Providing Clarity on Big Data: Discussing Its Definition and the Most Relevant Data Characteristics

Matthias Volk[a], Daniel Staegemann[b] and Klaus Turowski
Magdeburg Research and Competence Cluster VLBA, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Keywords: Big Data, Definition, Data Characteristics, Overview.

Abstract: For many years now, the domain of big data has received lots of attention, as numerous studies, reports, and research articles reveal. Up to this date, a multitude of different definitions, guidelines, technologies, architectures, and best practices appeared that were supposed to provide clarity in the jungle of existing solutions. Instead, it led to further confusion regarding the general nature and the applicability of big data. To overcome those obstacles in detail, in the following, a thorough description of the term big data. To elucidate those obstacles, in the following, some of the most important aspects regarding the terminology as well as the essence of big data as a concept are discussed. Hereby, subsequent researchers as well as practitioners interested in the domain are provided with a strong base that further considerations and future projects can be built upon. This, in turn, hopefully helps to facilitate the progression of the domain as a whole as well as its further proliferation.

1 INTRODUCTION

For many years now, the domain of big data has received lots of attention, as numerous studies, reports, and research articles reveal (Staegemann et al. 2019). Up to this date, a multitude of different definitions, guidelines, technologies, architectures, and best practices appeared that were supposed to provide clarity in the *jungle* (Volk et al. 2019) of existing solutions and help to harness the immense potential of the concept (Müller et al. 2018). Instead, this deluge led to further confusion regarding the general nature and the applicability of big data. To overcome those obstacles, in the following, some of the most important aspects regarding the terminology as well as the essence of big data as a concept are discussed. First, a short historical outline is presented at which existing definitions are discussed. Since the underlying data characteristics are not only the foundation for most of these approaches but also essential to the overall understanding of the domain itself, these are discussed afterward. Hereby, subsequent researchers as well as practitioners interested in the domain are provided with a strong base that further considerations and future projects can be built upon. This, in turn, hopefully helps to facilitate the progression of the domain as a whole as well as its further proliferation.

2 DEFINITION

In the yearly published hype cycle of the business analytics company Gartner, the maturity and adoption of recent technologies are graphically represented according to the five key-phases that each of those experience during their life cycle. Namely, those are the innovation trigger, the peak of inflated expectations, the trough of disillusionment, the slope of enlightenment, as well as the plateau of productivity (Gartner 2022b). Throughout those stages, each depicts whether new technology can be seen as hype or an ideal market solution that creates revenue. An example of the hype cycle in 2011 is provided in Figure 1.

In 2011 the term big data appeared for the first time in that life cycle (Gartner 2011). After becoming a hype topic quickly in the following years, in 2014 it passed the peak of inflated expectations (Gartner 2014) before it vanished completely in 2015 (Gartner 2015).

[a] https://orcid.org/0000-0002-4835-919X
[b] https://orcid.org/0000-0001-9957-1003

Volk, M., Staegemann, D. and Turowski, K.
Providing Clarity on Big Data: Discussing Its Definition and the Most Relevant Data Characteristics.
DOI: 10.5220/0011537500003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 3: KMIS*, pages 141-148
ISBN: 978-989-758-614-9; ISSN: 2184-3228

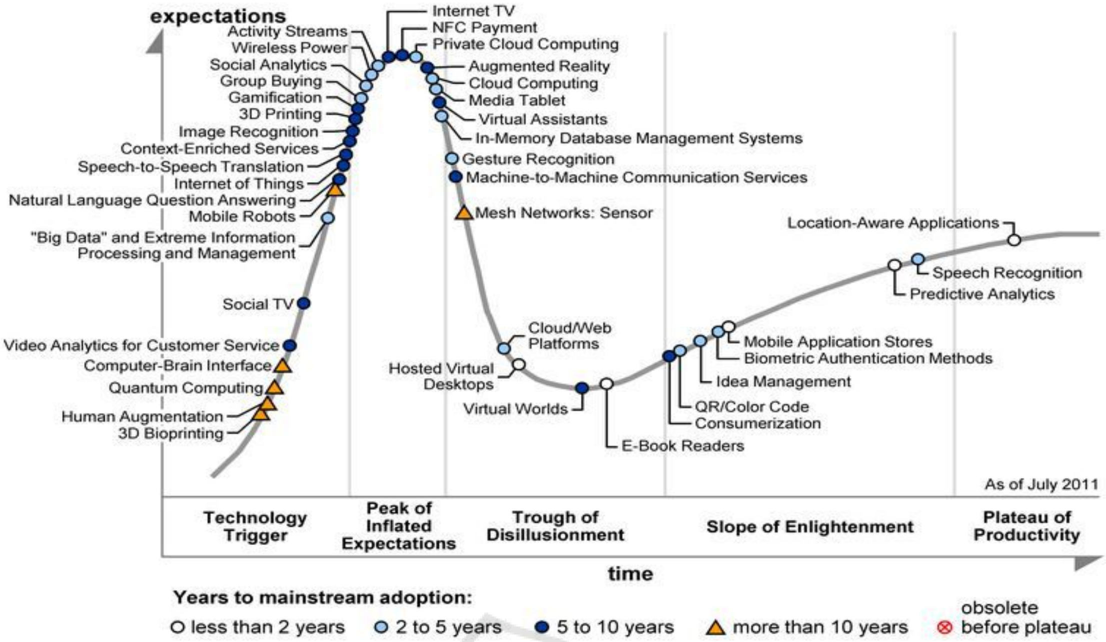

Figure 1: The Gartner hype cycle 2011 (Gartner 2011).

One of the main reasons lies in the tremendous effort put into researching and applying this topic in academia and industry, especially in the first years of its incorporation into the hype cycle. According to Betsy Burton, a business analyst of Gartner, the term big data “has become prevalent in our lives across many hype cycles” (Woodie 2015). This statement indicates the importance and interdisciplinary nature that big data quickly achieved. Compared to other technologies, it has become rapidly indispensable for today's organizations.

The definition of the term itself also constitutes this evolution. Since the initial mention (Diebold 2021) in a research article in (Cox and Ellsworth 1997), the definition and application of the term big data tremendously changed. According to Cox and Ellsworth (1997), big data can be distinguished in collections and objects. While collections refer to the data sets acquired and aggregated, the objects solely focus on single data elements. In both cases, these can sometimes be too big under the given context. Although some minor evidence was put on other relevant data characteristics, these are only implicitly mentioned, such as highlighting the relevance of meta-data that sometimes comes with an extensive collection or the origin out of heterogeneous databases. This trend slowly continued until the subsequent mention by Roger Mouglas from O’Reily (Mishra et al. 2021), who emphasized that big data “refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools” (Dontha 2017). Concurrently to that, the first release of Apache Hadoop, which is today known as one of the core technologies in the domain of big data, emerged. Other approaches that arose afterward didn’t exclusively consider the volume of the data (Chen et al. 2014). Starting from that, other data characteristics were recognized (Al-Mekhlal and Ali Khwaja 2019; Grandhi and Wibowo 2018), and sometimes additional specificities are highlighted, such as the need for scalable technologies (Chang and Grady 2019) or the challenges with established technologies (Manyika et al. 2011). For that reason, it is not surprising that many different research articles almost exclusively deal with the exploration of an applicable definition, such as (Ward and Barker 2013; Ylijoki and Porras 2016). Some of those definition approaches, over the course of the years, are depicted in Table 1.

Those definitions showcase the maturation of the term. However, at the same time, it becomes evident that many discrepancies still exist, and the domain is undergoing continuous transformations. Although the data characteristics, or the nature of the data regarding different dimensions, are still the central focus today, it is above all the technologies, techniques, and general paradigms that make big data appear remarkable. Therefore, it serves today as a technical foundation for many different data-intensive application scenarios requiring sophisticated technological concepts. The specificities, which are often differently discussed, are presented in the following.

Table 1: Selected definitions of the term big data.

Reference	Year	Definition
(Cox and Ellsworth 1997)	1997	„Big data objects are just that -- single data objects (or sets) that are too large to be processed by standard algorithms and software on the hardware one has available.”
(Manyika et al. 2011)	2011	„Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.“
(BITKOM 2012)	2012	„Big Data refers to the analysis of large amounts of data from diverse sources at high speed with the aim of generating economic benefits“
(Chen et al. 2012)	2012	Big data summarizes technological developments in the area of data storage and data processing that provide the possibility to handle exponential increases in data volume presented in any type of format in steadily decreasing periods of time”
(Mayer-Schönberger and Cukier 2013)	2013	“Big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more”
(Hashem et al. 2015)	2015	„Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale“
(Chang and Grady 2019)	2015[1]	„Big Data consists of extensive datasets primarily in the characteristics of volume, velocity, variety, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis“
(Grandhi and Wibowo 2018)	2018	“Big data is defined as a large collection of multifaceted data sets, which can also be described as being high volume, variety and velocity, making difficult to move and process instantly with the traditional database management systems”
(Gartner 2022a)	2022	“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.”

3 DATA CHARACTERISTICS

The nature of the data plays a decisive role in this big data domain and is often referred to by the term data characteristics. As already highlighted by a multitude of different definitions (cf. Table 1) and contributions, as thoroughly investigated in (Volk et al. 2016), the quantity (volume), variety (variety), and speed (velocity) of the data are the most widely acknowledged ones. Notably, all of those start with the letter V, which is an implicitly followed property of the data characteristics formulation in the domain of big data. Presumably, this can be traced back to their origin. The main characteristics, which are commonly abbreviated as the 3V´s go back to the former Gartner (formerly META) data analyst Doug Laney and his report from 2001, titled 3D Data Management: Controlling Data Volume, Velocity, and Variety (Laney 2001). In this report, he discussed the potentials and challenges associated with considering these three dimensions. In the following years, these characteristics were widely applied and further developed in the context of a continuous data increase. To this day, the 3V´s are probably the most critical differentiators when it comes to the consideration of a data-intensive endeavor. However, even though a broad acceptance was achieved by many researchers and practitioners, sometimes different descriptions of each of them can be found. Generally speaking, as highlighted by the given definitions before, this circumstance is notable for almost all specificities of the domain of big data.

As a result of the ongoing global digitalization, for the year 2025 a volume of 175 Zettabytes (ZB) of data is expected to be generated worldwide, compared to 33 Zettabytes in 2018 (Reinsel et al. 2018).

To give perspective: One zettabyte is equal to one trillion gigabytes. If you would store the total amount of 175 ZB on DVD blanks, you would be able to create 23 stacks, each of which would reach from the earth to the moon (Reinsel et al. 2018).

Volume, as the most prevailing data characteristic, stands eponymous for the amount of data that must be acquired, stored, and processed. Within the literature, as most of the other characteristics too, it is considered heterogeneously and refers to the number of data elements and their sheer size (Al-Mekhlal and Ali Khwaja 2019; Chang and Grady 2019; Demchenko et al. 2013). However, at what point in time one can speak of large volumes of data has not yet been clearly defined (Assunção et al. 2015). Individual studies repeatedly deal with different definitions and metrics for the analysis and assessment of the volume, as found out in previous research (Volk et al. 2016). While many of them attempt to provide generic definitions, such as *too huge to handle with established databases*, others indicate data in terabytes, like (Arockia et al. 2017). However, those often don't follow any processes that provide enough evidence or a clear argumentation.

Variety, as another core characteristic, refers to the diversity of the data in terms of structure. Usually it can be distinguished into structured, semi-structured and unstructured data (Erl 2016; Gani et al. 2016). The first type describes data sets with a fixed scheme and can be stored, managed, and analyzed without great effort. A common example is the relational data model, in which information is stored row-based in a tabular format (Erl 2016). The semi-structured data partially contain information about the underlying structure. Still, they are further modifiable and extendable, for example, when using the exchange format of the Extensible Markup Language (XML) (Erl 2016). Contrary to the types above, unstructured data can mostly be handled by special procedures only (Gandomi and Haider 2015). This circumstance is mainly due to the fact that, despite an alleged similarity, great differences in the ability to process the data may occur, for example, when different file formats are present. Image files are a good example of this. While common formats, such as Joint Photographic Experts Group (JPEG), Portable Network Graphics (PNG), or bitmap (BMP), can be used to display the images, vector graphics, or program-specific formats with additional meta-information are also capable. Many authors understand diversity not only in terms of pure structure. For instance, in (Chang and Grady 2019), the incorporation from various sources is meant. Further factors such as the content of the data, the underlying context, the used language, units, or formatting rules are also addressed. In the case of the combination of differently structured data, e.g., through the merging of internal and external holdings, poly-structured data is also sometimes referred to (BITKOM 2014).

Velocity, as the last of the three big data core characteristics, refers to the speed of the data. As in the case of the previous characteristics, there is no universally accepted definition. While many experts in this field address the speed of the pure data processing (Arockia et al. 2017; Khan et al. 2019), others also refer to the speed with which the data arrives (Erl 2016; Gandomi and Haider 2015). Notwithstanding that, velocity is usually expressed in batch, (near) real-time processing as well as streaming. While batch processing is a sequential and complete processing of a certain amount of data (BITKOM 2014; Ghavami 2021), (near) real-time processing is described as an almost continuous processing method. Especially when the speed requirement increases, the processing becomes a non-trivial task. Hence, various technologies, frameworks, and architectures have been developed to deal with this challenge. A graphical overview containing the referred core data characteristics that build the foundation of big data as well as some of the arguably most common supplementary ones, is depicted in Figure 2.

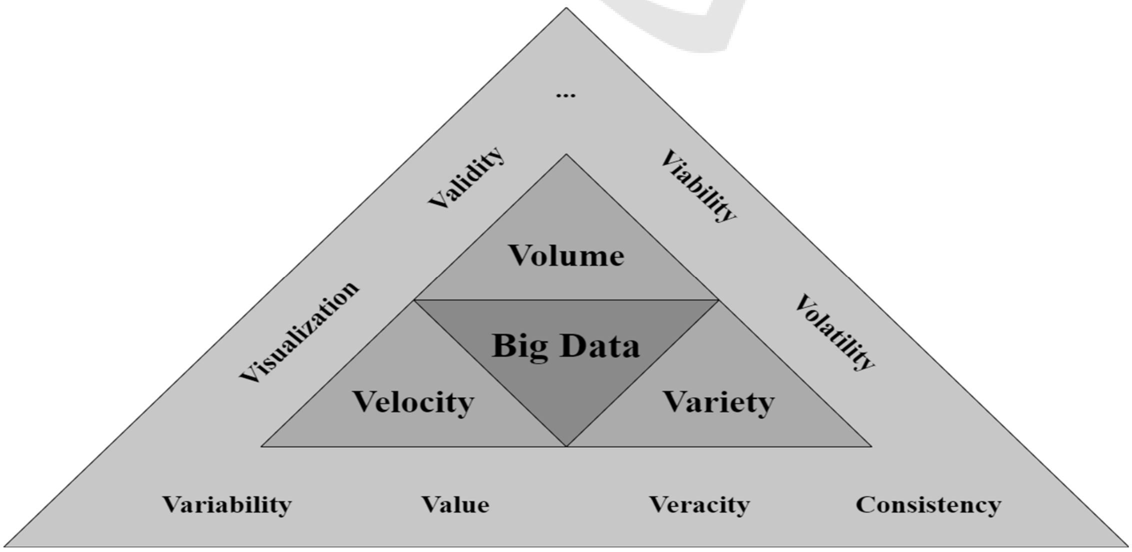

Figure 2: Overview of existing characteristics in the domain of big data, based on (Volk et al. 2020).

Table 2: Data characteristic definitions based on (Volk et al. 2019).

Characteristic	Definition
Volume	Volume indicates the amount of data that has to be handled.
Variety	Variety refers to the heterogeneity of data and its sources.
Velocity	Velocity denominates the speed at which data are incoming and the speed at which received data must be processed.
Volatility	Volatility refers to the tendency for data structures to change over time.
Variability	Variability corresponds to the change of the other characteristics.
Veracity	Veracity reflects the reliability and trustworthiness of the data.
Value	Value refers to economic value that emerges out of the processing of the data.
Consistency	Consistency refers to the data that flows among various sources and is shared by multiple users. In doing so the data needs to be in a consistent status all time.

Noteworthy, apart from these characteristics, many more have emerged since the origin of the term big data and the first mentioning of the data characteristics. Some of these have become widely accepted and used by researchers and practitioners today.

Volatility, as another characteristic, “refers to the tendency for data structures to change over time” (Chang and Grady 2019). While this statement may lead to the general idea of a data structure change, as described by the variety, it rather focuses on the overall changes of the data over time and its useability. The analysis of the data that was collected, stored, and processed over time may experience a certain shift regarding its level of detail or use (Chang and Grady 2019; Kapil et al. 2016). Hence, a holistic consideration of novel and historical data is required.

Variability, as an additional characteristic, refers to fluctuations that may occur in the context of the processed data. While some authors, such as (Hussein 2020), focus on the changes of the data itself, others understand by it the variations of the three core characteristics either in an isolated or compound way (Chang and Grady 2019; Gandomi and Haider 2015; Katal et al. 2013). Because of that, an impact on the overall data processing may occur that needs to be counteracted. This includes the data flow rate, structure, and volume. In doing so, sophisticated approaches are required to overcome emerging problems and increase the system's flexibility. One of them is the scaling of related systems, horizontally and vertically, for instance, by harnessing the capabilities of cloud computing (Chang and Grady 2019).

Veracity provides information about the data quality and thus the reliability of the raw data. It was first coined by IBM report in 2013 (IBM 2013). Especially when it comes to the necessity to utilize unreliable data for big data projects, both the correct context and the used analysis method are of utmost importance (Gandomi and Haider 2015). In the case of unreliable data, one measure could be to merge several sources in order to increase the overall quality, and thus the trustworthiness (IBM 2013; Izadi et al. 2015).

With the value of data, Oracle (Dijcks 2013) introduced a meta-data characteristic, rather than a stand-alone data characteristic, as it is heavily influenced by other characteristics, such as veracity, velocity, volume, and variety (Erl 2016). It focuses on the economic value of the data to be processed. It is possible to extract information and gain previously hidden knowledge, especially with semi- and unstructured data, which differ from traditional data structures. Often, however, the data in its pure form do not contain any significant benefit, which means that preceding processing steps and analyses are necessary to generate such value (Gandomi and Haider 2015).

Apart from those existing V’s, other concepts were introduced in recent years, providing further data characteristics that do not follow this pattern. The 3C model addresses the cost, complexity, and consistency in the context of the data to be processed. While the first focuses on the overall monetary expenditures required for a technological realization, the second indicates the severity of connections and relations between the data. The last C, the consistency, refers to the “data that flows among various sources and is shared by multiple users” (Bedi et al. 2014). Generally speaking, due to the distributed nature of data and related systems in big data, keeping a consistent state can be demanding. In many cases, simultaneous writes and reads have to be carefully observed to avoid adhering problems (Bedi et al. 2014; Chang and Grady 2019).

While the characteristics mentioned above have been widely recognized today, numerous other papers, studies, and reports have attempted to establish additional data characteristics that are less widely used. As a result, reports emerged that present 10V's (Kapil et al. 2016), 17V's (Arockia et al. 2017), 42 V's (Schafer 2017), 51V's (Khan et al. 2019) or even 56V's (Hussein 2020).

However, as many authors already highlight by themselves, not all of them are always important or even applicable. An overview of the arguably most essential definitions is given in Table 2.

4 CONCLUSION

Big data has become one of the most important technological trends of our time. With its tremendous influence on almost all aspects of today's society, the concept has sparked the interest of scientists and practitioners alike. This not only led to numerous publications dealing with the topic, but also to a variety of definitions and explanations of the term itself as well as the associated characteristics. However, to facilitate a fruitful discourse, it is necessary to have a common understanding of the corresponding terminology. For this reason, the publication at hand discusses the term, its origins as well as the most relevant data characteristics that can be found in the literature. Furthermore, some exemplary works that extend the set of data characteristics beyond the most common ones are highlighted.

Hereby, subsequent researchers as well as practitioners interested in the domain are provided with a strong base that further considerations and future projects can be built upon. This, in turn, hopefully helps to facilitate the progression of the domain as a whole as well as its further proliferation.

REFERENCES

Al-Mekhlal, M., and Ali Khwaja, A. (2019). "A Synthesis of Big Data Definition and Characteristics," in *Proceedings of the 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),* New York, NY, USA. 01.08.2019 - 03.08.2019, IEEE (doi: 10.1109/cse/euc.2019.00067).

Arockia, P. S., Varnekha, S. S., and Veneshia, K. A. (2017). "The 17 V's Of Big Data," *International Research Journal of Engineering and Technology* (4:9), pp. 329-333.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., and Buyya, R. (2015). "Big Data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing* (79-80), pp. 3-15 (doi: 10.1016/j.jpdc.2014.08.003).

Bedi, P., Jindal, V., and Gautam, A. (2014). "Beginning with big data simplified," in *Proceedings of the 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC),* Delhi, India. 05.09.2014 - 06.09.2014, IEEE, pp. 1-7 (doi: 10.1109/ICDMIC.2014.6954229).

BITKOM. (2012). "Big Data im Praxiseinsatz - Szenarien, Beispiele, Effekte," BITKOM (ed.), Berlin.

BITKOM. (2014). "Big-Data-Technologien–Wissen für Entscheider," BITKOM (ed.), Berlin.

Chang, W. L., and Grady, N. (2019). "NIST Big Data Interoperability Framework: Volume 1, Definitions," NIST - National Institute of Standards and Technology.

Chen, H., Chiang, R. H., and Storey, V. C. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quaterly* (36:4), p. 1165 (doi: 10.2307/41703503).

Chen, M., Mao, S., and Liu, Y. (2014). "Big Data: A Survey," *Mobile Networks and Applications* (19:2), pp. 171-209 (doi: 10.1007/s11036-013-0489-0).

Cox, M., and Ellsworth, D. (1997). "Managing big data for scientific visualization," *ACM Siggraph* (97:1), pp. 21-38.

Demchenko, Y., Grosso, P., Laat, C. de, and Membrey, P. (2013). "Addressing big data issues in Scientific Data Infrastructure," in *2013 International Conference on Collaboration Technologies and Systems (CTS),* San Diego, CA, USA. 20.05.2013 - 24.05.2013, IEEE, pp. 48-55 (doi: 10.1109/CTS.2013.6567203).

Diebold, F. X. (2021). "What's the big idea? "Big Data" and its origins," *Significance* (18:1), pp. 36-37 (doi: 10.1111/1740-9713.01490).

Dijcks, J.-P. (2013). "Oracle: Big Data for the Enterprise," Oracle Corporation (ed.), Redwood.

Dontha, R. (2017). "Who came up with the name Big Data?" *TechTarget*.

Erl, T. (2016). *Big data fundamentals*: *Concepts, drivers & techniques*, Boston: Prentice Hall ServiceTech Press.

Gandomi, A., and Haider, M. (2015). "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management* (35:2), pp. 137-144 (doi: 10.1016/j.ijinfomgt.2014.10.007).

Gani, A., Siddiqa, A., Shamshirband, S., and Hanum, F. (2016). "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and Information Systems* (46:2), pp. 241-284 (doi: 10.1007/s10115-015-0830-y).

Gartner. (2011). "Hype Cycle for Emerging Technologies, 2011," available at https://www.gartner.com/en/docum ents/1754719/hype-cycle-for-emerging-technologies-2011, accessed on Nov 25 2020.

Gartner. (2014). "Hype Cycle for Emerging Technologies, 2014," available at https://www.gartner.com/en/documents/2809728, accessed on Jun 14 2022.

Gartner. (2015). "Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor," available at https://www.gartner.com/en/newsroom/press-releases/2015-08-18-gartners-2015-hype-cycle-for-emerging-technologies-identifies-the-computing-innovations-that-organizations-should-monitor, accessed on Nov 25 2020.

Gartner. (2022a). "Definition of Big Data - Gartner Information Technology Glossary," available at https://www.gartner.com/en/information-technology/glossary/big-data, accessed on Mar 5 2022.

Gartner. (2022b). "Gartner Hype Cycle Research Methodology," available at https://www.gartner.com/en/research/methodologies/gartner-hype-cycle, accessed on Mar 15 2022.

Ghavami, P. (2021). *Big data management*: *Data governance principles for big data analytics*, Berlin: De Gruyter.

Grandhi, S., and Wibowo, S. (2018). "A Multi-criteria Group Decision Making Method for Selecting Big Data Visualization Tools," *Journal of Telecommunication, Electronic and Computer Engineering* (10:1-8), pp. 67-72.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems* (47), pp. 98-115 (doi: 10.1016/j.is.2014.07.006).

Hussein, A. A. (2020). "Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2)," *Journal of Information Security* (11:04), pp. 304-328 (doi: 10.4236/jis.2020.114019).

IBM. (2013). "Analytics: The real-world use of big data,"

Izadi, D., Abawajy, J. H., Ghanavati, S., and Herawan, T. (2015). "A data fusion method in wireless sensor networks," *Sensors (Basel, Switzerland)* (15:2), pp. 2964-2979 (doi: 10.3390/s150202964).

Kapil, G., Agrawal, A., and Khan, R. A. (2016). "A study of big data characteristics," in *Proceedings of the 2016 International Conference on Communication and Electronics Systems (ICCES),* Coimbatore, India. 21.10.2016 - 22.10.2016, pp. 1-4 (doi: 10.1109/CESYS.2016.7889917).

Katal, A., Wazid, M., and Goudar, R. H. (2013). "Big data: Issues, challenges, tools and Good practices," in *Sixth International Conference on Contemporary Computing*, Parashar (ed.), Noida, India. 08.08.2013 - 10.08.2013, IEEE, pp. 404-409 (doi: 10.1109/IC3.2013.6612229).

Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., and Qamar, S. (2019). "The 51 V's Of Big Data," in *Proceedings of the 2019 International Conference on Omni-Layer Intelligent Systems,* Crete Greece. 05.05.2019 - 07.05.2019, New York,NY,United States: Association for Computing Machinery, pp. 19-24 (doi: 10.1145/3312614.3312623).

Laney, D. (2001). "3D data management: Controlling data volume, velocity and variety," *META group research note* (6:70).

Manyika, J., Chui, M., Brown, B., Bughin Jacques, Dobbs, R., Roxburgh, C., Byers, and Angela Hung. (2011). "Big Data: The next frontier for innovation, competition, and productivity," McKinsey (ed.), McKinsey.

Mayer-Schönberger, V., and Cukier, K. (2013). *Big data*: *A revolution that will transform how we live, work, and think*, Boston: Houghton Mifflin Harcourt.

Mishra, B. K., Kumar, V., Panda, S. K., and Tiwari, P. (2021). *Handbook of Research for Big Data*: *Concepts and Techniques*, Milton: Apple Academic Press Incorporated.

Müller, O., Fay, M., and Vom Brocke, J. (2018). "The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics," *Journal of Management Information Systems* (35:2), pp. 488-509 (doi: 10.1080/07421222.2018.1451955).

Reinsel, D., Gantz, J., and Rydning, J. (2018). "The Digitization of the World - From Edge to Core," available at https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, accessed on Mar 5 2020.

Schafer, T. (2017). "The 42 V's of Big Data and Data Science," available at https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html, accessed on Jun 12 2022.

Staegemann, D., Volk, M., Nahhas, A., Abdallah, M., and Turowski, K. (2019). "Exploring the Specificities and Challenges of Testing Big Data Systems," in *Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems,* Sorrento. 26.11.2019 - 29.11.2019.

Volk, M., Hart, S. W., Bosse, S., and Turowski, K. (2016). "How much is Big Data? A Classification Framework for IT Projects and Technologies," in *Proceedings of the 22nd Americas Conference on Information Systems, AMCIS 2016,* San Diego, CA, USA. 11.08.2016 - 14.08.2016, AIS.

Volk, M., Staegemann, D., Pohl, M., and Turowski, K. (2019). "Challenging Big Data Engineering: Positioning of Current and Future Development," in *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security,* Heraklion, Crete, Greece. 02.05.2019 - 04.05.2019, SCITEPRESS - Science and Technology Publications, pp. 351-358 (doi: 10.5220/0007748803510358).

Volk, M., Staegemann, D., and Turowski, K. (2020). "Big Data," in *Handbuch Digitale Wirtschaft*, T. Kollmann (ed.), Wiesbaden: Springer Fachmedien Wiesbaden, pp. 1-18 (doi: 10.1007/978-3-658-17345-6_71-1).

Ward, J. S., and Barker, A. (2013). "Undefined by data: a survey of big data definitions," *arXiv preprint arXiv:1309.5821*.

Woodie, A. (2015). "Why Gartner Dropped Big Data Off the Hype Curve," available at https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/, accessed on Nov 25 2020.

Ylijoki, O., and Porras, J. (2016). "Perspectives to Definition of Big Data: A Mapping Study and Discussion," *Journal of Innovation Management* (4:1), pp. 69-91 (doi: 10.24840/2183-0606_004.001_0006).