Towards a Data Collection Quality Model for Big Data Applications

Mohammad Abdallah(✉), Alaa Hammad, and Wael AlZyadat

Faculty of Science and Information Technology,
Al-Zaytoonah University of Jordan, Amman, Jordan
{m.abdallah,wael.alzyadat}@zuj.edu.jo

Abstract. Big Data and its uses are widely used in many applications and fields; artificial information, medical care, business, and much more. Big Data sources are widely distributed and diverse. Therefore, it is essential to guarantee that the data collected and processed is of the highest quality, to deal with this large volume of data from different sources with caution and attention. Consequently, the quality of Big Data must be fulfilled starting from the beginning; data collection. This paper provides a viewpoint on the key Big Data collection Quality Factors that need to be considered every time the data are captured, generated, or created. This study proposes a quality model that can help create and measure data collection methods and techniques. However, the quality model is still introductory and needs to be further investigated.

Keywords: Big Data · Data collection · Data quality · Data collection quality · Data collection quality model

1 Introduction

Collecting a huge amount of data about the business, health, or anything else. Then analyze this huge amount of data will give a better understanding of the field and will help in taking the right decisions and predictions. Therefore, Big Data plays a significant role in life.

Big Data (BD) is a technological word for the large quantities of heterogeneous data that are quickly generated and distributed, and for which traditional methods for processing, analyzing, retrieving, storing, and visualizing such huge sets of data are now unsuitable and insufficient. This can be seen in a variety of fields, including sensor-generated data, social networking, and digital media uploading and downloading. Big Data can be structured, semi-structured, or unstructured.

The data in any system or application is a vital part. It can be collected from different sources or created in many ways. So, the data must be correct, accurate, and complete to get the maximum benefits out of it [1].

Data collection is the procedure of collecting, measuring, and analyzing accurate insights for research using standard validated techniques. So, Data collection quality becomes an essential part of any data processor data management. In the Big Data period, data quality considerations and problems have already been addressed which

W. Abramowicz et al. (Eds.): BIS 2021 Workshops, LNBIP 444, pp. 103–108, 2022.
https://doi.org/10.1007/978-3-031-04216-4_11

need to be evaluated. The Big Data quality variables also include the factors focused on the data itself and the management of Big Data and customer needs [2, 3].

However, the data source or data collection quality has not been questioned, discussed, and measured deeply in literature. In this research, we have proposed a quality model to measure the data collection process.

In Sect. 2, some related work about data collection quality factors and models are explored. In Sect. 3, a data collection quality models was proposed and finally in Sect. 4 the conclusions and future direction of the research are highlighted.

2 Related Works

The data is a set of qualitative or quantitative variables, it can be in many different forms and formats. Big Data is distinguished from any other data type by the 3V's dimensions; Volume, Velocity, and Variety [4]. The 3V's of Big Data was also the start of Big Data quality factors. Volk et al. [5] have summarized the quality issues that can be raised by the 3 V's of Big Data as: "*Handling huge volumes of data in different formats at high speeds while maintaining resiliency and data security, can be very challenging*".

Big Data and apps are a problem of the consistency of Big Data. Any program that uses it must ensure that data have high-quality requirements to deliver a good quality system. Especially the factors of quality that take into account Big Data are the same as for conventional sets of data. Furthermore, certain additional quality factors deal with many of the facts, such as data management and repair [6].

The consistency factors of data have been debated by most researchers in past studies. However, the data and Big Data have several common indicators of consistency and vary in quality and calculation applications [7].

Data collection is one of the processes that are essentials in any data sets or Big Data application. Big Data Accuracy, Completeness, Redundancy, Readability, Accessibility Consistency, Trusts are the primary data and Big Data Quality Factors. To fulfill the quality factor, every factor is connected to one or more quality parameters or criteria. However, these factors have discussed the data quality from all perspectives; the Data perspective, Management perspective, Processing, and Service perspective [8].

The Big Data quality measurement is not only focused on data quality. Data Quality Management (DQM) is also a quality challenge that must be considered [9], which intersects with data collection in its aspects. DQM has five main aspects [10, 11]: People, Data Profiling, Defining Data Quality, Data Reporting, Data Repair.

Data creation may introduce a group of mistakes. This may be caused by human influence such as typos, misunderstanding, or misrepresentation [12]. As well as data collection tools and techniques [13].

To decrease their influence, several strategies have been applied. Those include procedures to reduce possible causes of mistakes, such as better instructions or the simplification of forms [12]. Heinrich et al. [14] have an application of rule sets and statistical analyses. There metric makes it possible to regard laws with a unique probability that are likely to be followed. The resulting metric values are likely to free the evaluated dataset from internal inconsistencies about the ambiguous rules and thus provide a consistent understanding.

Savosin et al. [15] raised issues that edit Big Data professional sources without authorization might cause them to shut down, although the information itself can be used freely. Ordinary users will, on the other hand, offer a great deal of information with a little pause in different fields of expertise. However, depending on the subject and external factors such as the precise usage of sensor instruments, the distribution of users in that region, etc., their efficiency and precision can vary.

Liu et al. [16] proposed a non-linear optimization programming model with resource management constraints by creating a data-quality Petri net to catch the mechanism by which the information system produces, spreads, and builds problems in the data quality.

As seen from the literature there is a lack of a general model that deals with data collection quality issues. The researchers only try to avoid or tolerate the data collection or creation quality issues. In the next section, a proposed quality model for data collection will be introduced. This model should help in preventing the data from being collected in the wrong way and can help to produce clean, reusable Big Data sets.

3 The Proposed Data Collection Quality Model

The process of data collection is one of the basic stages in improving and developing the quality of the data and the resulting information. The process of data collection is not limited to specific areas, but it is used in many sectors such as Technology, Health care, Engineering, and many others. Therefore, it is become an important step to make sure that the collected data is not only correct but also of good quality. So, this stage requires a lot of accuracy, time, effort, and sometimes cost but to obtain the highest benefit and quality of data, there is a set of proposed conditions for application in the data collection stage.

The proposed quality model for collecting data to be used in Big Data systems is accumulative different quality models for data collection for different purposes. Therefore, we hope this quality model can help any data collection applications that collect data for any reason. The proposed quality model has 7 quality factors, as shown in Table 1.

Reliability of the data source: the data must be correct, complete, and consistent, and coherent with the 4c's of data. A dependable source presents a well-researched and well-supported hypothesis, claim, or discussion based on strong evidence [17]. Reliability is a cornerstone in any quality model, since it is related to the correctness and dependability.

Trustworthy: The data providers will not give the right data if they do not trust the data recipient and they are not convinced about the purpose of data collection. Therefore, it is important to work to build trust between the data providers (the first part) and the data recipient (the second part). As a result of building trust; the effort, time, and cost will be reduced [18]. It also considers the data legitimate and the data provider must be aware of why the data is collected and how it will be used [19].

Table 1. Data collection quality factors

Quality factor	Quality dimensions	References
Reliability of the data source	Correctness, completeness, consistent, coherent, dependability	[17]
Trustworthy	Accuracy, legitimate, intrinsic	[18, 19]
Data suitability	relevance, fitting, usefulness	[20, 21]
Data preservation	Availability, validity, accessibility	[22]
Data integrity	Security, privacy	[23]
Rapid data collection	Timelines	[24]
Data reusability	Understandable, renewable and composable	[25–27]

Data Suitability: The data collected should be related to the purpose of the data use. Collecting data from unrelated sources will produce a huge amount of unfitting data, which increases the time, effort, and cost of data cleansing and analysis [20, 21]. If the data is irrelevant then for sure the decisions that are made depending on it will be faulty and misleading.

Data Preservation: The data should be ensured to be available, valid, and accessible in the long run. Preserving efforts should ensure that the content is accurate, secure, and accessible while maintaining its dignity, such as authentication, signing metadata for the protection, assigning representation records, and ensuring appropriate data structures and file formats [22]. However, the data provider and the data recipient should have an agreement about how long and where the data will be preserved. However, the data can become outdated if it was kept for too long time. Therefore, the data preservation should be correctly applied.

Data Integrity: The data must be secured and encrypted during data collection and transmission. In some cases, the data is collected by a third party. Thus, the data must be not accessible or readable by the third party and only the data recipient can decrypt the data and read it. This means the data must keep its integrity from the source till the last destination [23]. Data leak can cause many problems. Therefore, the data integrity is one of the important quality factors that must be in any data quality model.

Rapid Data Collection: Real-time data (RTD) is the data that is delivered immediately after its collection. For navigation and tracking, real-time data is often used. Real-time computation is usually used to process such data, but it can also be saved for review later or offline [24]. Collecting huge amount of data needs to be fast to get the most of the data needed in a short time. Otherwise, the system will take ages to collect the data needs and that may effects its accuracy.

Data Reusability: The collected data should be understandable, renewable, and composable. The process of cleaning and converting raw data before processing and analysis are known as data preparation. It's a crucial step before processing that usually entails reformatting data, making data corrections, and merging data sets to enrich data. For data professionals or business users, data preparation may be time-consuming, but it is essential to place data in perspective to transform it into information and remove prejudice caused by poor data quality. Standardizing data formats, enriching source data, and/or eliminating outliers are all common steps in the data preparation phase. Also, Data reuse refers to the practice of repurposing data for a different task or purpose than it was created for. The use of suitable metadata schemas will help to describe datasets and enable them to be reused over time [25].

4 Conclusions and Future Directions

Big Data is one of the quickest tools used in many applications nowadays. Data inspires agriculture, healthy production, and a variety of market choices. Data quality must be taken into account and assessed because bad decisions are based on low data quality. In this research, a quality model for the data collection process was introduced. The model has 7 quality factors that measure the data collection process. The factors start with the reliability of the data source then the regulations of data collecting and then the nature of the targeted data.

In the future, the model may be expanded and evaluated against other quality models. It also may be applied in some experiments as well. Also, the quality model can be used to produce Big Data area-specific data collection quality models.

References

1. Staegemann, D., Volk, M., Nahhas, A., Abdallah, M., Turowski, K.: Exploring the specificities and challenges of testing big data systems. In: 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS 2019), Sorrento, Italy, pp. 289–295. IEEE (2019)
2. Batini, C., Rula, A., Scannapieco, M., Viscusi, G.: From data quality to big data quality. J Database Manage. **26**(1), 60–82 (2015). https://doi.org/10.4018/jdm.2015010103
3. Staegemann, D., et al.: A preliminary overview of the situation in big data testing. In: In 6th International Conference on Internet of Things, Big Data and Cyber Security (IoTBDS 2021), pp. 296–302. SciTePress (2021)
4. Laney, D.: 3D management: controlling data volume, velocity, and variety. Application delivery strategies. META Group Res. Note **6**(70), 1 (2001)
5. Volk, M., Staegemann, D., Pohl, M., Turowski, K.: Challenging big data engineering: positioning of current and future development. In: In 3rd International Conference on Internet of Things, Big Data and Cyber Security (IoTBDS 2019), pp. 351–358 (2019)
6. Staegemann, D., et al.: Challenges in data acquisition and management in big data environments. In: 6th International Conference on Internet of Things, Big Data and Security, pp. 193–204. SciTePress (2021)
7. Abdallah, M.: Big data quality challenges. In: The International Conference on Big Data and Computational Intelligence (ICBDCI 2019), Pointe aux Piments, Mauritius, pp. 1–3. IEEE (2019)
8. Abdallah, M., Muhairat, M., Althunibat, A., Abdalla, A.: Big Data quality: factors, frameworks, and challenges. COMPUSOFT Int. J. Adv. Comput. Technol. **9**(8), 3785–3790 (2020)

9. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM **41** (2), 58–65 (1998). https://doi.org/10.1145/269012.269022
10. Lebied, M.: The Ultimate Guide to Modern Data Quality Management (DQM) For an Effective Data Quality Control Driven by the Right Metrics. https://www.datapine.com/blog/data-quality-management-and-metrics/. Accessed 18 July 2021
11. Galin, D.: Software Quality: Concepts and Practice, 1 edn. Wiley-IEEE Computer Society (2018)
12. Janssen, M., van der Voort, H., Wahyudi, A.: Factors influencing big data decision-making quality. J. Bus. Res. **70**, 338–345 (2017)
13. Izadi, D., Abawajy, J.H., Ghanavati, S., Herawan, T.: A data fusion method in wireless sensor networks. Sensors **15**(2), 2964–2979 (2015)
14. Heinrich, B., Klier, M., Schiller, A., Wagner, G.: Assessing data quality – a probability-based metric for semantic consistency. Decis. Support Syst. **110**, 95–106 (2018)
15. Savosin, S.V., Mikhailov, S.A., Teslya, N.N.: Systematization of approaches to assessing the quality of spatio-temporal knowledge sources. In: Journal of Physics: Conference Series, vol. 1801, no. 1, p. 012006 (2021)
16. Liu, Q., Feng, G., Zhao, X., Wang, W.: Minimizing the data quality problem of information systems: a process-based method. Decis. Support Syst. **137**, 113381 (2020)
17. Abowitz, D.A., Toole, T.M.: Mixed method research: fundamental issues of design, validity, and reliability in construction research. J. Constr. Eng. Manag. **136**(1), 108–116 (2010)
18. Byabazaire, J., O'Hare, G., Delaney, D.: Using trust as a measure to derive data quality in data shared IoT deployments. In: 29th International Conference on Computer Communications and Networks (ICCCN), Honolulu, HI, USA, pp. 1–9 (2020)
19. Berger, C., Stefani, P.D., Oriola, T.: Legal implications of using social media data in emergency response. In: 11th International Conference on Availability, Reliability and Security (ARES 2016), Salzburg, Austria, pp. 798–799 (2016)
20. Khayyat, Z., et al.: BigDansing: a system for big data cleansing. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, pp. 1215–1230 (2015)
21. Maletic, J., Marcus, A.: Data cleansing: beyond integrity analysis. Iq: Citeseer, pp. 200–209 (2000)
22. Berman, F.: Got data? A guide to data preservation in the information age. Commun. ACM **51**(12), 50–56 (2008)
23. Lebdaoui, I., Hajji, S., Orhanou, G.: Managing big data integrity. In: International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco, pp. 1–6 (2016)
24. Zeng, Y., Jiang, W., Wang, F., Zheng, X.: Real-time data collection and management system for emergency spatial data based on cross-platform development framework. In: IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC 2019), Chengdu, China, pp. 260–265 (2019)
25. Mehboob, B., Chong, C.Y., Lee, S.P., Lim, J.: Reusability affecting factors and software metrics for reusability: a systematic literature review. Softw. Pract. Exp. **51**, 1416–1458 (2021)
26. Alzyadat, W., AlHroob, A., Almukahel, I.H., Muhairat, M., Abdallah, M., Althunibat, A.: Big data, classification, clustering and generate rules: an inevitably intertwined for prediction. In: The International Conference on Information Technology (ICIT 2021), Amman, Jordan, pp. 149–155. IEEE (2021)
27. Staegemann, D., Volk, M., Lautenschläger, E., Pohl, M., Abdallah, M., Turowski, M.: Applying test driven development in the big data domain–lessons from the literature. In: The International Conference on Information Technology (ICIT 2021), Amman, Jordan, pp. 511–516. IEEE (2021)