diversity,
– the possibility to endlessly replicate and move it across time and space,
– subjectivity of assessments.

One of the most important features distinguishing information from other resources is the fact that it does not get depleted whilst processing. This means that it can be accessed by multiple subjects simultaneously without any need to renew it. Another unique characteristic is its virtual nature, which means that it is not permanently bound to any one storage device – it may be stored in multiple places without losing its value to the user. What enables endless replication in time and space is precisely the virtual nature of information – it can be transferred between users and storage devices even across large distances.

Treating information as a resource also requires adequate logistics, which term refers to any intentional human activity, business-related or otherwise. The core idea of logistics is controlling the information flow both within and between cooperating business organisations along the logistics chains and channels (Chaberek, 2002). Currently, every action taken by a company involves a wave of information. The utilization of this information should be done in the most beneficial, efficient and effective way possible. Since contemporary, turbulent economy makes this particularly challenging, the role of logistics is of crucial importance as the factor that rationalizes the rudimentary processes (Szmelter, 2013). What handling information within the logistics framework primarily does is create systems enabling circulation, mining, storage and processing of information; it also ensures that both the primary and auxiliary processes are performed correctly. As a result, a company obtains the tools required to run the processes, as well as adequate technical and organisational solutions. The primary goal of these activities is to accomplish the objectives of logistics in reference to information handling, which includes supplying the appropriate information (that is also of sufficient quality), in the appropriate place and time, as well as at a reasonable cost. In order for the information resource to utilised in a profitable, efficient and effective way, it is paramount to design a system of information delivery within a company. Said system will aid the decision-making processes by supplying the appropriate information (i.e. resource), which is vital for the realisation of this process. It is impossible to acquire the appropriate information resources without establishing their source. By analysing the information that companies dealing in e-commerce are handling, one might observe that only a fraction of it is correctly understood and processed. The information handled by a company may be likened to an ice-berg, of which only a small portion is known, while the rest remains invisible and thus unutilized (Wit, 2008). From the point of view of logistics, it is pivotal to make companies realise the extent of the potential that information holds, as well as to demonstrate the means of its acquisition and processing that aim at building competitive advantage. Big Data is one of the sources of such knowledge, which may be viewed as an information handling tool (once the information supplying system is in place).

## 2. Big Data

Big Data, just like many other IT terms, may be defined in a number of ways. The available literature offers at least several dozen different definitions, which leads to the conclusion that the development of modern technologies made it compulsory to alter the definitions in consonance with the current situation and the technical and technological possibilities of qualitative and quantitative processing of large cashes of data of various kinds (McKinsey, 2011). One of the most common definitions of Big Data was introduced in 2001 by an analyst working for Gartner. It centres around three primary properties that characterise Big Data: volume, variety and velocity, or '3V' in short (Doug, 2001; Russom, 2011; Kwon, Sim, 2012; McAfee, Brynjolfsson, 2012; Gartner, 2017). It is precisely this definition that has been updated most frequently through the addition of new components, such as veracity (IBM, 2012), variability, complexity (SAS, 2014). The assumption made within the framework of this article is that Big Data equals immense amounts of data displaying such characteristics as volume, variety, velocity, veracity, variability and complexity (Doug, 2001; IBM, 2012; SAS, 2014), and thus requiring modern architectures and technologies to be obtained and processed (Katal et al., 2013) in order to acquire new information relevant for the decision-making processes, which could not have been utilised had they not been aggregated by means of Big Data. Examples definitions of the term "Big Data" are shown in Table 1.

Table 1. Examples definitions of the term "Big Data"

| Author(s), date | Attributes |
|---|---|
| Doug, 2001<br>Russom, 2011<br>Kwon, Sim, 2012<br>McAfee, Brynjolfsson, 2012<br>Gartner, 2017 | 3V: Volume, Variety, Velocity |
| IBM, 2012<br>IDC, 2012<br>Oracle, 2013<br>Forrester, 2012 | 4V: 3V + Veracity |
| SAS, 2014 | 5V + C: 4V + Variability + complexity |
| Other definitions | |
| Author(s), date | Definition |
| Johnson, 2012 | Big Data: extremely large sets of data related to consumer behaviour, social network posts, geotagging, sensor outputs |
| Davenport et al., 2012 | Big Data: data from everything including click stream data from the Web to genomic and proteomic data from biological research and medicine |
| Manyika et al., 2011 | Big Data: datasets with a size that is beyond the ability of typical database software tools to capture, store, manage, and analyse |
| Rouse, 2011 | Big Data: description of the voluminous amount of unstructured and semi-structured data a company creates or data that would take too much time and cost too much money to load into a relational database for analysis |

Source: (own elaboration based on: Wamba Fosso et al., 2015)

For Big Data to be properly understood, it is crucial to clearly define each of its constitutive features.

'Volume' is characterized by very high dynamics of data acquisition. In 2013, IDC[1] predicted that the amount of global data would rise from 4.4. zettabytes to 44 zettabytes by the year 2020. The current prediction made by IDC estimates that by 2025, the combined data from across the globe will amount to 180 zettabytes. This substantial increase is primarily due to the rapidly growing number of devices that have Internet access. At present, there are approximately 11 billion such devices, which number is estimated to rise to 30 by 2020 and 80 by 2025 (Kanellos, 2016).

'Variety' pertains mainly to the diverse types of data storage devices and such file formats as audio, video, documents, geolocational data, online logins, links to texts, browser queries, etc., all of which derive from a number of different sources (Chang et al., 2006).

The term 'velocity' is to be understood as data that rapidly arises into existence and is characterized by a stream-like flow. The pace at which data is created makes it necessary to constantly update it in such a way as to preserve its value-edgendering properties and its capacity to contribute to the decision-making processes. The pace of data flow has to be divided into two sub-categories, each selectively applied to a different type of streams:

– those data that are newly-created in various places and originate from diverse sources, evolving into an accretion over the already-existing data and data sets;
– and those data that are created as a result of queries made by widely-dispersed users and streams of updates to the already-stored data (Palańska, Wassilew, 2015).

'Veracity' pertains to the matters of trust and uncertainty when it comes to mass data. This hold for both data themselves as well as the results of their analyses (Ward, Barker, 2013). This property is also connected with the value of information. What is meant by 'value' here is the unique information value implicit in large, complex sets of data, which value could not be extracted through a singular analysis of specific items of data (Jinchuan et al., 2013).

The term 'variability' denotes the multiplicity of hidden, potential meanings within the same data sets, which meanings may be extracted depending on the perspective from which said data is analyzed. In other words, certain items of data may yield different information within different contexts of analysis (Vorhies, 2014).

The next property, 'complexity', points to the diversity of data, it covers data organization, data structure, mutual relations and hierarchies of information. By viewing data through the prism of its complexity, it is possible to reveal new, hitherto unrecognized inter-relations between individual items of data.

Proper understanding of Big Data and implementation of mass data sets management within e-commerce may aid companies in coming up to the expectations set before them. Consequently, these companies start building a competitive advantage in both the domestic and international market. If the data sources are

[1] IDC (International Data Corporation) – the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets.

correctly localized, understood and processed, both the customer and the company will find it easier to make decisions (Brown, Chul, Manyika, 2011). The big challenge for the information logistics within Big Data is creating the underlying architecture of information systems that acquire and process these data, and identifying sources from which data may be obtained. Since the number of devices with Internet access is growing and new sources of information are being created, the above definition of Big Data as well its implementations are constantly evolving. Hence, it will be necessary to create systems that will enable adjustments to the ever-changing conditions.

## 3. Big Data as a source of information in the e-commerce sector

Supplying decision-making processes with information is, to a large extent, dependent on access to said information. In order for the information to arise, it is first necessary to acquire adequate data, which will be turned into information in the course of processing. In literature, very often the word "information" is replaced with related terms such as knowledge or data. The relationship between them is shown in Figure 1.

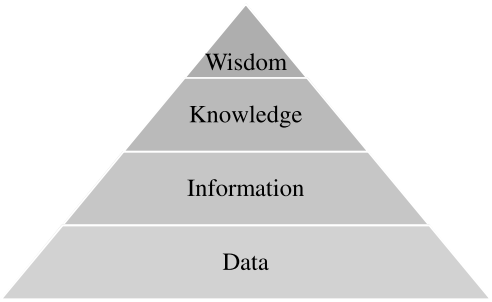


Figure 1. Hierarchy of data, information, knowledge, and wisdom
Source: (own elaboration based on: Heracleous, 1998, p. 155)

The term 'data' encompasses all forms of unstructured, individual facts about a certain object or event. Data may manifest as individual items, which makes their further processing significantly easier, or in data sets (large streams of data), whose processing and rapid changeability require special tools. These might be results of observations or research, raw statistical data or even records of commercial transactions, as well as information pertaining to Internet users – their preferences, interests, etc. Results such as these might constitute raw material used in the process of creating information, which – by contrast with data – has both purpose and meaning. One of the means of transforming data into information is processing it in terms of its intended goal, classifying it, subjecting it to mathematical, statistical and econometrical analysis, and, finally, presenting it visually in form of graphs and charts. While data is being transformed into information,

it is imbued with acquired value, which gives it greater priority and emphasizes its superiority over data.

The e-commerce market in Poland is primarily composed of small and medium-sized businesses, which do not possess adequate financial resources and which often lack the knowledge that is necessary to conduct detailed analyses of information or to implement the usually costly information management systems. However, a properly organized and planned information logistics may allow such companies to build a competitive advantage at a relatively low expense. As previously observed, the information that every e-commerce company has to handle may be compared to an ice-berg. This problem also results from the fact that companies are often incapable of specifying which information is important for them and which is not, they also do not know where information may be obtained from and how to process it. The significance and usefulness of a given piece of information is determined by the following properties: being up to date, relevance, completeness, assimilability and trustworthiness (Bukowski, 2004). The simplest, and yet most commonly underused, sources of information include:

– Google Analytics – this statistics system allows the user to acquire the basic information about the customer, such as: their browsing history, the means through which they found out about a given website, how much time they spent visiting it, what they looked at and how much they spent doing so, what region they accessed it from, what operating system they were using, what are their gender, interests, what were they looking at and clicking. This is one of the first implementations of the Big Data e-commerce technology, and it consists in building algorithms, which recommend goods and services on the basis of prior customer experience, such as the scores they gave to certain products and whether they bought them at all (Cho, Kim, 2004). The next step is to expand on these algorithms by inputting data from Google AdWords and AdSense;
– Advertising systems Google AdWords, AdSense – these make it possible to determine what it is that the customers look for the most, which form of an ad is most effective (graphics, text, video), what the best time to air an ad is and whom the ad attracts – their interests, etc.;
– CRM system – this system allows companies to get to know their customers, build relations with them and strengthen the customer-company bonds. It may also supply information regarding our current customers: who they are, the number of times they contacted us, the time and place of the purchases they made, their birthday, etc. (Kozłowski, Sikorski, 2013);
– Mailing systems – the efficacy of individual topics, which mailings were opened by the customers and when, whether they proved to be effective, what was it exactly that the recipient clicked, what device/software did he use to do this;
– Social Media (Faceboook, Tweeter, etc.) – who is our customer, what did they like, what are their interests, what activities do they engage in;
– Other systems, such as Google Trends, etc.

An e-commerce company will find a variety of applications for the data acquired from the sources listed above. Better adjustments to the pricing across time. Thanks to Big Data and the analysis of data acquired in real-time, it is possible to dynamically

adjust the prices of products to the current situation on the market. Companies have full knowledge of the costs of purchasing the product, as well as of the additional expenses relating to product handling they will have to cover (such as the cost of storage, packaging, transportation, etc.); and thanks to the information acquired through such means as price-comparison websites and applications, a company is able to monitor its prices and those of their competitors, thus becoming better equipped to adjust appropriately.

More efficient utilisation of the advertising budget. Google AdWords and AdSense, to name just a few, made it possible to manage advertising more satisfactorily, as the Google algorithms enable making such adjustments to a given advert so as to make it displayable only to those who are interested in the type of products or services that it is representative of. By obtaining the information about the customer's interests, the websites they frequently visit, their age and combining it with the information acquired through, for instance, remarketing, the advertisement that is displayed on their screen may be perfectly fitted to their needs. Thus, these systems inform the decisions regarding what should be advertised and what form said advertisement ought to assume (text adverts, graphics, video). Additionally, the Cost Per Click system facilitates detailed advertising budget planning, which results in a company paying only for that advertising which proves to be effective, which means the one that attracts potential customers to the shop.

Learning about the clients and their preferences. The analysis of data acquired through such systems as CRM, Google Analytics, AdWords and AdSense, it is possible to get a holistic perspective on the client's sphere of interests. As a result, the decisions regarding, for instance, the release of new products, will be made on the basis of whether the customers are likely to find it appealing and whether it comes up to their expectations.

Creating a personalised offer. Google algorithm-based remarketing allows for the displayed products to be tailored specifically to the customer's requirements. It also furnishes the opportunity to generate networks of related products; for instance, a customer who bought a pair of shoes may be interested in purchasing shoe protection and maintenance products. With the aid of customer information, a personalized sales offer may be designed, one that is informed by both the client's preferences as well as the company's potential to respond to the client's needs.

Optimisation of the sales route. The information acquired through the use of Big Data may facilitate the decision-making processes regarding the sales route. In simple terms, this refers to the path that the customer has to follow in order to finalise their purchase. It is, of course, best if the sales path is optimized in such a way as to meet the customer's expectations. By following customer online behaviour one may acquire the information detailing the specific point at which a given customer refrained from purchase, thus laying the groundwork for the optimization of this precise step so as to encourage them not to quit in the process of completing a purchase.

Improving the management of the supply chain. By utilizing the data from the transportation companies, it is possible to monitor the flow of resources out

of and into a corporation, which enables optimization of said resource flow and close scrutiny of the so-called last mile of the supply chain.

Companies are able to build a competitive advantage through the acquisition and utilisation of information from Big Data. Having the information acquired through Big Data at their disposal, companies are capable of building their advantage. This, coupled with comprehensive understanding of the fundamental nature of today's e-commerce sector and of the expectations that customers have of the corporations, will allow the companies to get ahead of their competition in an ever-changing environment. The primary function of information logistics is to support the decision-making processes on part of the customer and the company. Applying the main objectives of logistics (5R) to information handling facilitates smooth functioning of said processes as well as the accomplishment of the company's goals.

## 4. The challenges for the application of information logistics in e-commerce

Big Data is undoubtedly of great benefit to the companies that use it, and not just those from the e-commerce sector. Virtually any business organisation can identify the most suitable sources of mass data and process said data into information. Yet despite multiple benefits offered by the utilization of mass data in e-commerce, this tools is burdened with many limitations, which are primarily due to the imperfections of the information technology systems within information supply frameworks. Because the amounts of accessible data are so immense and data themselves constantly undergo rapid change, these systems are often stifled and cannot work to their fullest potential. It is necessary to remodel anew the processes of acquiring, processing and utilizing data for the informational purposes and to transform these models into efficiently working information systems. While a great deal of the data obtained from the Internet are described as open-access[2], there are there also a great number of data that are inaccessible and invisible online[3], generating such data is often very costly due to their large size and the necessity to first establish a suitable data infrastructure (servers, networks, etc.), as well as the need to ensure maximum safety and privacy for the sensitive data (Greenberg et al., 2008). The large size of the sensitive data results in high storage, processing and analysis costs, which makes it necessary to seek new (cheaper) means of storing data, taking into account their nature and properties (such as variability, in-flow pace, etc.). The first step towards solving this problem is to implement appropriate technology, such as Apache Hadoop (hadoop.apache.org). It is an open-source technology – which means it is being created by an open community – it is free of charge, though it is frequently plagued by a number of faults, the primary of which is the fact that

[2] Data that is generally available. It is necessary to have the appropriate skills to "see" and to acquire them.

[3] Deep web data, these data are not indexed by standard search engine algorithms, and therefore are not available in standard search results (Wei et al., 2010).

it may only be used by programmers with a lot of experience. Another downside is that its functionality is limited in terms of the queries that may inputted. Perhaps the coming years and the development of Big Data platform will make it more accessible to regular Internet users, but this cannot happen unless there is strict cooperation between programmers and logisticians.

Another important challenge is data cleansing (also known as data cleaning), which is absolutely crucial whenever multi-source data are utilised. When a business organization is analysing date acquired through the Internet, it cannot differentiate between a product that it itself has on offer and a similar one, but belonging to its competition. This results from a lack of appropriate procedures that would enable identification of commodities sold online. Thus, the data cleansing process very time- and labour-consuming, and constitutes one of the greatest obstacles for the practical implementation of Big Data (Rahm, Do, 2000).

It is not only data storage and acquisition that cause the costs of mass data utilization to be so high, it is the need to process them in order to obtain information. Both basic, descriptive, statistical analysis and more elaborate econometric models may be hindered by the vastness of the amounts of data available and their high degree of changeability.

What also proves to be a problem in many of these cases is current technology and the omission of many significant components of information logistics. However, it is highly likely that once these processes are explored and understood on a deeper level, as well as re-modelled by logisticians cooperating with programmer, the functionality and adaptability of Big Data will vastly expand.

## Conclusions

Professional utilisation of the Big Data tool in e-commerce is still plagued by a number of limitations resulting from the fact that this technology is in early stages of development. The tools that are currently used are still insufficiently explored and complicated. Nonetheless, the following report attempted to demonstrate the possibility of implementing Big Data in e-commerce without necessarily incurring great costs. Big Data already is an ideal source of information supply in the decision-making processes. The benefits that Big Data offers for a company include better-informed decisions regarding the selection of products on offer, their prices, advertising costs, client service, etc. It is also of great value to a customer, since a company that can handle information efficiently is able to aid its customers while they are deciding which product to choose, where to buy it, etc.

Further research into the means of acquiring and also processing and analyzing large sets of data may make Big Data more successful in future, thus creating something that not only bolsters e-commerce and constitutes a source of information, but also shapes new strategies in the e-commerce sector.

## References

Brown, B., Chul, M. and Manyika, J. (2011), Are you ready for the era of 'big data'?, *McKinsey Quarterly*, October.

Bukowski, L. (2004), *Problemy przetwarzania informacji logistycznych w zintegrowanych systemach produkcyjnych. Wybrane zagadnienia logistyki stosowanej, Materiały VII Konferencji Logistyki Stosowanej – Total Logistic Management*, Oficyna Wydawnicza TEST, Kraków.

Chaberek, M. (2002), *Makro- i mikroekonomiczne aspekty wsparcia logistyczneg*o, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, p. 15.

Chaberek, M. (2010), Praktyczne i teoretyczne aspekty kontaminacji i atomizacji logistyki i informatyki ekonomicznej, In: Chaberek, M. and Jezierski, A. (Eds.), *Informatyczne narzędzia procesów logistycznych*, CeDeWu, Warszawa, p. 13.

Chang, C., Kayed, M., Girgis, M.R. and Shaalan, K.F. (2006), A survey of web information extraction systems, *IEEE Transactions on Knowledge and Data Engineering*, 18(10), pp. 1411–1428.

Cho, Y.H. and Kim, J.K. (2004), Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce, *Expert systems with Applications*, 26(2), pp. 233–246.

Davenport, T.H., Barth, P. and Bean, R. (2012), How Big Data Is Different, *MIT Sloan Management Review*, 54(1), pp. 43–46.

Doug, L. (2001), *Data Management: Controlling Data Volume, Velocity, and Variety*, "Application Delivery Strategies", META Group (currently with Gartner).

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), How 'big data' can make big impact: findings from a systematic review and a longitudinal case study, *International Journal of Production Economics*, 165, pp. 234–246.

Forrester (2012), *The Big Deal About Big Data For Customer Engagement*. Available from https://www.forrester.com/report/The+Big+Deal+About+Big+Data+For+Customer+Engagement/-/E-RES72241 [Accessed 15 September 2017].

Gartner (2017). Available from http://www.gartner.com/it-glossary/big-data/ [Accessed 15 September 2017].

Greenberg, A., Hamilton, J., Maltz, D.A. and Patel, P. (2008), The cost of a cloud: research problems in data center networks, *ACM SIGCOMM Computer Communication Review*, 39(1), pp. 68–73.

Heracleous, L. (1998), Better than the rest: Making Europe the leader in the next wave of innovation and performance, *Long Range Planning*, February.

IBM (2011). Available from https://www-01.ibm.com/software/in/data/bigdata/ [Accessed 28 April 2017].

IDC (2012). Available from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf [Accessed 15 September 2017].

Jinchuan, C., Yueguo, C., Xiaoyong, D., Cuiping, L., Jiaheng, L., Suyun, Z. and Xuan, Z. (2013), Big data challenge: a data management perspective, *Frontiers of Computer Science*, 7(2), pp. 157–164.

Johnson, B.D. (2012), The Secret Life of Data, *The Futurist*, 46(4), pp. 20–23.

Kanellos, M. (2016), *Amount of Data Created Annually to Reach 180 Zettabytes in 2025*. Available from https://whatsthebigdata.com/2016/03/07/amount-of-data-created-annually-to-reach-180-zettabytes-in-2025/ [Accessed 28 April 2017].

Kozłowski, R. and Sikorski, A. (2013), *Nowoczesne rozwiązania w logistyce*, Wolters Kluwer business, Warszawa.

Kwon, O. and Sim, J.M. (2012), Effects of data set features on the performances of classification algorithms, *Expert Systems with Applications*, 40(5), pp. 1847–1857.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), *Big data: the next frontier for innovation, competition and productivity*, Report McKinsey Global Institute, May.
McAfee, A. and Brynjolfsson, E. (2012), Big data: the management revolution, *Harvard Business Review*, 1.
McGrath, R.G. (2013), Transient Advantage, *Harvard Business Review*, 91(6), pp. 62–70.
Oracle (2013). Available from http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf [Accessed 15 September 2017].
Polańska, K. and Wassilew, A. (2015), Analizy "big data" w serwisach społecznościowych, *Nierówności Społeczne a Wzrost Gospodarczy*, 44(4), p. 118.
Rahm, E. and Do, H.H. (2000), Data cleaning: Problems and current approaches, *IEEE Data Engineering Bulletin Issues*, 23(4), pp. 3–13.
Rouse, M. (2011), *big data (Big Data).* Available from http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data [Accessed 15 September 2017].
Russom, P. (2011), *The Three Vs of Big Data Analytics*, TDWI.
SAS (2014). Available from https://www.sas.com/en_us/insights/big-data/what-is-big-data.html [Accessed 28 April 2017].
Szmelter, A. (2013), Business Intelligence jako element systemu zaopatrzenia informacyjnego, *Roczniki Wyższej Szkoły Bankowej w Toruniu*, 12, p. 129.
Vorhies, W. (2014), *How many V's in Big Data? The Characteristics that Define Big Data.* Available from http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data [Accessed 28 April 2017].
Ward, J.S. and Barker, A. (2013), *Undefined by Data: A Survey of Big Data Definitions.* Available from https://arxiv.org/pdf/1309.5821v1.pdf [Accessed 28 April 2017].
Wei, L., Xiaofeng, M. and Weiyi, M. (2010), ViDE: A Vision-Based Approach for Deep Web Data Extraction, *IEEE Transactions on Knowledge and Data Engineering*, 22(3), pp. 447–460.
*What is Apache Hadoop?* (2014). Available from https://hadoop.apache.org [Accessed 28 April 2017].
Wit, B. (2008), *Electronic commerce – budowanie konkurencyjności przedsiębiorstwa w Internecie*, Wydawnictwo Politechniki Lubelskiej, Lublin.

## Corresponding author

Dariusz Weiland can be contacted at: d.weiland@ug.edu.pl