Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 532–537

Procedia Computer Science

www.elsevier.com/locate/procedia

The 6th International Symposium on Emerging Information, Communication and Networks (EICN 2019)
November 4-7, 2019, Coimbra, Portugal

Risk Analysis of Using Big Data in Computer Sciences

Jesus Silva[a]*, Omar Bonerge Pineda Lezama[b], Ligia Romero[c] , Darwin Solano[d] , Claudia Fernández[e]

[a]*Universidad Peruana de Ciencias Aplicadas, Lima 07001, Peru*
[b]*Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula 21101, Honduras*
[c,d,e]*Universidad de la Costa, Cl. 58 # 55 – 66, Barranquilla 080001, Colombia*

Abstract

Today, as technologies mature and people are encouraged to contribute data to organizations' databases, more transactions are being captured than ever before. Meanwhile, improvements in data storage technologies have made the cost of evaluating, selecting, and destroying legacy data considerably greater than simply letting it accumulate. On the one hand, the excess of stored data has considerably increased the opportunities to interrelate and analyze them, while the moderate enthusiasm generated by data warehousing and data mining in the 1990s has been replaced by a rampant euphoria about big data and data analytics. But, is this as wonderful as seems? This paper presents a risk analysis of Big Data and Big Data Analytics based on a review of quality factors.

Keywords: Data management, data quality, decision making, data analysis.

1. Introduction

Big data characteristics are commonly described as volume, speed, and variety [1], leading to the widespread definition that they are very large, very fast, or very difficult to process [2]. Some specialists later added value to it, and occasionally, a fifth characteristic is mentioned: veracity [3]. It is important to distinguish the notion of a very

* Corresponding author. Tel.: +51-975737103.
E-mail address: jesussilvaUPC@gmail.com

1877-0509
10.1016/j.procs.2019.11.052

large dataset from a consolidation of two or more datasets from different sources into a single collection, whether physical or virtual nature. A broad way of referring to them is that they are massive quantities of data, for which, the applied mathematics replaces any other tool that might be needed, out of any theory of human behavior, from linguistics to sociology. It makes forget taxonomy, ontology, and psychology because when faced with massive data the old scientific approach, hypothesis-modelproof, becomes obsolete. Petabytes allow to say today that correlation is enough [4].

While the level of enthusiasm in business schools is scarcely less effusive [5-7] and observers from the information technology industry have expressed concern, there is little concern in the literature of Computer Science and Information Systems. However, some authors have expressed concerns, such as [2], [8], [9]. This paper analyzes the question of what risks arise from the lack of attention to quality factors in big data and data analytics, based on a review of data quality and decision factors in the big data context.

2. Method

For developing the document, a descriptive-type qualitative research based on the bibliographic review was proposed. According to the contributions of authors such as Stage & Manning [3] in the education sector, the qualitative methodology facilitates the analysis of characteristics, attributes, and particularities with the support of the theoretical foundation, making space for the contrast of postulates derived from research aligned with the identified problem and also promotes the reflection and deepening of interest topics for the generation of alternatives of continuous improvement and structuring of new processes. For the documentary review, parameters were established for the inclusion and selection of documents such as articles and research papers from the last six years, English and Spanish language literature, analysis of information from local and foreign universities, documents derived from databases as Scopus, Emerald, Scielo, Springer, and Redalyc. The gathered information was collected, organized, and summarized for the presentation of the following results, following a conceptual and theoretical order to provide a better understanding on the research topic [8].

3. Big Data Risks

Quality problems are evident when handling data, but the purpose is to show how risks are exacerbated in the context of big data. Table 1 describes the primary quality factors divided into two categories: 1) data quality, which can be assessed at the moment of the collection, and 2) information quality, which is not assessable until the data are used. Quality factors provide a framework within which analysis can be carried out.

The quality items of D2, D3 and D4 data are key factors in the meaning of the data, which is determined at the moment they are collected. However, since the cascade method of resource-intensive software development has declined, it is less common to establish and maintain a data dictionary. As a result, data definitions can be unclear, ambiguous, and even implicit. Lack of clarity about the original meaning increases the likelihood that it will change over time and that there will be different, and even mutually inconsistent uses of the same data element. At the time it is used, a more or less clear definition is overlapped by users' perspectives and interpretations.

Table 1. Quality factors

Data quality		Information quality	
ID	**Factor/Description**	**ID**	**Factor/Description**
D1	Syntactic validity: Conformity of the data content with the domain in which the data item is defined.	I1	Theoretical relevance: A demonstrable ability of the data element to make a difference in the decision-making process in which it is to be used.
D2	Appropriate entity association: A high level of confidence that the data item is associated with the particular real-world identity or entity whose attributes or representations it intends to represent.	I2	Practical relevance: A demonstrable ability of the content of the data element to make a difference in the decision-making process in which it is to be used.

Data quality		Information quality	
ID	**Factor/Description**	**ID**	**Factor/Description**
D3	Appropriate attribute association: On which real-world attributes is the data item intended to represent the lack of ambiguity	I3	Circulation: The absence of a material delay between a real-world occurrence and the recording of corresponding data
D4	Appropriate meaning of attributes: The absence of ambiguity about the state of the world's particular attribute that the content of the data intends to represent	I4	Comprehensiveness: The availability of sufficient contextual information so that the content of the data cannot be misunderstood
D5	Accuracy: A high degree of correspondence of data content with the real-world phenomenon it intends to represent, typically measured by a confidence interval.	I5	Controls: The application of business processes that ensure that data quality and information quality factors have been considered prior to the use of the data.
D6	Accuracy: The level of detail at which data content is captured, reflecting the domain in which valid content for that data element is defined.	I6	Auditability. Availability of metadata evidencing data quality and information quality factors
D7	Temporal applicability: The absence of ambiguity to the date or the period of time in which the content of the data represents or represented a real-world phenomenon		

Particularly, when data are frequently collected over time, the act of collection may involve compression of the data by sampling, filtering, and averaging. These actions affect the D5 (Accuracy) factor of data quality and the I4 (Completeness) of information quality. When looking for outliers, compression is likely to ensure that the most potentially relevant data are not in the collection.

The D2 factor of data quality requires that they have a reliable association with a particular real-world identity or entity. Data about any particular entity represent its digital person [10, 11]. The reliability of the association between a real-world phenomenon and a data shadow depends on the attributes that are selected as identifiers, but the association process presents error factors. In some circumstances, the link between the digital character and the underlying entity is challenging (pseudo-anonymity) and, in other cases, no link can be achieved (anonymity). In fact, in order to protect important interests and comply with relevant laws, it may be necessary to break any link (disidentification/anonymity) [12, 13].

On the other hand, rich data sets are vulnerable to re-identification procedures [14-16]. These problems affect all large data collections intended to assist in the management of long-term relationships. The problems are compounded by the expropriation of data for support purposes outside the original context of use, such as longitudinal research studies. Reliable identification processes are quite difficult in individual systems, because the challenges are multiplied when data from multiple sources are combined, particularly when the identifiers used by the underlying systems are not the same. Risks are particularly acute when data are sensitive. Social big data is one of such cases, but serious consequences also arise in other areas, such as health. Commonly, the big data movement includes the use of data for purposes other than their original purpose.

The quality problems of the data identified are exacerbated by the loss of context (Data Quality Factor I4), including the lack of clarity about the trade-offs applied at the time of collection, because the absence of this information greatly increases the likelihood of misinterpretation. This movement also often involves the further step of consolidating data, physically or virtually, from multiple sources. It depends on the links between data items whose semantics and syntax are substantially or subtly different, and the extent of misunderstandings and misinterpretations is multiplied. On the other hand, there are deficiencies in most datasets. Over time, however, many other integrity problems arise.

A common problem is the metadata loss, such as the scale at which they were originally collected, the definition at the time of collection, the source and any supporting evidence for the quality of the data, as well as the loss of contextual information including undocumented changes in its meaning over time. It undermines the quality factors of information I4 (completeness), I5 (controls) and I6 (auditability) and considerably increases the likelihood of inappropriate interpretation. To address perceived data integrity deficits, analysts design and deploy data debugging, cleansing, or purification processes [8]. Some of these processes use an external, authorized reference point, such as

a database of recognized location names and addresses. However, most of them lack an external key and simply rely on the logical quality of the data, i.e., the internal consistency within the consolidated datasets [9].

4. Risks of Big Data Analytics

There are several ways in which analytical tools can be applied on Big Data in order to test hypotheses, which can be predictions of theory, correlations that arise in other data sets, existing heuristics, or intuitions. Inferences can be made about digital relationships that can also be applied to the population of entities purporting to represent the data, or to segments of that population. Profiles can be constructed by categories of entities of specific interest, such as weather incidents or slimming diets, which can be isolated from many different types to be identified. Inferences can be made about individual entities, directly from a given digital relationship, or about apparent inconsistencies within a consolidated digital relationship or by comparing a digital relationship with similar entity relationships, or against a previously defined profile, or against a profile created through the analysis of that big data collection.

Big data analytics can be used as a decision system, which can be done formally, for example, by automatically sending infringement notifications or demonstrative arguments. However, it is also possible that it may become a decision system not through conscious decisions of the organization, but by default. This can happen when a decision maker is lazy or is replaced by a less experienced person who is not in a good position to check the reasonableness of the inferences obtained by the software. When decisions are made by analytics, or derived inferences are very influential in decision making, perhaps to the point of being a default decision that a person has to override, a number of concerns arise about the quality of the decision: does it reflect the scale, accuracy, and precision of the data that were decisive in making it? Was the inference mechanism actually used applicable to those categories of data? Does the data mean what the inference mechanism implicitly treated as meaning? To the extent that the data were consolidated from multiple sources, were those sources compatible with respect to scale, accuracy, precision, meaning, and completeness of the data? Alternatively, rather than as a form of automated decision making, big data analytics can be used as a support system, with a human decision maker who evaluates inferences before applying them to a real purpose.

However, the individual may have great difficulty in understanding the details of origin, quality, meaning, and relevance of the data, as well as the nature, prerequisites, characteristics, and limitations of the analytical technique, as well as the basements behind the recommendation. When each data element is collected, it represents a measure against a scale and some of them arise from measuring against a reason scale and can be analyzed using powerful statistical tools. However, data collected from cardinals, and even simple ordinal scales, are often happily assumed on a proportional scale in order to justify the application of statistical inference. Meanwhile, a large amount of data is collected against nominal scales, including text and images, which only support weak analytical tools, fuzzy concordance, and fuzzy logics. Another challenge arises when data that have been measured with different types of scales are consolidated and analyzed, since the applicability of the available analytical tools requires careful consideration.

A particularly serious challenge exists in the case of mixed-scale data, whose joint analysis is more a shady art than a precise science. All of these issues arise in the context of individual datasets, but get bigger when multiple datasets are consolidated [10]. A common feature of these circumstances is that decisions are made about complex real-world phenomena and therefore need to be represented by models that integrate the required variety, including explicit recognition of confusing, intervening, and missing variables. In practice, however, the models applied in big data analytics may be unduly simple and even merely implicit. A related concern is that correlations may be, and commonly are, low-grade, but are treated, perhaps implicitly, as if the relationships were causal, but causal in one direction and not in the other [12].

5. Discussion

Given the uncertainty of data quality and decision-making processes, many big data inferences currently offer greater credibility than they really justify, and resources will inevitably be misallocated. Within corporations, the

impact will ultimately be felt in a lower return on investment, while in public sector contexts there will be negative impacts on public policy outcomes. When big data analytics are inadequately applied to population inference and profiling, the harms that can occur include not only misallocation of resources, but also unjustified discrimination by and against segments of the population. In addition, when the profiles obtained are applied to generate suspicion, the result is an obscure predetermined characterization or pattern of infraction [13], based on a probabilistic cause rather than a probable cause [14].

This means unjustified impositions because the costs are borne by people, perhaps in the form of inconveniences, but sometimes with financial or psychological dimensions. The lack of transparency in relation to data and decision-making criteria gives rise to mysterious accusations, often indefensible, which can lead to unfair deprivation of rights. This represents a denial of natural justice and, in countries that are signatories to international conventions, a violation of the right to information about the nature and cause of the charge. In the literature on big data, the absence of discussion of the issues and impacts described in this paper is evident. Worse still, many documents dealing with these issues do not reflect the cumulative misunderstanding of the data and the quality of the decisions. Previously, the notion of data quality mining was essentially reduced to a mere internal consistency within data collection [6]. Even balanced big data assessments [4] do not address these issues and it is common that quality factors are not fully addressed in the specialized media, or worse still, that these issues are not even mentioned.

The limited guidance regarding the appropriate process is limited to dealing with internal consistency controls, overlooking data and information quality factors (see Table 1), and omitting controls and auditing [7]. Similarly, Marchand and Peppard [8] propose some process guidelines, but omit any meaningful consideration of processes to ensure the quality of the data and decision processes that are applied. Big Data success is inevitably linked to clear rules about data quality, and the high quality of data requires that they be consistent with time, content, meaning, and allow for unique and reliable identification [12]. This call has been almost entirely ignored to date.

6. Conclusions

Given that Big Data presents great problems as well as great opportunities, should authors recommend measures to avoid the bad and potentiate the good? Alternatively, can computer science disciplines and professions avoid these issues, believing that the responsibility lies in others? Management disciplines study these issues in the abstract, and managers and executives take responsibility for decisions in the real world. It is therefore debatable whether they are the ones who must be concerned about quality assurance and risk assessment and management. Professional associations considerably vary in the extent to which they impose responsibilities on their members, determining obligations in relation to the research process, but not to the impacts of their studies; so they must accept responsibility for making decisions consistent with the safety, health, and public welfare, and quickly revealing factors that may endanger the people or the environment.

References

[1] Schroeck, M. et al. (2012). Analytics: The real-world use of big data. IBM Institute for Business Value. University of Oxford.

[2] McAfee, A. & Brynjolfsson, E. (2012). Big data: The management revolution. Harvard Business Rev. 90, 61–68.

[3] Clarke, R. (2014). Promise unfulfilled: the digital persona concept, two decades later. Information Technology & People 27(2), 182–207

[4] Jagadish, H. et al. (2014). Big data and its technical challenges. Communications of the ACM 57(7), 86–94

[5] Boyd, D. & Crawford, K. (2012). Critical questions for big data. Information, Comm. & Society 15(5), 662–679. [22] Clarke, R. (2014). What drones inherit from their ancestors. Computer Law & Security Review 30(3), 247–262

[6] Amelec, V. (2015). Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. Advanced Science Letters, 21(5), 1406-1408.

[8] Guo, P. (2013). Data science workflow: Overview and challenges. Communications of the ACM Blog.

[9] Ariza, P., Pineres, M., Santiago, L., Mercado, N., & De la Hoz, A. (2014, November). Implementation of moprosoft level I and II in software development companies in the colombian caribbean, a commitment to the software product quality region. In 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV) (pp. 1-5). IEEE.

[10] Lis-Gutiérrez JP., Lis-Gutiérrez M., Gaitán-Angulo M., Balaguera MI., Viloria A., Santander-Abril JE. (2018) Use of the Industrial Property System for New Creations in Colombia: A Departmental Analysis (2000–2016). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[11] Gaitán-Angulo M., Cubillos Díaz J., Viloria A., Lis-Gutiérrez JP., Rodríguez-Garnica P.A. (2018) Bibliometric Analysis of Social Innovation and Complexity (Databases Scopus and Dialnet 2007–2017). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[12] Jesus Silva, Jenny Cubillos, Jesus Vargas Villa, Ligia Romero, Darwin Solano, Claudia Fernández. (2019). Preservation of confidential information privacy and association rule hiding for data mining: a bibliometric review. Procedia Computer Science 151; 1219–1224.

[13] Adhvaryu R, Domadiya N (2012). An Improved EMHS Algorithm for Privacy Preserving in Association Rule Mining on Horizontally Partitioned Database. In: Security in Computing and Communications Springer Berlin Heidelberg, pp: 272-280.

[14] G Li, M Xi. An Improved Algorithm for Privacy-preserving Data Mining Based on NMF. In: Journal of Information & Computational Science, 12(9) (2015), pp. 3423-3430

[15] Lis-Gutiérrez J.P., Henao C., Zerda Á., Gaitán M., Correa J.C., Viloria A. (2018) Determinants of the Impact Factor of Publications: A Panel Model for Journals Indexed in Scopus 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[16] Isasi, P., Galván, I.: Redes de Neuronas Artificiales. Un enfoque Práctico. Pearson. ISBN 8420540250 (2004).