**UDC 311.21:004**

**M. Tashkova**,
*PhD (Economics), Head Assistant,*
*D. A. Tsenov Academy of Economics, Bulgaria*

# BIG DATA – THE NEW CHALLENGE FACING BUSINESS

It is almost certain that in the following years Big Data will be among the leading information technologies when discussing IT trends. Moreover, it is one of the pillars of the 'third platform'[1] as defined by IDG. While in a sense this term is new, large volumes of data are not infrequently handled in computer processing. Companies are aware that data – internal and external – is an important source of self-knowledge, which would help them to improve their business processes and productivity. Therefore, they are looking for technologies that will enable them to collect and analyze this data. This makes the issue of information technologies designed to manage large volumes of data both topical and important in the future.

**Specifics of Big Data**

The term 'Big Data' has been in use quite recently. 'Big Data' was first mentioned in 1997 in the paper 'Application-Controlled Demand Paging for Out-of-Core Visualization', presented at the eighth Conference on Visualization organized by IEEE[2]. In a very short time it gained popularity, but big data and related technologies became popular after 2008. The catalyst of this popularity is the thematic issue of the Nature Journal in 2008, which was entirely devoted to big data. The initiative was followed by a number of prominent journals which came out with separate issues devoted to Big Data[3]. In 2011, Gartner analysts emphasized Big Data in their traditional report. In it they noted that big data technologies focus attention, but even the IT industry itself cannot predict what potential lies in them. Several years later, these technologies turned from emerging into promising for the world of technology.

Today there are many sources that create large volumes of data. On the one hand, these are the people themselves, who are able to generate information on their own using the Internet and social networks. An example of how much information is created today and at that globally is the fact that the information generated all through 2000 was less than the information generated in one minute during the past 2016, and nearly 90% of the files available in the world were created in the last two years.

On the other hand, with the advent of the Internet of Things it is not only users that generate information, but also most of the devices that are used daily. Sensors are a great source of data. Only in half an hour the sensors of jet engines generate about 10 TB of data. Roughly the same data streams are generated by sensors installed on drilling oil rigs.

In order to make the above stated more vivid, it should be added that worldwide[4]:

- corporate users store nearly 7 EB[5] of information and personal users – 6 EB annually;
- 30 billion new sources of information are published in the social network Facebook each month;
- one of the leading sources of big data is the Twitter short message service, through which data of about 8 TB per day is created, despite the restrictions on message length (140 characters);
- the mobile operators networks worldwide serve more than 5 billion phones and manage the audio-video stream generated by them;
- only US companies from 15 sectors create data that is larger than the data in the Library of Congress[6].

All this logically leads to the question – what is meant by the term 'big data'? A simple question with many answers, since different users see 'big data' differently. For one company, a large volume is 10 TB, while for another it may be 100 TB. And does only the volume determine whether certain data is big?

In order to give a simple definition of the nature of big data the paradigm Big Data '5V' is usually referred

[1] The 'Third Platform' concept was presented by the IDG analysts in 2012 and the aim was to highlight the global transformation of information technologies. The four technologies gaining popularity – mobile applications and devices, cloud services, analyses of large volumes of data and social networks – underlie the platform.

[2] The idea of 'Big Data', however, emerged considerably earlier. In 1975 the Japanese Ministry of Posts and Telecommunications started a quantitative study of the information flow in Japan, but the idea for this quantification was proposed as early as 1969.

[3] These were CACM (2008); TheEconomist (2010); Science (2011).

[4] McKinsey& Company, Big data: The next frontier for innovation, competition, and productivity [Electronic resource]. – Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[5] EB – exabyte of data, where 1 EB =$10^{18}$ B.

[6] The Library of Congress of the USA stores 235 TB of data.

to. According to it data that has simultaneously at least two of the following specific features[1] is considered big data:

– *Volume of data*. The possibility to handle large volumes of information is a key feature of Big Data, and the sources for generating information are different – business transactions, social media, information from sensors, data transmitted from machine to machine;

– *Velocity of data accumulation*. No less important parameter in big data is the frequency of their creation and change, as well as the speed of processing them and obtaining the results in near real-time;

– *Variety of data*. Big data is not always structured. It can also include video, audio, email, unstructured documents, messages from social services and media. That is why organizing it in relational databases is in fact already very difficult;

– *Veracity of data*. This refers to the purity and authenticity of the information generated. Big data must have the necessary reliability in order for the analysis to be accurate. This is achieved by pre-filtering of data to overcome the noise and anomalies in it;

– *Value of data*. This feature renders how effective data processing is in relation to investments made since the realization of infrastructure, the systems for storage and processing of big data are a relatively large expense for companies.

In a synthesized form, however, big data could be defined as hardware and software products for processing huge volumes of structured and unstructured data, which are characterized by great diversity. Using different methods of processing and analyzing data new hypotheses and models are discovered, which could be used effectively to optimize the business processes of companies.

**IT for big data management**

It is a fact that large volumes of data are generated today and this is an upward trend. The challenge now is – what technologies these large volumes of data could be managed with. The popular technologies for handling big data involve:

- *NoSQL databases*

NoSQL technologies gained recognition after the large IT manufacturers – Google with Big Table; Facebook with Cassandra; Amazon with SimpleDB; Twitter and SourceForge with MongoDB; Yahoo with Sherpa; Adobe with Hbase – got interested in them. These IT manufacturers practically need to overcome the limitations of relational databases and to move to a new model of data organization and storage, as unlike relational databases, NoSQL is characterized by:

– absence of a predefined schema. With NoSQL there is no need to know in advance what data will be stored in the database and hence it is not compulsory to create a schema. This allows NoSQL to maintain dynamic data changes as it is not tied to a specific structure;

– easy scalability. NoSQL has distributed and fault-tolerant architecture as the information sites are stored on several servers. The advantage of NoSQL is that it can easily be expanded horizontally by adding additional servers without using additional logic applications for this;

– finer control over available information.

According to their purpose NoSQL databases can be:

– Key-value. These databases provide generally one operation – retrieving a unit value through its key.

– Column. In these databases the records are stored in columns rather than in rows as in relational databases.

– Graph. Here each element can be connected to an unlimited number of relationships;

– Document. In document-based databases the values in the primary key value are referred to as a document. An identifier is used in place of the key.

NoSQL databases share common principles, but solve various needs. Therefore a certain NoSQL is assigned to perform one or more specific tasks, but the core functionality is implemented with a relational database or another NoSQL database.

- *Apache Hadoop*

The Hadoop platform is designed to organize distributed processing of large volumes of data, and uses the model of separation and collection, i.e. each task is divided into smaller parts and each part of the set is performed on a separate node of the cluster. Hadoop is written in Java and consists of different components. The main ones are:

– MapReduce is a combination of two interrelated functions performing data processing on a particular totality. First the Map-function is performed, which reads data from an input file, performs the necessary filtering and transformation, then generates a set of input records consisting of data and its assigned keys. Each of the Map programs operates independently of the others on its node in the cluster. Its task is to retrieve data, search and sort. Reduce has the opposite task – to unite, summarize, filter or modify the data processed and to record the results.

– Hadoop Distributed File System (HDFS) is a distributed file system for data storage. HDFS is designed to store very large files with streaming access

[1] The paradigm passed through the 3V and then 4V stages, and in less than ten years two new features characterizing big data were added.

In its study 'Big and open data: A growth engine or a missed opportunity' the Warsaw Institute for Economic Research added these two new features in order to measure the need for the 'Big Data' accumulation in the EU economy.

data patterns. It is also designed to work on clusters of inexpensive hardware, ensuring data availability without losses, even when some servers do not work.

– Pig and Hive – SQL-like languages for constructing MapReduce applications on large volumes of data. Pig requires MapReduce and HDFS, while Hive – Data Warehouse and HDFS.

The use of Hadoop technologies in handling large volumes of data reduces the time for their processing and their equipment costs, increases sustainability and performs painless horizontal scalability.

**Applied aspects of big data in business**

A year ago, the major users of big data technologies were IT giants such as Facebook or Yahoo, willing to analyze their Web users' routing information. Today these technologies are used by any company successfully positioned in the market. Modern electronic users readily leave their personal data in the digital space. Their purpose is to obtain certain information or to gain access to the desired content. Through forums, social networks, online polls, they make their likes and dislikes public. Using this data by means of the big data tools allows for carrying out various business analyses and forecasts. The major *economic industries* that are increasingly interested in the possibilities of big data at the moment are:

- *commerce*. Or rather retailing in realizing sales to the end user. Here, the use of big data is directed to:

– automatic forecasting of consumer demand;

– optimal planning of promotional campaigns;

– conducting an effective pricing policy and marketing strategies;

– responding to any market fluctuation if necessary.

- *banks*. In banks big data is used in:

– assessing the creditworthiness of a bank's customer;

– offering banking products personally;

– receiving information promptly;

– preventing suspicious transactions.

- *telecommunications*. Here big data is used for:

– customer segmentation;

– studying the preferences and assessing the profitability of different user groups;

– managing customer loyalty.

The power of big data technologies involves working and making business decisions using the maximum amount of data. The more data, the more substantiated analyses and forecasts. In this regard some good practices in the use of big data can be presented.

- *Amazon*. Creating and changing business models.

Until recently shopping from Amazon was the typical online shopping. The power of big data, however, made it possible to offer customers an electronic shop assistant through which information about each transaction with each buyer is stored. This information is used in real time to identify and offer what the customer needs. Amazon is expected to provide a service for predelivery of goods[1]. It is used to send goods for which the Big Data analyst has forecasted that the customer will also pay.

- *Facebook and Google*. Data collection and analysis.

Facebook and Google have turned big data collection and analysis into business models. Both giants use the available data to attract analysts and advertisers. Much of the revenues of the companies are namely from advertising and sale of data to carry out market research. The difference between the business models of the two IT giants is in the data source. With Facebook this is the personal information posted by users in the social network. With Google the data is from the free services that the corporation offers its users such as GoogleSearch, Gmail, YouTube, GoogleTalk, GoogleDocs, Google+, GoogleMaps.

**Conclusion**

The problem is not that organizations create huge volumes of data, but the fact that they cannot use them optimally. A large part of this data is not structured and the traditional relational databases do not have tools to process them effectively. Taking into account the frequent data update, it could be concluded that an alternative to the traditional methods of data analysis today is the technologies to manage big data. Knowing and using those leads to the following advantages:

- the possibility to analyze huge volumes of information, unlike the currently known approach to draw conclusions on limited data (excerpts)[2];
- management of information in its actual state and not of purified (ideal) data[3];
- finding out the interdependences in the data and not looking for a specific causality.

Big data is and will continue to be important for the business and the IT sector. The combination of social data, mobile applications and CRM records, allows to create forecasts and successful models of corporate behaviour. Using the technologies for processing big data, companies get the IT tools which make it possible to:

- analyze large volumes of data related to:

– profitability and customer behaviour;

– operational analyses.

- create predictive models for:

– the customer's market interests;

[1] The grounds for this are the patents, which are registered by Amazon.

[2] Weaknesses of random sampling – up to 3 % error; lack of detail.

[3] Data warehouse (DW) is often called online analytical processing (OLAP).

– the volumes of production and sales of goods and services;

– the creditworthiness of corporate and personal customers.

- optimize business processes based on predictive models.

**References**

1. **Atanasova, G. et al.** Tehnologichni aspekti na modela '3Vs' za predstavyane na golemi obemi ot danni. // Scientific Works of the University of Ruse, 2014, Vol. 53, series 6.1. 2. **Krasteva, N.** Platformata Apache Hadoop v1.0 priema predizvikatelstvoto na 'golemite danni' // CIO, Issue 2-2012. 3. **Leboeuf, K.** The 5 vs of Big Data: predictions for 2016. 4. http://www.excelacom.com/resources/blog/the-5-vs-of-big-data-predictions-for-2016. 5. **McKinsey & Company**, Big data: The next frontier for innovation, competition, and productivity http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. 6. **Yossi, A.** What happens in an Internet minute? How to capitalize on the Big Data explosion, 2015 http://www.excelacom.com/resources/blog/what-happens-in-an-internet-minute-how-to-capitalize-on-the-big-data-explosion. 7. **Apache Hadoop**, http://hadoop.apache.org/.

**Ташкова М. А. Великі дані – новий виклик, що стоїть перед бізнесом**

Сьогодні існує багато джерел, які створюють великі обсяги даних. З одного боку, – це самі люди, які здатні генерувати інформацію самостійно, використовуючи Інтернет і соціальні мережі. З іншого боку, з появою Інтернету речей ними стають не тільки користувачі, які генерують інформацію, але й більшість пристроїв, які використовуються користувачами щодня. Сила технологій переробки великих даних передбачає роботу і прийняття бізнес-рішень, використовуючи максимальну кількість даних. Чим більше даних, тим більше обгрунтовані аналізи і прогнози. Виклик полягає у такому: за допомогою яких технологій переробка цих великих обсягів даних може здійснюватися.

*Ключові слова:* великі дані, обсяг, швидкість, різноманітність, достовірність, вартість.

**Ташкова М. А. Большие данные – новый вызов, стоящий перед бизнесом**

Сегодня существует множество источников, которые создают большие объемы данных. С одной стороны, это сами люди, которые способны генерировать информацию самостоятельно, используя Интернет и социальные сети. С другой стороны, с появлением Интернета вещей ими становятся не только пользователи, которые генерируют информацию, но и большинство устройств, которые используются пользователями ежедневно. Мощь технологий обработки больших данных подразумевает работу и принятие бизнес-решений, используя максимальное количество данных. Чем больше данных, тем более обоснованными будут анализы и прогнозы. Вызов заключается в следующем: с помощью каких технологий переработка этих больших объемов данных может осуществляться.

*Ключевые Слова:* большие данные, объем, скорость, разнообразие, достоверность, стоимость.

**Tashkova M. A. Big data – the new challenge facing business**

Today there are many sources that create large volumes of data. On the one hand, these are the people themselves, who are able to generate information on their own using the Internet and social networks. On the other hand, with the advent of the Internet of Things it is not only users that generate information, but also most of the devices that are used daily. The power of big data technologies involves working and making business decisions using the maximum amount of data. The more data, the more substantiated analyses and forecasts. The challenge now is – what technologies these large volumes of data could be managed with.

*Keywords*: big data, NoSQL, hadoop, volume, velocity, variety, veracity, value.

Received by the editors: 02.12.2016
and final form 28.12.2016