

Brief overview of the concept of big data

Stella Hrehova¹
{stella.hrehova@tuke.sk}

Faculty of Manufacturing Technologies with seat in Presov, Technical University in Kosice, Bayerova 1,
080 01 Presov, Slovakia¹

Abstract. The term "Industry 4.0" originates from a project in the high-tech strategy of the German government, which promotes the computerization of manufacturing. The vision of this concept is digital enterprises that will manage their processes through data. This data is usually owned, but it is not processed and evaluated enough. The success of the enterprises in conjunction with Industry 4.0 therefore depends on the data. With new options, the volume of this data is growing and structure is changed and term "big data" is used. The concept of big data is not new, but with the development of new technologies its significance changes. The aim of this paper is to provide basic information on this phenomenon at the time of ongoing technological and application changes in the company.

Keywords: big data, definitions, dimensions

1 Introduction

The ability to effectively manage information and extract knowledge is now seen as a key competitive advantage [1]. The growth of the Internet of Things (IoT) in commercial and industrial usage is one of the main drivers of Big Data [2]. With the development of new, more powerful data collection, processing and data acquisition technologies the more things are changed, the more changes are captured and recorded as data [3].

- Most data collected now is unstructured and requires different storage and processing than that found in traditional relational databases.
- Available computational power is sky-rocketing, meaning there are more opportunities to process big data.
- The Internet has democratized data, steadily increasing the data available while also producing more and more raw data.

So, why does the traditional data differ from so-called big data? According of [4], [5] there are these features:

- big data can be an entirely new source of data,
- huge in volume, consisting of terabytes or petabytes of data,
- the speed of data feed has increase to such an extent that it qualifies as a new data source, high in velocity, being created in or near real-time,
- increasingly more semi-structured and unstructured data are coming in,
- exhaustive in scope, striving to capture entire populations or systems (n=all),
- fine-grained in resolution and uniquely indexical in identification,

- relational in nature, containing common fields that enable the conjoining of different data sets,
- flexible, holding the traits of extensionality (can add new fields easily) and scalability (can expand in size rapidly).

2 Definitions

Over the past few years, the use of this term has become more and more frequent in terms of the need to define it. Various definitions can be found in the literature and on the Internet. The following table presents some of them [1], [3], [6], [7].

Table 1. Chosen definitions of term “Big Data”.

Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization	Laney (2001), Manyika et al. (2011)
Big Data is “data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time	Jacobs (2009)
When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam	Loukides (2010)
Big Data is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies.” By extension, the platform, tools and software used for this purpose are collectively called “Big Data technologies	NESSI (2012)
Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.	Gartner (2012)
The term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications	Wikipedia (2014)
Big data refers to data sets so large and complex that they are difficult to process using traditional ICT applications.	European parliament (2016)
Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which uses massive parallelism on readily-available hardware.	Techopedia 2018

In the initial descriptions, the big data was defined by the three essential dimensions. Those dimensions are marked as 3Vs.

- **VOLUME** - is the V most associated with big data because, volume can be big. What we're talking about here is quantities of data that reach almost incomprehensible proportions (tera, peta, zetta byte) [8].
- **VARIETY** - data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications, structured data as database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two [4].
- **VELOCITY** - refers to the speed with which the data is generated, analysed and reprocessed. Today this is mostly possible within a fraction of a second, known as real time [9].



Fig. 1. Three essential dimensions of big data.

These three basic features were complemented by another feature in 2012 which was presented by Schroeck in his work – VERACITY [6].

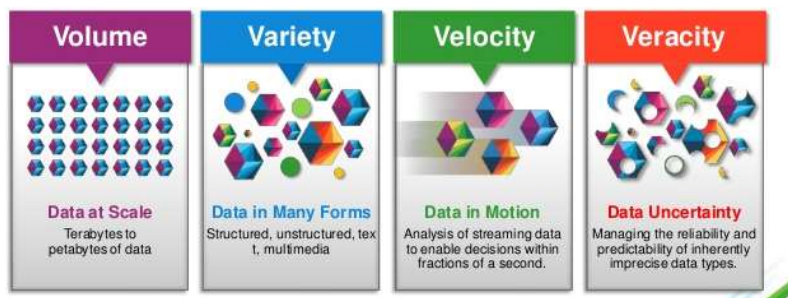


Fig. 2. Newer form of V's attributes of big data.

Its description according to [10] is:

- VERACITY refers to the noise and bias in the data as one of the biggest challenges to bringing value and validity to big data.

Having endless amounts of data is one thing, but unless it can be turned into value it is useless [11]. For this reason has been added the next dimension of big data:

- VALUE - Of course, data in itself is not valuable at all. The value is in the analyses done on that data and how the data is turned into information and eventually turning it into knowledge. The value is in how organisations will use that data and turn their organisation into an information-centric company that relies on insights derived from data analyses for their decision-making.

With the increasing areas of use of this data, additional and further "dimensions" have gradually begun to be added. The 10 V's shows the next figure.



Fig. 3. 10 V's attributes of big data [12].

3 Conclusion

The power of big data is in the analysis we do with it and the actions we take as the result of the analysis. Big data or small data does not in by itself possession any value. It is valuable only when we can get some insight out of the data. And that insight can be used to guild our decision making [4].

This paper offered you a brief overview of the concept of big data to understand its power and possibilities of its use in different area.

Acknowledgments. This work was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0602.

References

- [1] Curry, E.: The Big Data Value Chain: Definitions, Concepts, and Theoretical approaches. In: *New Horizons for a Data-Driven Economy*, ed. J.M. Cavanillas, E. Curry, W. Wahlster, Springer, pp. 29 – 37 (2016)
- [2] Kho, N. D.: *The State of Big Data*. Ecotent, Vol. 41, pp.11-12, (2018)
- [3] Techopedia <https://www.techopedia.com/definition/27745/big-data> (accessed 22 August 2018)
- [4] Su, X. Introduction to Big Data. <https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf> (accessed 22 August 2018)
- [5] Kitchin, R.: Big Data and Official Statistics: Opportunities, Challenges and Risks, *Statistical Journal of the International Association of Official Statistics* Vol.31(3) pp. 471-481 (2015)
- [6] Ylijoki, O., Porras J.: Perspectives to Definition of Big Data: A Mapping Study and Discussion, *Journal of Innovation Management, JIM* 4,1 pp. 69-91 (2016)
- [7] European parliament, (2016) [http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589801/EPRS_BRI\(2016\)589801_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589801/EPRS_BRI(2016)589801_EN.pdf) (accessed 22 August 2018)
- [8] Gewirtz, D.: Volume, velocity, and variety: Understanding the three V's of big data [Online] <https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/>
- [9] What is Big Data? – A definition with five Vs. <https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs> (accessed 22 August 2018)
- [10] Hammer, C.L., Kostroch, D.C., Quirós, G.: Big Data: Potential, Challenges, and Statistical Implications, (2017), [Online]: <https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2017/09/13/Big-Data-Potential-Challenges-and-Statistical-Implications-45106>
- [11] Cano, J. The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity, (2014) <https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity> (accessed 22 August 2018)
- [12] Arockia P.S., Varnekha, S.S., Veneshia, K.A.: The 17 V's of Big Data. *International Research Journal of Engineering and Technology (IRJET)*, Vol.4 (9) pp. 329 – 333 (2017)