On Definition of BigData

Oleg Zolotov[1,*], *Yulia* Romanovskaya[2], and *Varvara* Rzhannikova[2]

[1]Murmansk Arctic State University, Near-Earth Environment Computer Modelling Laboratory, RU-183025, Murmansk, Russia
[2]Murmansk State Technical University, Department of Math, Information Systems and Software Engineering, RU-183010, Murmansk, Russia

Abstract. The term Big Data (or BigData) is widely used in scientific, educational, and business literature; however, there does not exist a single definition that can be unreservedly called “canonical”. A careless use of Big Data term to promote commercial software further emphasizes the importance of this issue. In this paper, we have performed a review of definitions of Big Data and highlighted the principal features that are attributed to Big Data. We compared all these principal features with features of databases compiled using Edgar F. Codd’s publications, and showed that they are not unique and can also be attributed to the databases. Having studied C. Lynch original work, we proposed the definition of Big Data based on the so-called conservation institution. The key point of this definition is a shift from purely technical attitude towards public institutions. Since the current use of the Big Data term may lead to a loss of meaning. There is a need not only to spread out best practices but also to eliminate or minimize the use of dubious or misleading ones.

1 Introduction

Specific study of a given phenomenon requires determination of a common terms dictionary that ensures consistent communications and understanding of the object being investigated. The Big Data term is widely used in relation to scientific, educational and business tasks but there is no single specific definition that can be unreservedly called as “canonical” Big Data definition. The use of Big Data term to promote commercial software intelligence solutions further exaggerates the situation.

Clifford Lynch is considered the person who firstly introduced the term Big Data [1]. Curiously, his paper does not provide explicit definition of the Big Data. Instead, it discusses the challenges that appear due to a significant increase of the data volumes and considers new solutions that allow to obtain, transform, store, and analyze those huge datasets. As the key solution C. Lynch formulated a foundation of what he called “preservation institutions”.

Attempts to generate added value from the data, to produce new knowledge and methods to deal with the data, are, in particular, reflected in the development of information theory as well as the database theory. For example, Edgar F. Codd in 1970 published the article [2] entitled “A relational model of data for large shared data banks”. Keeping in mind that large is larger than big and huge is bigger than large, a few mostly rhetorical questions might arise. Do the large data banks from 1970th refer to bigger things in contrast with nowadays’ Big Data? Should we expect the appearance of Huge Data in the nearest future?

This paper presents our attempt to formalize principle features of the data that make the Data Big by the nature.

2 Methods

In this investigation we perform a review of recently used Big Data definitions and use-cases, and contrast them with each other to discriminate commonly accepted features. A few Big Data definitions are summarized in Table 1. If the definition is given not in English, a translation to English is provided. After that the discriminated Big Data features are discussed.

3 Results

A few typical examples of modern Big Data definitions one may see in Table 1. The following features are usually declared to make the data Big:

- large volumes of the data;
- the required large-scale computer power;
- the lack of the structure of the data;
- the need for specialized hardware, software, and algorithms to deal with,
- the requirement of innovations in hardware, software, algorithms and means to perform analysis,
- the requirement to get the result in reasonable time,
- the requirement to unlock the business value.

[*] e-mail: ZolotovO@gmail.com

The mentioned above is usually provided in relative units, i.e. in contrast with currently available solutions.

Table 1. Big Data definitions

Definition of Big Data	**Ref.**
"The term BigData is used to describe massive digital datasets that require innovations in analytical techniques in order to exploit them and create new forms of value. Big data's vastness is not about absolute size but about the required scale of analysis"	Ref.[3]
Big Data is often understood as: large volumes of data arrays and the need to use large-scale computing power, custom software and methods for extracting value from data in a reasonable amount of time. *(Translated by V. Rzhannikova, see original Russian text in [4])*	Ref.[4]
Big data is a term that defines not only the size of data sets that exceeds the capabilities of conventional databases, but also unstructured information, which can't be process and analyze by traditional algorithms. *(Translated by V. Rzhannikova, see original Russian text in [5])*	Ref.[5]
The term Big Data refers to data sets whose size exceeds the capabilities of typical databases for storing, managing and analyzing information. *(Translated by V. Rzhannikova, see original Russian text in [6])*	Ref.[6]
"Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value"	Ref.[7]
"Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results"	Ref.[8]
"Big Data is a data that's too big, too fast, or too hard for existing tools to process"	Ref.[9]
"Datasets which could not be captured, managed, and processed by general computers within an acceptable scope."	Ref.[10]
"Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources"	Ref.[11]

4 Discussion

As it follows from Table 1, the most often mentioned Big Data feature is the size of the data that we are able to process on a "regular" computer. This criterion is not very stable. Dating back to 1980th, a typical random access memory (RAM) capacity increased from units of kB to GB nowadays, i.e. 10^6 times. Similarly, a persistent storage (like hard disk drives, tapes, etc…) capacity increased for more than 9 orders – from kB to TB, or even more if consider special devices or cloud storage solutions. Thus, classification of the data as Big in this case depends on the currently available hardware, and it will be eventually changed.

Another widely used criterion is the required computational power (CPU- or machine-power). It does not make any significant changes to the mentioned above because the performance of the computing machines increased greatly till nowadays. For example, the CPU clock frequencies raised from MHz to GHz, i.e. 103 times or 3 orders increase.

The requirement for a specific hardware and / or software to process Big Data is also not unique. For example, Edgar F. Codd considered specific problems of multiprogramming scheduling [12-14].

Definition of the term Big Data basing on the only structure of the considered data is also incomplete. Problem of complexly structured, unstructured, or semi-structured data representation is the well-known topic in the frame of, e.g., relation databases [2, 15-16] from the time of their appearance.

The requirement to get the result in reasonable time does not introduce any new features and seems to be like an attempt to define technical (software and hardware) characteristics without their explicit formulation. Such criterion may be considered as an attempt to move from low-level technical domain into the constrained with business requirements domain.

The requirement for innovations to deal with data is also not innovative. For example, in the frame of relational databases a relational algebra was developed to describe and investigate the properties of relations and corresponding operations. Moreover, techniques [2, 17-19] were developed that served as a guide for application of newly developed relational databases and relational theory for practical business-purposes. Besides, a few most popular for-that-time-innovative products were evaluated against the requirements for databases to be relational one [20].

The requirement to unlock the business value is not specific to the Big Data neither original. In a slightly old-fashioned manner the same problem was discussed by Edgar F. Codd [21] in term of "productivity" in his 1981 ACM Turing Award lecture entitled "Relational Database: A Practical Foundation for Productivity". A generation of value is sometimes explained in terms of new knowledge extraction that triggers, e.g., new use-cases and user experience or significantly change the way the user interact with. Similar topics were also covered by Edgar F. Codd in relation to the databases. For example, an ability to use natural languages as a database query language was considered in [22]. Specific problems related to the "semantic models" representation and extraction were covered in [23]. A related but different problem to describe, denote and manipulate the (representation of) missing information was discussed in [24-25] on the basis of three-valued logic.

An attempt to classify the data as big basing on its origin (nature) is also not reliable. Data from any research field (including but not limited to chemistry, physics, computer vision investigations like, e.g., in [26-28]) might produce huge and small pieces of data depending on the considered spatial scale and time-step.

Below we discuss a few concrete Big Data definitions. Considering Big Data as data "that's too big, too fast, or too hard for existing tools to process" [9] is likely to be a motto but it is not a robust definition. It might reveal its' place to attract attention, to promote the technology, or to involve a community around an ecosystem. Being understood literally, it implies that Big Data can't be processed at all. Such assumption does not seem reasonable.

Another definition [10] stating that Big Data could not be "… processed by general computers…" requires further explanation. For example, the term "general computers" might refer to general-purpose computers as well as to a typical "averaged" computer in use for a given architecture or a use-case.

As one may see, a common problem to define the Big Data term arises from an attempt to build a reference frame basing on relative conditions. The relative nature of Big Data is explicitly noted, e.g., in [3]. It means, any Big Data definition that directly or indirectly refers to the currently available hardware and software abilities will eventually become outdated.

Keeping in mind the original paper [1], we define Big Data as the data that requires the "preservation institutions" to deal with. It is important that preservation institutions are not about hardware or software, i.e., technical requirements (only). They are strongly linked to the organizational and legal issues as well as with authorities and public society communications. An example of such a mature institution with a long history is the libraries. The challenge is to guarantee operations for periods exceeding human life or even the time of some countries existence. Such challenge can't be addressed by an individual or a not-specialized organization. There is also well-known Internet-related example, i.e., Web Archive (https://web.archive.org/). By the way, the Web Archive – Internet Archive is officially registered as the library.

Preservation institutions in contrast with libraries in addition should provide guarantees on specific means to deal with the data (to generate added value or knowledge). The value of such operation may be demonstrated by the following case. Users of the PyGlow (https://github.com/timduly4/pyglow) geophysical package were surprised when NGDC (National Geophysical Data Center) NOAA (National Ocean and Atmosphere Administration) discontinued to provide (update) a few geophysical indexes. PyGlow is designed to provide python wrappers to a set of well-known geophysical models (IRI – International Reference Ionosphere, HWM – Horizontal Wind Model, etc…) and by design depended on those indexes to perform models' runs.

In this paper we considered Big Data term definition from scientific publications only. Definitions provided with commercial products by corporations (like Amazon, Google, IBM, Microsoft, Yandex, etc) are out of the scope of this article. The revealed from the available publications principle Big Data features we evaluated against their ability to be unique on specific to Big Data. Basing on the E.F. Codd's publications only [2, 12-25] we clearly demonstrated that all the "specific" to the Big Data features were considered long before, e.g., in the frame of database theory and database management systems' implementation. We intentionally analyzed decades-aging publication to demonstrate that the "specific" to Big Data problems appeared long before the Big Data term appearance. Despite the relative youth of the computer science, it is possible to illustrate similar problems with even earlier publications. But we could not find a researcher (except Edgar F. Codd) who touched all those problem jointly and consistently.

5 Conclusion

In this paper we present a review of typical Big Data definitions. They all rely on the following features or a combination of them. (1) A volume of the data. (2) Technical characteristics of the required hardware and software. (3) A few 'natural' or business-like characteristics similar to time-to-process or time-to-deliver. (4) The above characteristics are usually nominated in relative units win contrast with currently available data-processing means. Some researchers denote the relative nature of the Big Data explicitly.

We demonstrated that it is impossible to define the Big Data term in absolute units because almost any data being classified as the Big one at a given moment of time will eventually become un-Big due to the hardware and software facilities improvements.

Basing on the initial paper by C. Lynch [1], we proposed the definition of Big Data as the data that requires the preservation institutions to deal with, i.e. to generate added value or to extract a new knowledge. This definition also has a kind of relative nature but principally shifts the key features from purely technical or business domains into the institutional domain.

Big Data term is also used as an "umbrella" term which hides different brunches of IT-technologies. Moreover, it is often impossible to recognize the specific technology hidden behind. This usage of the term Big Data supports the "hype" around corresponding technologies and results in significant development in the educational and business applications. But it also makes harder to build a common terms' dictionary and spoils the understanding of the object being considered.

Thus, like Edsger W. Dijkstra's letter "Go To Statement Considered Harmful" [29] triggered the revolution in software development aimed to abolish the harmful practices, today there is the strong need to eliminate harmful practices in the field of the Big Data.

References

1. C. Lynch, Nature **455**, 28-29 (2008) doi: 10.1038/455028a
2. E.F. Codd, Commun. ACM **13**(6), 377-387 (1970), doi: 10.1145/362384.362685
3. C. Pentzold, C. Brantner, L. Fölsche **21**(1), 139-167 (2019), doi: 10.1177/1461444818791326
4. K.S. Juchinson, J. Higher School Econ. **1**, 216-245 (2017), doi: 10.17323/2072-8166.2017.1.216.245 [in Russian]
5. O.Y. Denisova, A.E. Mukhutdinov, Bull. Kazan Tech. Univ. **18**(4), 226-230 (2015), URL: https://cyberleninka.ru/article/n/bolshie-dannye-eto-ne-tolko-razmer-dannyh [in Russian]
6. P.D. Ivanov, V.Zh. Vampilova, Engineering J.: Sci. Innov. **8**(32), (2014), doi: 10.18698/2308-6033-2014-8-1228 [in Russian]
7. S.R. Qureshit, A. Gupta, CSIBIG, 1-6 (2014), doi: 10.1109/CSIBIG.2014.7056933
8. S. Sagiroglu, D. Sinanc, In : Int. Conf. Collab. Tech. Syst., 42-47 (2013), doi: 10.1109/CTS.2013.6567202
9. S. Madden, IEEE Internet Comp. **16**(3), 4-6 (2012), doi: 10.1109/MIC.2012.50
10. M. Chen, S. Mao, Y. Liu, Mobile Netw. Appl. **19**, 171-209 (2014), doi: 10.1007/s11036-013-0489-0
11. X. Wu, X. Zhu, G. Wu, W. Ding, IEEE Trans. Knowledge Data Engineering **26**(1), 97-107 (2014), doi: 10.1109/TKDE.2013.109
12. E.F. Codd, E.S. Lowry, E. McDonough, C.A. Scalzi, Commun. ACM **2**(11), 13-17 (1959), doi: 10.1145/368481.368502
13. E.F. Codd, Commun. ACM **3**(6), 347-350 (1960), doi:10.1145/367297.367317
14. E.F. Codd, Commun. ACM **3**(7), 413-418 (1960), doi: 10.1145/367349.367356
15. E.F. Codd, Commun. ACM **6**(3), 40-42 (1974), doi: 10.1145/983076.983079
16. E.F. Codd, IEEE Softw. **5**(4), 4-6 (1988)
17. E.F. Codd, C. J. Date, In: Proc. of the 1974 ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control: Data Models: Data-structure-set Versus Relational, 11-41 (ACM 1975), doi: 10.1145/800297.811529
18. C.J. Date, E.F. Codd, In: Proc. ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control: Data Models: Data-structure-set Versus Relational, 83-113 (ACM 1975), doi: 10.1145/800297.811532
19. E.F. Codd, SIGPLAN Not. **16**(1), 112-114 (1980), doi: 10.1145/960124.806891
20. E.F. Codd, In: Proc. 2nd Int. Conf. Data Engineering**,** 720-729 (IEEE Computer Society 1986)
21. E.F. Codd, Commun. ACM **25**(2), 109-117 (1982), doi: 10.1145/358396.358400
22. E.F. Codd, SIGART Bull. **61**, 31-32 (1977), doi: 10.1145/1045283.1045298
23. E.F. Codd, In: *Proc. ACM SIGMOD*, 161 (ACM 1979), doi: 10.1145/582095.582122
24. E.F. Codd, SIGMOD Rec. **15**(4), 53 (1986b), doi: 10.1145/16301.16303
25. E.F. Codd, SIGMOD Rec. **16**(1), 42-50 (1987), doi: 10.1145/24820.24823
26. J. Herb, A.B. Nadykto, K.M. Nazarenko, N.A. Korobov, F. Yu, Comp. Theor. Chem. **1133,** 40-46 (2018), doi: 10.1016/j.comptc.2018.04.012
27. O. V. Zolotov, M. A. Knyazeva, Yu. V. Romanovskaya, Russian J. Phys. Chem. B **13**, 681-684 (2019), doi: 10.1134/S1990793119040146
28. V. Voronin, M. Pismenskova, A. Zelensky, Y. Cen, A. Nadykto, K. Egiazarian, In: Proc. SPIE Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies II **10802**, 204-211 (SPIE 2018), doi: 10.1117/12.2326801
29. E.W. Dijkstra, Commun. ACM **11**(3), 147-148 (1968), doi: 10.1145/362929.362947