Challenges According to the ‘7Vs’ of Big Data

Subjects: Computer Science, Theory & Methods | Computer Science, Information Systems

Contributor: Cristian González García , Eva Álvarez-Fernández

Big Data has challenges that sometimes are the same as its own characteristics. These characteristics are known as the Vs. The researchers describe ‘7Vs’ that they have found and are the most used and more related to general data used in Big Data, including an explanation and challenges that exist according to them. Some of them have different subtypes according to the differences detected in the literature and working with them. Some authors create a different V for similar purposes, and here, the researcher gather them into one due to the similarities.

Big Data survey challenges Data Mining KDD Vs

1. Volume

The volume or quantity of data is enormous and gigantic. This feature describes exactly that: the large amount of data that is coming, the large amount that is working daily ranging from gigabytes to exabytes and more. Storage capacity has been doubling approximately every 3 years, yet in many fields this has been insufficient, as has been the case in medical and financial data [1].

Initially, these data came from databases. However, after the ‘boom’ in improved technologies, more sites are now taken into account, such as the Internet, smart objects and sensors and actuators [2], applications, and even photos and videos, or the Internet of Things (IoT) [3], or Online Social Networks. In addition, some researchers should deal with datasets thousands of times larger than those generated only a decade ago, such as satellite procedures, telescopes, high-performance instruments, sensor networks, particle accelerators and supercomputers [4]. This also makes us human beings into walking data generators [5], since we are the ones who continuously generate many of these data with the use of mobile devices, among others.

Many companies currently try thousands of terabytes (10^{12}) daily to obtain useful information for their businesses and about their users. The amount of data is growing because every day there is more information and more existing users, which implies, as some already estimated, that it has exponential growth. It is estimated that there will be 500 times more data in 2020 than in 2011 [6].

Often this knowledge must be efficient as it needs to be real time due to problems with the space to store it [7]. An example of this is the Square Kilometre Array (SKA) radio telescope [8], which will become the largest radio telescope in the world and aims to discover the beginning and end of the universe. Researchers estimate that SKA will produce approximately 4.1 pebibits (2^{50}) per second or 562 terabytes (10^{12}) per second. As can be seen, there

will be a future with even more data, with huge amounts to analyse. Some estimates indicate that the amount of new information will double every three years [9].

Another example of the amount of data generated occurred on 4 October 2012, when the presidential debate between the President of the United States of America, Barack Obama, and Governor Mitt Romney had Twitter users post more than ten million tweets in an hour and a half [10]. Studying this Twitter data about the US presidency debate [10], it can be observed that people tweeted more when talking about the health insurance of older people. Thus, based on this, it can be determined what mattered to people at that time.

Flickr, a social network of photos, is also widely used in investigations because of the information that can be obtained from its images. About 70 million new public photos are currently uploaded monthly [11]. Thus, thanks to the analysis of these photos, they could be used to study human society, social events or even disasters [7].

On the other hand, the scientific advisor to the President of the United States of America, John Paul Holdren, said in 2011 that Big Data was an important issue because every year about 1.2 zettabytes (10^{21}) of electronic data are created. The equivalent in terabytes is 1,200,000,000, ranging from scientific experiments to telescope data and tweets [12]. This is certified by other estimates made in 2012 [5], where they predicted the creation of 2.5 exabytes (10^{18}) each day, equivalent to 2,500,000 terabytes, but that this creation capacity would double every 40 months, approximately. Thus, this prediction is quite similar to that made by John Paul Holdren.

However, this amount of data is not currently being created because, years ago, there were already certain applications that generated large amounts of data. In 1990, satellites orbiting Earth generated one terabyte (10^{15}) of information per day. This meant that if a photo were taken every second and a person was asked to analyse all these photos, assuming he worked at night and on weekends, it would take him many years to analyse all the photos of a day [13]. This is useful to emphasise that now, 26 years later and with improved technologies, both hardware and software, we can make faster and automatic analysis, also on images with better resolution and more data.

Another relevant project, both in terms of importance and data size, is the project 'The Genome 1000 Project', which deals with the human genome. In this project, two hours of Solexa execution created 320 terabytes of information. This made it impossible to save and compute in 2008 [14].

In relation to this evolution, it can be seen the prediction of Eron Kelly, who predicted that in the next five years, we will generate more data than all those generated by humanity in the last 5000 years [15]. Meanwhile, the NIST expects data generation to double every two years, reaching 40,000 exabytes by 2020, of which one-third is expected to be useful if analysed. This is something that highlights the evolution of the data humans generate and its exponential growth throughout history, as well as what is expected to happen.

To handle all this information, some years ago the different systems have started to be migrated to the Cloud and perform what is known as Cloud Computing, whether in a private, public or hybrid system. This resulted in a saving

of money and a new way of processing data, which facilitated the use of Big Data in companies thanks to the different technologies provided by companies and the scaling of these tools [16].

However, it should be kept in mind that not always because of having larger datasets will one get better predictions; this can be classified as arrogance, because Big Data is not the substitute for data collection and Data Mining [17]. Therefore, despite having these large sets, one should not forget the basic techniques of measurement and construction of reliable and valid data and the dependencies between these data [18].

2. Velocity: Reading and Processing

Velocity describes how fast data is processed, because in Big Data, the creation of this data is continuous; it never ceases. On the Internet, whether through an Online Social Network such as Twitter, or on Facebook, or by different services such as blogs or video services, there are people at all times writing information, uploading a video, sending emails or accessing web pages. This happens all the time, thousands of datasets every minute. This makes for a great velocity of content creation or reading.

The vast majority of these services require data from their users, how they use their service or their preferences in order to adapt their content or ads to their users. This creates a great need for data processing velocity. This velocity can be of four types: batch or interval, near time or nearly time, which is almost real time, real-time, and streaming.

As can be seen, this "V" has two types of velocities, the one of reading or content creation and the one of processing, which can be independent or dependent, according to the requirements of the application.

In addition, there are applications that need more processing velocity than data volume and thus allow a company to be much more agile than its competitors by offering real-time or near-real-time applications [5][6]. It is important, at this point, to bear in mind that this does not refer to bandwidth or protocol issues, but to the velocity of content creation and the manageability of these, which is divided into their storage, analysis and visualisation, hoping that with the use of Big Data this time will be minimum [6][19].

An example of this "V" is the competition that took place at the SC08 International Conference for High Performance Computing congress in 2008 [20]. Participants had to consult with the Sloan Digital Sky Survey (SDSS). The SDSS is a space research project in which three-dimensional images of space are taken in order to map it [21]. The winner took 12 min to make a query in a parallel cluster, while the same query without using parallelism took 13 days [4]. This highlights the importance of processing velocity when computing large amounts of data.

3. Variety

Variety describes the organisation of data, whether structured, semi-structured, unstructured or mixed. Currently, there are countless possible sources of data given by the wide variety of existing sources for collecting them, being in many cases unstructured, difficult to handle and very noisy data [5]. The researchers are talking about millions of sensors around the world, tens of social networks in which millions of messages are published daily and hundreds of different formats ranging from plain text to images.

There are websites where users write their opinions: Twitter, Facebook, LinkedIn, Instagram, YouTube, Tumblr, Flickr and other social networks, which are considered the most valuable resources [22]. Another very important variety of data is that offered by different governments when they provide different open data, as these data are offered in different formats such as Excel, CSV, Word, PDF, JSON, RDF-N3, RDF-Turtle, RDF-XML, XML, XHTML, plain text, HTML, RSS, KML and GeoRSS. Other types of data that are interesting are videos, in different formats such as FLV, WMV, AVI, MKV or MOV, which are also often accompanied by comments; in this case, the researchers find services such as YouTube and Vimeo, among others. A similar case is that of audio services or radios, which have different formats, such as MP3, OGG, FLAC, MIDI or WAV.

The latest technologies, such as different mobile devices, televisions, cars, tablets, smartphones or any other Smart Object [2], offer many data through the different means they have, either through their sensors or their GPS systems. Mentioning sensors, they are all available in the market and they are opening up more and more thanks to their ease of use through an Arduino or a Raspberry Pi. To this must be added other IoT-supporting devices [23], which are one of the cornerstones as a source of information, both structured and semi-structured or unstructured, from Big Data [16].

It should not be forgotten other very important data to measure and used to watch user interaction, such as mouse clicks, keystrokes, page scrolling, reading time in an article, shared content, or interactions with content, as many social networks do, such as Facebook, Twitter, and Weibo. All this is often accompanied by photos, which further expand the multitude of formats and treatments with JPG, PNG, TIFF, or BMP.

Moreover, it is sometimes necessary to deal with legacy data in different databases, whether SQL or NoSQL, documents, emails, telephone conversations or scanned documents [24]. Other times, it may be information from web pages that are in HTML or well-structured XMLs or PDFs that do not have a structured way of displaying data. At other times, there are different data groups of different types in compressed files in RAR, RAR4, ZIP, 7Z, or TAR, which must first be tried to decompress and analyse its contents.

As can be seen, there are many formats, and here there are only a few examples of the most used ones. Each of these files also needs special treatment, even if they are of the same type. For example, in the case of images, not all formats have the same properties and small differences. In addition, every year, we have new devices or services that provide new useful data or the same data, but in other formats, or modifying how that information was shown, which implies modifying certain parts of the data-reading software. It must also be borne in mind that, in addition to the formats, all this also depends on the application, since it is not the same to analyse telescope or satellite images as social images or to analyse user data with time data.

Thus, this heterogeneity of data, incompatible formats and inconsistent data is one of the greatest barriers to effective data management [19]. In addition, these data are constantly changing and new systems are being added that either require or modify the way in which existing information is provided to the end user, contrary to the way it was done in the past, where there was only one structured database with a well-defined and slowly changing schema [25]. Some authors call it viscosity due to the fact that we have to use data from different sources, and sometimes it requires a transformation to use it [26]. This one is another of the problems; not every data we need has the same structure or format file if we require it for different sources, such as different governments, enterprises or portals. Then, we have to create different parsers and translate all the information into one common format. Other times, some of this information can be in a hard format to use, such as PDF or HTML, and the transformation is required to work easier with it.

4. Veracity: Origin, Veracity and Validity

Veracity is given because not everything that exists is true and false or erroneous data may be being collected. Therefore, when working, one should be careful with data sources and check that they are true data, thus trying to obtain accurate and complete information. This, in turn, can be divided into three sub-sections, namely, origin, veracity and validity.

Data origin is important to maintain quality, protect security and maintain privacy policies. It should be borne in mind that the data in Big Data move from the individual limits to those of groups or communities of interest and that these range from a regional or national limit to the international one. For this reason, the source helps to know where these data come from and what their original source is, for instance, by using metadata. This is very important because, knowing its origin, you can maintain the privacy of this data and thus be able to make some important decisions, such as the right to be forgotten. Other data to be inserted could be supply chain-related, such as calibration, errors, or missing data (timestamps, location, equipment serial number, transaction number, and authority).

Veracity includes the guarantee of the information of the means used to collect the information. A clear example is if sensors were used, which should include traceability, calibration, version, sampling time and device configuration. The reason for this can be a malfunctioned or uncalibrated device [27]. On the other hand, Online Social Networks give to us the opinion of the users, but maybe we cannot trust it [28].

Another example of a lack of or problems with veracity is the one mentioned in [29]. In this research, it is mentioned that, as a preliminary study to verify a certain assumption in 2006, the author used Google Trends to find out if the president who had been elected to the Real Madrid Football Club was the president with most queries on Google search engines. As the author points out, the surname of this president was Calderón. The problem, as he stated, is that on the same day, a president with the same surname was elected in Mexico. Thus, the problem was that Google Trends did not differentiate what Calderón was meant by each query; thus, it merged them. In other words, there was a lack of complementary data, a lack of veracity.

Validity refers to the accuracy and precision of the data, or rather the quality of the data. Examples can be given if, in continuous and discreet data on the gender of people and being male = 1 and female = 2, a 1.5 is received because this does not mean a new gender, but an error. Another type of error is the fraud of clicks on pages, clicks made by robots, hidden ads, etc.

An example of veracity is what the author of this article commented [30]; in it, he gave an example of several problems that had occurred in the United States of America. The first one is that politicians always like to talk about more data, but ultimately these are never among their priorities, as this has to compete with other things that offer a more immediate impact, which makes the money insufficient to keep the data accessible and digestible. The second problem is that data are institutionalised when they should be isolated from politicians, for example, with the alleged creation of the Environmental Statistics Agency (BES), which, in addition to collecting data, would analyse them. The third and final problem arises when chemical companies refuse to present their pollution data based on else; they reveal much useful data for their competitors. As stated in the article by David Goldton, it can be deduced that these data may not be easily accessible, outdated, incorrect, biased, or difficult to understand.

Another example, although some add it as part of the Variety and noise type, is the case of the Fukushima nuclear power plant, when people began to write about it and to share data from poorly calibrated sensors and/or that were not exactly in the said area [31]. In this case, it is a problem of Veracity due to the origin because of the lack of calibration and error, of veracity for being incorrect data and of validity for not being precise.

Clearly, if a large dataset is available, for instance, to see a tendency and there are few "bad" data, these will be lost and automatically ignored, thanks to the fact that they will be hidden among "good" data. However, if it is something casual, maybe this "bad" data can spoil the experiment, which makes the veracity of these data extremely important [6].

5. Variability: Structure, Time Access and Format

Variability is due to changes that the data have over time.

Maybe the data structure and how users want to interpret that data can change with time, modifying its structure, which would modify the way of parsing the data tree, perhaps by modifying the model in XML.

Other times, the time to access this data is different because they take more time to create the data or to update it, not always updating the information constantly. For instance, they do not update the information of that file more than once, as happens with some files from some Open Data portals. Another example is when they can delete the data, such as some companies or governments, when they are not obliged to keep the data for more than one year [28].

Another possible problem is when they change the format in which they are offered by migrating from one format, such as XML, to another, such as JSON, or the composition of these, adding or removing internal elements of their

structure.

Because of these problems, one has to be aware of these changes whenever one wants to update them, but the original data should also be kept unchanged, that is, to know how data were changing over time. This has the drawback of data redundancy and the necessary storage space. In addition, of course, the necessary system for monitoring changes and comparing them with existing data already stored in our system.

This point is very important because, having tools that process this data, it will be necessary to adapt them over time, as well as they should be able to detect new changes in the file; thus, a human being makes decisions about those changes: to add them, modify the programme, to avoid them, etc. Or maybe we have to change the source of data if they stop updating it or just delete it.

6. Value

The value of the data lies in what they can bring to the company and whether they can provide an advantage, as they help to make decisions based on questions that can now be answered thanks to the analysis of this data or in the discovery of trends. However, this value is variable because it depends on the data available, the hidden information they contain and whether it is found and extracted. Therefore, the challenge is to identify this value and extract it to perform an analysis of the data provided to us and to find this value. According to the survey analysed in [32], which corresponds to that carried out by MIT and IBM to 3000 executives, analysts and managers of more than 30 companies in 100 countries, one in five said they were concerned about data quality or ineffectiveness of government data as a major concern.

Thus, if these data with which one works have no value or have insufficient value for the company, project, or research, they will create a monetary loss because of storage, processing and human resources, as well as time. This is why, probably, it is the most important V, according to [28].

7. Visualisation

This V focuses on representing the data obtained in something that is readable, according to [31]. When dealing with so much data, many companies have had to hire people who are dedicated only to these visualisations; thus, they offer added and visual value to their employees. In addition, they created new tools for viewing these that worked in a correct and fast way. Moreover, due to the large size of the data to work, and sometimes at velocity, it becomes very difficult to create visualisation because the current tools have poor performance in terms of functions, scalability, and response time [1][23][33]. This is especially complicated when making real-time applications. Other times, with so much data available, the visualisation can happen to be difficult to understand.

An example of visualisation is the creation of Hive by Facebook [34], although now it is the one from the Apache Foundation, and it allows SQL-style queries to be performed using a command line. On the other hand, we have

Hue [35] that offers a graphical interface that gives Hive support as well as to other tools of the Hadoop ecosystem, in addition to providing graphics, monitoring and management.

Another case is that of eBay, where its employees can see the relevant data of users of the platform and thus be able to perform sentiment analysis on this data [31].

Then, we see that this V is important because it represents the challenges of visualising useful information for a user or company in a clear and fast way that allows the visualisation or decision making of the processed data.

References

1. Philip Chen, C.L.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. 2014, 275, 314–347.
2. González García, C.; Meana-Llorián, D.; G-Bustelo, B.C.P.; Lovelle, J.M.C. A review about Smart Objects, Sensors, and Actuators. Int. J. Interact. Multimed. Artif. Intell. 2017, 4, 7–10.
3. González García, C.; García-Bustelo, C.P.; Espada, J.P.; Cueva-Fernandez, G. Midgar: Generation of heterogeneous objects interconnecting applications. A Domain Specific Language proposal for Internet of Things scenarios. Comput. Netw. 2014, 64, 143–158.
4. Bell, G.; Hey, T.; Szalay, A. Beyond the Data Deluge. Science 2009, 323, 1297–1298.
5. McAfee, A.; Brynjolfsson, E. Big data: The Management Revolution. Harv. Bus. Rev. 2012, 90, 60–68.
6. NIST Big Data Public Working Group: Definitions and Taxonomies Subgroup. NIST Big Data Interoperability Framework: Volume 1, Definitions; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015; Volume 1.
7. Wu, X.; Zhu, X.; Wu, G.-Q.; Ding, W. Data mining with big data. IEEE Trans. Knowl. Data Eng. 2014, 26, 97–107.
8. Dewdney, P.E.; Hall, P.J.; Schilizzi, R.T.; Lazio, T.J.L.W. The Square Kilometre Array. Proc. IEEE 2009, 97, 1482–1496.
9. Greiner, L. What is Data Analysis and Data Mining? Available online: https://www.dbta.com/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx (accessed on 27 October 2022).
10. Sharp, A. Dispatch from the Denver debate. Available online: https://blog.twitter.com/2012/dispatch-from-the-denver-debate (accessed on 27 October 2022).

11. Michel, F. How Many Public Photos are Uploaded to Flickr Every Day, Month, Year? Available online: https://www.flickr.com/photos/franckmichel/6855169886/ (accessed on 27 October 2022).
12. Mervis, J. Agencies Rally to Tackle Big Data. Science 2012, 336, 22.
13. Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. Knowledge Discovery in Databases: An Overview. AI Mag. 1992, 13, 57–70.
14. Doctorow, C. Big data: Welcome to the petacentre. Nature 2008, 455, 16–21.
15. Howie, T. The Big Bang: How the Big Data Explosion Is Changing the World. Available online: https://blogs.msdn.microsoft.com/microsoftenterpriseinsight/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world/ (accessed on 27 October 2022).
16. Chen, M.; Mao, S.; Liu, Y. Big Data: A Survey. Mob. Netw. Appl. 2014, 19, 171–209.
17. Lazer, D.; Kennedy, R.; King, G.; Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. Science 2014, 343, 1203–1205.
18. Boyd, D.; Crawford, K. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 2012, 15, 662–679.
19. Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Gr. Res. Note 2001, 6, 70.
20. ACM SC08 International Conference for High Performance Computing, Austin, TX, USA, 15–21 November 2008. IEEE Computer Society: Austin, TX, USA. Available online: http://sc08.supercomputing.org/ (accessed on 27 October 2022).
21. Astrophysical Research Consortium. The Sloan Digital Sky Survey SDSS. Available online: https://www.sdss.org/ (accessed on 27 October 2022).
22. Bello-Orgaz, G.; Jung, J.J.; Camacho, D. Social big data: Recent achievements and new challenges. Inf. Fusion 2016, 28, 45–59.
23. Fan, W.; Bifet, A. Mining Big Data: Current Status, and Forecast to the Future. ACM SIGKDD Explor. Newsl. 2013, 14, 1.
24. Intel IT Center. Big Data Analytics. Intel’s IT Manager Survey on How Organizations Are Using Big Data; Intel Corporation: Santa Clara, CA, USA, 2012.
25. Dijcks, J.-P. Oracle: Big Data for the Enterprise; Oracle: Redwood, CA, USA, 2013.
26. Fatima Ezzahra, M.; Nadia, A.; Imane, H. Big Data Dependability Opportunities & Challenges. In Proceedings of the 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 3–4 October 2019; pp. 1–4.
27. Deepa, N.; Pham, Q.-V.; Nguyen, D.C.; Bhattacharya, S.; Prabadevi, B.; Gadekallu, T.R.; Maddikunta, P.K.R.; Fang, F.; Pathirana, P.N. A survey on blockchain for big data: Approaches,

opportunities, and future directions. Futur. Gener. Comput. Syst. 2022, 131, 209–226.

28. Khan, M.A.; Uddin, M.F.; Gupta, N. Seven V’s of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3–5 April 2014; pp. 1–5.
29. Gayo-Avello, D. No, you cannot predict elections with twitter. Internet Comput. IEEE 2012, 16, 91–94.
30. Goldston, D. Big data: Data wrangling. Nature 2008, 455, 15.
31. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 2017, 70, 263–286.
32. Lavalle, S.; Lesser, E.; Shockley, R.; Hopkins, M.S.; Kruschwitz, N. Big Data, Analytics and the Path from Insights to Value. MIT Sloan Manag. Rev. 2011, 52, 21–31.
33. Assunção, M.D.; Calheiros, R.N.; Bianchi, S.; Netto, M.A.S.; Buyya, R. Big Data computing and clouds: Trends and future directions. J. Parallel Distrib. Comput. 2015, 79–80, 3–15.
34. Thusoo, A.; Sarma, J.S.; Jain, N.; Shao, Z.; Chakka, P.; Anthony, S.; Liu, H.; Wyckoff, P.; Murthy, R. Hive—A warehousing solution over a map-reduce framework. Proc. VLDB Endow. 2009, 2, 1626–1629.
35. Apache Software Foundation. Hue. Available online: http://gethue.com/ (accessed on 27 October 2022).